

Vision in Action

Supplementary Materials

Anonymous Author(s)

Affiliation

Address

email

1 Experiment Details

Please find the videos of our policy rollouts on: <https://via-corl.github.io/>.

1.1 Success Criteria

Bag Task: The task is divided into three stages: *open*, *grasp*, and *take out*. We define stage-wise success as: the *open* stage is successful if the robot successfully grasps the bag strap and uncovers hidden areas. The *grasp* stage is successful if the robot localizes and grasps the target object. The *takeout* stage is successful if the robot removes the object from the bag.

Cup Task: The task is divided into three stages: *find & pick up*, *handover*, and *place*. The *find & pick up* stage is successful if the robot moves its arm to the correct section and successfully grasps the cup. The *handover* stage is successful if the cup is transferred from the right hand to the left without being dropped. The *place* stage is successful if the cup is accurately placed on the saucer.

Lime & Pot Task: The task is divided into three stages: *find*, *pick up*, and *bimanual grasp & align*. The *find* stage is successful if the lime appears in the robot’s camera view and the robot initiates a arm reaching motion toward it. The *pick & place* stage is successful if the robot grasps the lime and places it in the pot. The *bimanual grasp & align* stage is successful if the robot grasps both pot handles simultaneously and places the pot on the mat with proper alignment.

1.2 Training and Evaluation Initial Configuration

Bag Task: During both training and evaluation, the bag remains in a fixed position with the zipper open. One object is placed inside the bag. For training, we use five different objects, each placed at random positions within the bag. For evaluation, we use two different objects, each placed at five fixed positions, resulting in five evaluation configurations per object. These setups are visualized in the main paper figures.

Cup Task: Shelves A and B are placed at fixed positions on the table to serve as occluders. During training data collection, a yellow cup is randomly initialized on Shelf A, which is divided into two sections: upper and lower. The saucer is randomly initialized in a hidden position beneath Shelf B. During data collection, the object is reset and placed randomly on the shelf. To avoid introducing bias, the teleoperator is instructed to close their eyes during this reset process.

During evaluation, we defined 10 test configurations for initializing the cup and saucer. Five locations on the upper section of shelf A (e.g. `upper_1` to `upper_5`) and five locations on the lower section (e.g. `lower_1` to `lower_5`) were used for the cup placement. Similarly, five locations beneath shelf B (e.g. `shelf_B_1` to `shelf_B_5`) were used for the saucer placement. Each test configuration paired a cup location with a saucer location, such as `upper_1` with `shelf_B_1`, `upper_2` with `shelf_B_2`, and so on, including combinations like `lower_1` with `shelf_B_1`

34 and `lower_2` with `shelf_B_2`. All models were evaluated under these 10 configurations with a
35 fixed robot’s initial pose. Each configuration is evaluated twice, resulted in 20 rollouts.

36 ***Lime & Pot Task:*** A shelf is placed on the right side of the table to serve as an occluder for the
37 lime. This shelf can be reconfigured into different shapes, creating various occlusion patterns (see
38 image). During training data collection, the shelf is randomly positioned on the right side of the
39 table, and its shape is randomly configured. The lime is placed at random 3D locations—such as
40 on the shelf, beneath it, or on the left side of the table. The pot remains at a fixed position, with
41 minimal variation introduced to support robustness learning. During data collection, the lime is
42 reset and placed randomly. To avoid introducing bias, the teleoperator is instructed to close their
43 eyes during this reset process.

44 During evaluation, we define 10 test configurations for initializing the lime and the shelf. Specifi-
45 cally, we use two shelf configurations, each paired with five distinct lime placements. The pot initial
46 configuration is fixed. All models are evaluated under these 10 configurations using a fixed initial
47 robot pose. Each configuration is tested twice, resulting in a total of 20 rollouts.

48 2 Policy Training Details

49 All models are trained for 500 epochs with a learning rate of 0.0001 and a batch size of 64. Following
50 the Diffusion Policy framework, we use a DDIM scheduler with 50 denoising steps during training,
51 and 16 steps during evaluation. The model observes a single frame, predicts 16 future actions, and
52 executes 8 actions per rollout.

53 3 VR Teleoperation Details

54 The VR device tracks the user’s head pose relative to their body frame. This body frame is initialized
55 at the beginning of each VR session, with its origin defined on the ground. Before starting a VR
56 teleoperation session, we ask the user to sit still and face forward to ensure a consistent reference.
57 Once the session begins, the head pose in the body frame is transformed into the robot’s world frame
58 by applying a height offset. This resulting absolute pose in the robot’s world frame is then used to
59 control the robot’s neck end-effector pose. For training data collection, the teleoperator is instructed
60 to adopt a consistent starting pose for both the neck and the arms, with minimal variation.

61 Empirically, we found that having the user hold a physical interface during teleoperation helps them
62 better coordinate their upper body, as opposed to waving their arms in the air without support. This
63 motivates us to use additional teaching arms for arm teleoperation instead of relying on VR hand
64 tracking.

65 4 User Study Details

66 **Post-Session Experience Survey.** We asked users to provide feedback on their teleoperation expe-
67 rience through a post-session survey, which is included at the end of this supplementary document.

68 **Stereo-RGB Streaming Setup.** We mounted a ZED Mini camera on the end-effector of the robot
69 neck to provide stereo RGB streaming. The rest of the setup remained consistent with our teleoper-
70 ation setup.

User ID___ Date___ Time___

User Background

1. VR experience: How familiar are you with using Virtual Reality (VR) devices?

- ☐ Never used VR before
- ☐ Tried it once or twice
- ☐ Occasionally use (a few times a year)
- ☐ Regular user (a few times a month)
- ☐ Frequent user (weekly or daily)

2. Robot Teleoperation experience: What is your level of experience with robot teleoperation?

- ☐ Never used robot teleoperation before
- ☐ Beginner (basic familiarity; used a few times)
- ☐ Intermediate (occasional use)
- ☐ Advanced (frequent use with confidence)
- ☐ Expert (researcher or professional experience)

choose a number on a scale from 1 to 5 for each of the questions:

1. How **motion-sick** did you feel during VR teleoperation with the following systems?

1 never feel motion sick, 5 high unable to tolerate

For system A:

For system B:

2. Mental Demand: How mentally demanding was the task?

1 Very low 5 Very high

For system A:

For system B:

3. Physical Demand: How physically demanding was the task?

1 Very low 5 Very high

For system A:

For system B:

4. Temporal Demand: How hurried or rushed was the pace of the task?
1 Very rushed 5 Not rushed

For system A:

For system B:

5. Performance: How successful were you in accomplishing what you were asked to do?
1 Very successful 5 Not successful

For system A:

For system B:

6. Effort: How hard did you have to work to accomplish your level of performance?
1 Very hard 5 Not hard

For system A:

For system B:

7. Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?
1 Very frustrated 5 Not frustrated

For system A:

For system B:

Preference:

If you were asked to deliver high-quality, high-volume demonstrations via teleoperation on a daily basis, which system would you prefer to use?

- ☐ System A
☐ Neutral (no preference)
☐ System B

Information for internal records only completed by the tester:

Time takes for completing task in system A _____

Time takes for completing task in system B _____

System A: RGB or Point Cloud _____