

## A Supplementary Material

We first go over policy architecture details in Section 4. We then present additional descriptions of the simulated experiment setup in Section 5.1 and the real robot experiment setup in Section 5.2. Lastly, we show results for one additional ablation experiment for the real robot experiments in Section 5.2. For visualization of the real robot experiments as well as the zero-shot skill chaining experiments mentioned in Section 5.2, please see the videos attached.

### A.1 Architecture Implementation Details

In Section 4 of the main paper, we overviewed the architectural choices. Here, we provide a more detailed description of the implementation details. In both simulated and real robot experiments, we train a separate policy for each task. For both environments, we use DINO-v2 with ViT-B/16 backbone to encode objects and parts. In simulated experiments, the policy network is implemented as a 4-layer MLP with hidden sizes [512,256,128], and the concatenation of all token outputs from the attention layer is taken in. In real robot experiments, we use a 3-layer MLP with hidden sizes [1024,1024] instead. Under the robot’s hardware constraint, we only input the CLS token into the policy MLP to reduce the number of parameters. All methods share the same policy network architecture.

### A.2 Simulated Experiments Setup

In Section 5.1 of the main paper, we briefly describe the five simulated tasks. Now we will go over a detailed description of each task and how the task-relevant objects are selected: In **OpenMicrowave**, the goal is to open a microwave sitting on a kitchen counter. It requires the agent to locate the microwave and its handle. In **SlideCabinetDoor** and **OpenCabinetDoor**, agents need to locate the handle of cabinet doors and open them. In **TurnOnLight** and **TurnKnob**, agents need to turn the perspective knobs on a panel. In **OpenMicrowave**, the task-relevant object is selected by prompting GroundedSAM with “microwave.” In **SlideCabinetDoor**, the task-relevant object is selected by prompting GroundedSAM with “cabinet.” For the rest of the tasks, we annotate the task-relevant object locations. Note that since the positions of objects in the environments are fixed, we only need to annotate the position of the task-relevant objects once.

### A.3 Real Robot Experiment Setup

We use a 7-DoF Franka robot arm with a continuous joint-control action space at 15 Hz. A Zed 2 camera is positioned on the table’s right edge, and only its RGB image stream—excluding depth information—is employed for data collection and policy learning. Another Zed mini camera is affixed to the robot’s wrist. We encode the wrist image with DINO-v2 and pass the CLS token as an additional token to the policy during training. Operating under velocity control, our robot’s action space encompasses a 6-DoF joint velocity and a singular dimension of the gripper action (open or close). Consequently, the policy produces 7D continuous actions.

### A.4 Additional Ablation Experiment for Real Robot Setup

Similar to the simulated experiments, we perform the ablation experiments Ours—multi-level where we remove object decomposition. Our main observation is that compared to this ablation, our method performs more robustly in more complicated tasks where identifying parts is crucial to the task’s success. Especially in `Pout Water From Kettle into Pot`, where a firm and secure grasp is needed to pick up the kettle and precise location of the pot is needed, Ours—multi-level succeed 11 times in IND setup and only 6 times in OOD setup, proving that having the ability to identify and locate the parts greatly improves the task success rate in both IND and OOD cases. Full results are in Table 2.

Method \ Task	Eggplant-Sink		Kettle-Stove		Faucet		Eggplant-Pot		Water-Pot		Overall	
	IND	OoD	IND	OoD	IND	OoD	IND	OoD	IND	OoD	IND	OoD
# of Trials	15	15	15	15	15	15	15	15	15	15	75	75
HODOR (Ours)	<b>14</b>	<b>12</b>	<b>13</b>	<b>12</b>	12	<b>8</b>	<b>13</b>	<b>8</b>	<b>12</b>	<b>9</b>	<b>64</b>	<b>49</b>
Ours—multi-level	14	12	11	11	12	8	10	8	11	6	58	45

Table 2: IND and OOD BC Results on Real Robot Tasks. We report the number of success of each task out of 15 trials.