

545 A Supplementary Material

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

551 A.1 Architecture Implementation Details

552 In Section 4 of the main paper, we overviewed the architectural choices. Here, we provide a more
553 detailed description of the implementation details. In both simulated and real robot experiments,
554 we train a separate policy for each task. For both environments, we use DINO-v2 with ViT-B/16
555 backbone to encode objects and parts. In simulated experiments, the policy network is implemented
556 as a 4-layer MLP with hidden sizes [512,256,128], and the concatenation of all token outputs from
557 the attention layer is taken in. In real robot experiments, we use a 3-layer MLP with hidden sizes
558 [1024,1024] instead. Under the robot’s hardware constraint, we only input the CLS token into
559 the policy MLP to reduce the number of parameters. All methods share the same policy network
560 architecture.

561

562 A.2 Simulated Experiments Setup

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

573 A.3 Real Robot Experiment Setup

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

581 A.4 Additional Ablation Experiment for Real Robot Setup

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

590

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)	14	12	13	12	12	8	13	8	12	9	64	49
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.