RAGNet: Large-scale Reasoning-based Affordance Segmentation Benchmark towards General Grasping
Supplementary Files

Dongming Wu[1], Yanping Fu[2], Saike Huang[3], Yingfei Liu[3], Fan Jia[3], Nian Liu[4], Feng Dai[2], Tiancai Wang[3], Rao Muhammad Anwer[4], Fahad Shahbaz Khan[4], Jianbing Shen[5]

[1] The Chinese University of Hong Kong, [2] Institute of Computing Technology, Chinese Academy of Sciences,

[3] Dexmal, [4] Mohamed bin Zayed University of Artificial Intelligence, [5] SKL-IOTSC, CIS, University of Macau

1. Details of Data Annotation

As the original data sources, such as HANDAL [3], Open-X [6], EgoObjects [13], GraspNet [2], provide original annotation information (*e.g.*, ground-truth boxes or masks), we make full use of them for minimal human intervention. From these data, we emphasize grasping-oriented objects, encompassing both those with handles and those without. Therefore, as described before, we design five annotation tools: ❶: Original mask, ❷: SAM2 [8], ❸: Florence2 [10] + SAM2, ❹: VLPart [9] + SAM2, ❺: Human (+ SAM2). The details of how to compose these tools for one specific dataset are shown in Table 2. In addition, Table 2 provides the reasoning instruction annotation details. *In summary, our dataset RAGNet includes a broad range of data domains, categories, and reasoning instructions, establishing a robust basis for open-world grasping applications.*

2. Affordance Annotation Examples

Since our benchmark RAGNet includes a significant number of grasping-oriented objects from various domains (like robot, wild, and ego-centric domains), we highlight this aspect by showcasing additional examples of affordance segmentation annotations in Fig. 2. For each affordance map annotation, the candidate objects are initially identified to determine if they possess a handle. Then, their affordance maps are carefully annotated according to our tool priority. For example, the banana from Open-X [6] is segmented using the combination of Florence2 [10] and SAM2 [8] according to its original grasping instruction. The knife handle from EgoObjects [13] can be accurately grounded using VLPart [9], and its output box can be further transformed into a pixel-wise mask using SAM2 [8]. Regarding the microwave handle in the EgoObjects dataset [13], it has been annotated manually because there is no suitable tool available for automated annotation. In conclusion, we collect a total of 273k diverse images along with their corresponding affordance annotations.

3. Reasoning-based Affordance Examples

More reasoning-based affordance segmentation examples are shown in Fig. 4. It contains two types of instructions, easy instructions and hard instructions. As seen, the easy instructions include the target object name, while the hard ones only include functional descriptions without the object name. These instructions are generated by GPT-4 and the corresponding prompts used in GPT-4 are listed in Table 3 and Table 4. The highlighted "words" are category names at most times. We sometimes provide additional keywords about potential grasping action for some categories, for aligning the instructions with the image content. For example, if the microwave is closed, we would assign the keywords "microwave, open the door".

4. Implementation Details of AffordanceNet

Beyond our reasoning-based affordance segmentation data, we also incorporate a variety of generic segmentation datasets into our training. This diverse generic set includes data for semantic segmentation (*e.g.*, ADE20k [12], COCO-Stuff [1], PACO [7]), referring segmentation (*e.g.*, RefCOCO [11]), VQA (*e.g.*, LLaVA-150k [5]) and reasoning-based segmentation (*e.g.*, ReasonSeg [4]). The data sampling ratios are presented in Table 1. We deploy eight NVIDIA A100 GPUs (80GB) to train our model, with a learning rate of 2e-5. The training loss follows [4], which uses binary focal loss for map prediction and cross-entropy loss for text output. We utilize a batch size of 40 without gradient accumulation.

5. More Results on Visual Affordance

We provide more visualization results of affordance segmentation from our AffordanceVLM model in Fig. 4. The testing images are selected from multiple validation

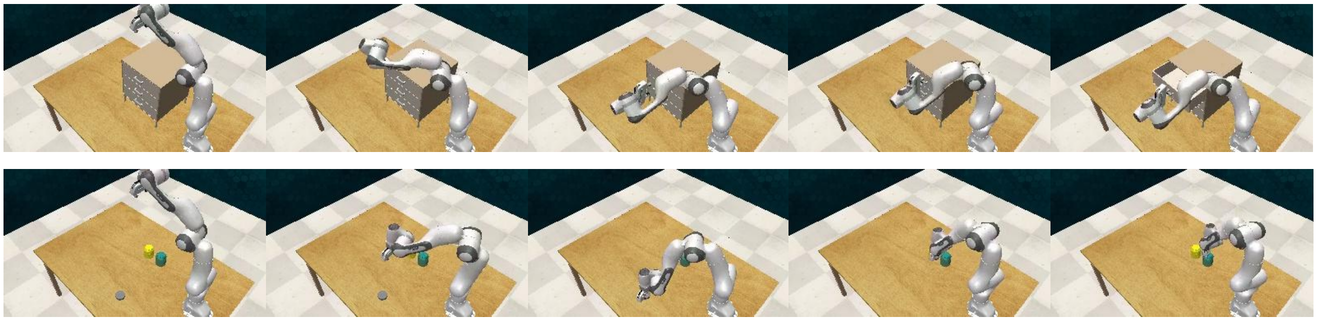


Figure 1. **Object grasping results from our AffordanceNet on RLBench**. The instruction of the top video is "*Open the top drawer*", and the bottom one refers to "*Close the green jar*".

Data	Semantic Seg	Referring Seg	Reasoning-based Seg	VQA	Affordance Seg	Reasoning-based Affordance Seg
Ratio	3	1	1	1	9	3

Table 1. **Data sampling ratios during training.**

sets, such as GraspNet `Novel`, 3DOI, and HANDAL. We employ template-based, easy reasoning-based, and hard reasoning-based instructions for affordance map prediction, respectively. It is obvious that our AffordanceVLM can understand these high-level human instructions, and transform them into precise affordance maps. Meanwhile, our model can deal with various challenging situations like unseen categories or domains. Both suggest that our model possesses robust open-world reasoning capabilities, which will significantly enhance subsequent object-grasping tasks.

6. More Results on Real Robot

Beyond the evaluation tasks in our main manuscript, such as grasping can, pen, screwdriver, hammer, and wok, we also evaluate the open-world generalization capabilities of our model by utilizing a broader range of instructions encompassing various unseen categories in real-robot environments, like panda, toy, circle and so on. The real-robot experiment videos are included within the same directory. These results demonstrate impressive open-world object perception and grasping proficiency.

7. More Results on RLBench

We present several visualization results from the simulation task RLBench in Figure 1. The top video demonstrates the task "open the top drawer," while the bottom video illustrates "close the green jar". As shown, our model successfully completes both tasks with high accuracy.

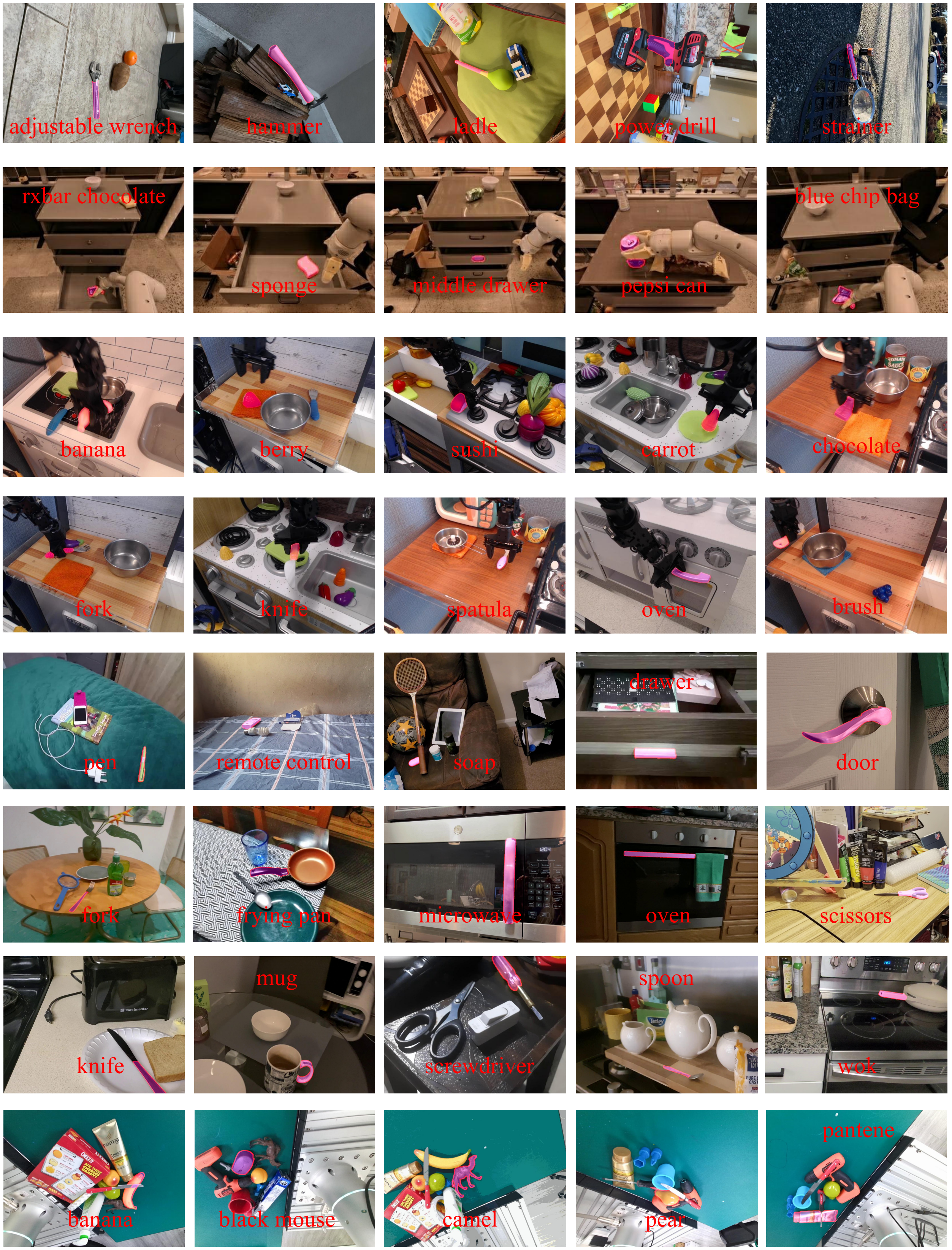
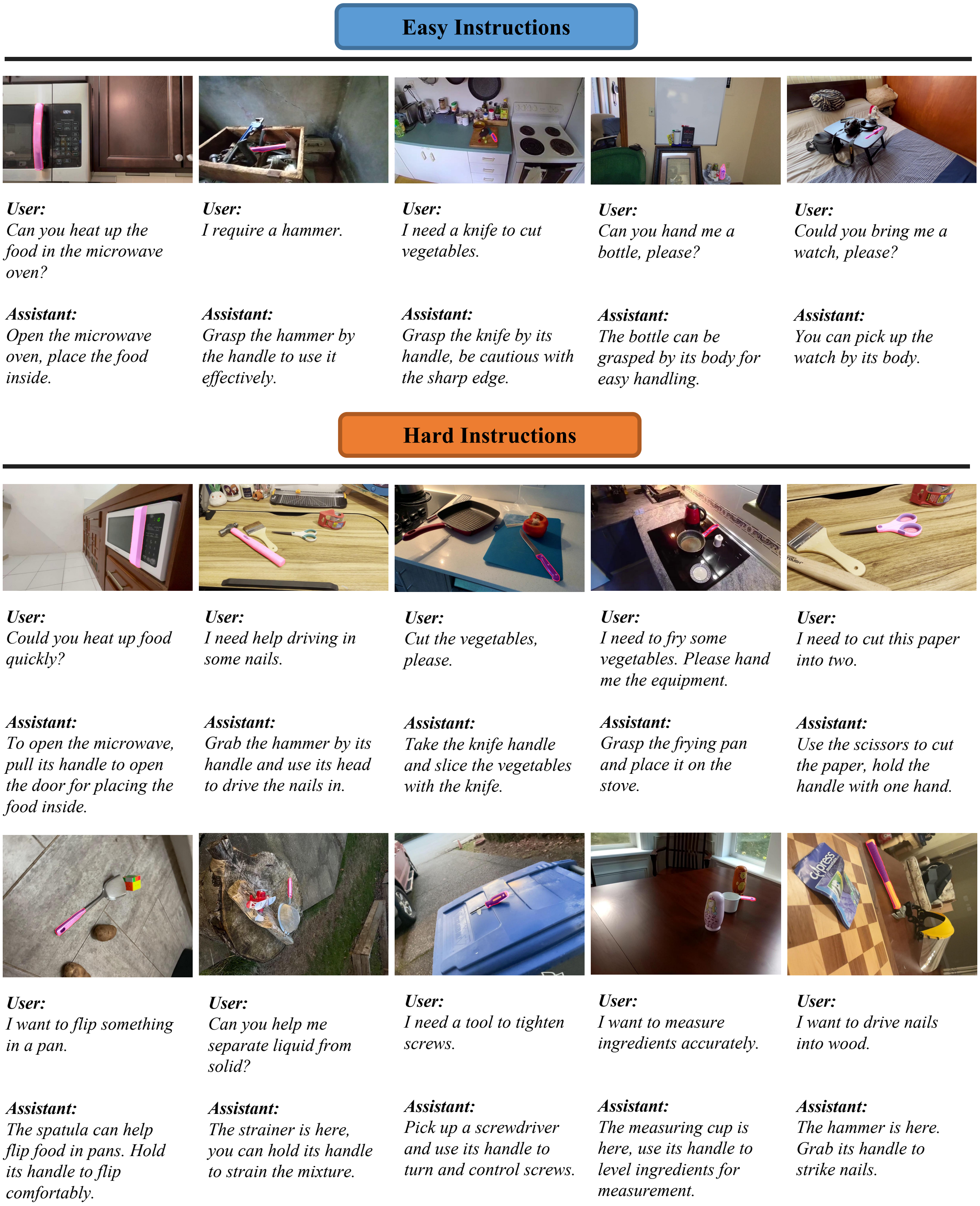


Figure 2. **More affordance segmentation examples from our RAGNet.** It covers various data sources, like robot, wild, and ego-centric domains.



User: *Can you heat up the food in the microwave oven?*

Assistant: *Open the microwave oven, place the food inside.*

User: *I require a hammer.*

Assistant: *Grasp the hammer by the handle to use it effectively.*

User: *I need a knife to cut vegetables.*

Assistant: *Grasp the knife by its handle, be cautious with the sharp edge.*

User: *Can you hand me a bottle, please?*

Assistant: *The bottle can be grasped by its body for easy handling.*

User: *Could you bring me a watch, please?*

Assistant: *You can pick up the watch by its body.*

User: *Could you heat up food quickly?*

Assistant: *To open the microwave, pull its handle to open the door for placing the food inside.*

User: *I need help driving in some nails.*

Assistant: *Grab the hammer by its handle and use its head to drive the nails in.*

User: *Cut the vegetables, please.*

Assistant: *Take the knife handle and slice the vegetables with the knife.*

User: *I need to fry some vegetables. Please hand me the equipment.*

Assistant: *Grasp the frying pan and place it on the stove.*

User: *I need to cut this paper into two.*

Assistant: *Use the scissors to cut the paper, hold the handle with one hand.*

User: *I want to flip something in a pan.*

Assistant: *The spatula can help flip food in pans. Hold its handle to flip comfortably.*

User: *Can you help me separate liquid from solid?*

Assistant: *The strainer is here, you can hold its handle to strain the mixture.*

User: *I need a tool to tighten screws.*

Assistant: *Pick up a screwdriver and use its handle to turn and control screws.*

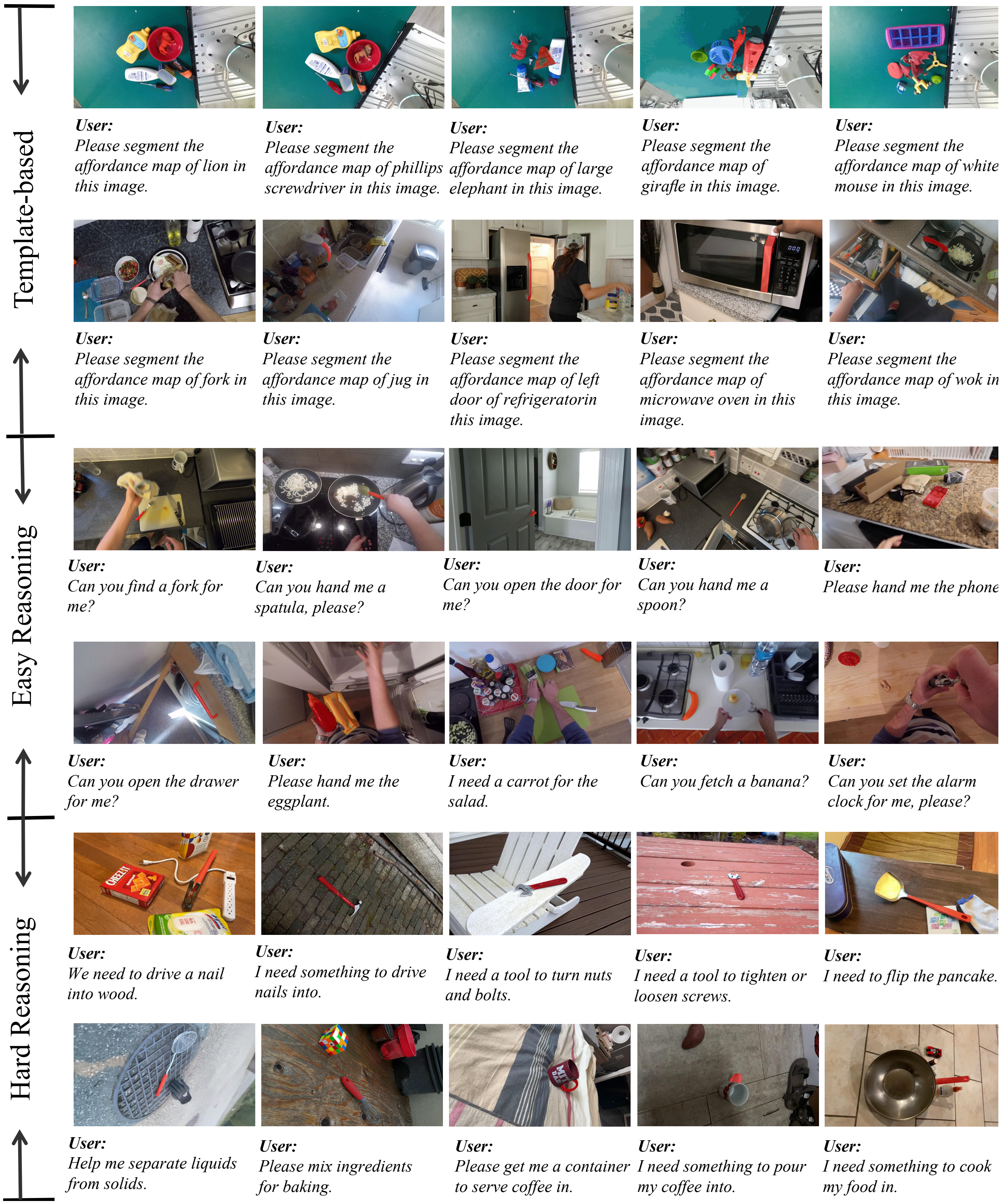
User: *I want to measure ingredients accurately.*

Assistant: *The measuring cup is here, use its handle to level ingredients for measurement.*

User: *I want to drive nails into wood.*

Assistant: *The hammer is here. Grab its handle to strike nails.*

Figure 3. **More reasoning-based affordance segmentation examples from our RAGNet.** It includes two types of reasoning instructions: easy instructions and hard instructions.



User: *Please segment the affordance map of lion in this image.*

User: *Please segment the affordance map of phillips screwdriver in this image.*

User: *Please segment the affordance map of large elephant in this image.*

User: *Please segment the affordance map of girafle in this image.*

User: *Please segment the affordance map of white mouse in this image.*

User: *Please segment the affordance map of fork in this image.*

User: *Please segment the affordance map of jug in this image.*

User: *Please segment the affordance map of left door of refrigeratorin this image.*

User: *Please segment the affordance map of microwave oven in this image.*

User: *Please segment the affordance map of wok in this image.*

User: *Can you find a fork for me?*

User: *Can you hand me a spatula, please?*

User: *Can you open the door for me?*

User: *Can you hand me a spoon?*

User: *Please hand me the phone*

User: *Can you open the drawer for me?*

User: *Please hand me the eggplant.*

User: *I need a carrot for the salad.*

User: *Can you fetch a banana?*

User: *Can you set the alarm clock for me, please?*

User: *We need to drive a nail into wood.*

User: *I need something to drive nails into.*

User: *I need a tool to turn nuts and bolts.*

User: *I need a tool to tighten or loosen screws.*

User: *I need to flip the pancake.*

User: *Help me separate liquids from solids.*

User: *Please mix ingredients for baking.*

User: *Please get me a container to serve coffee in.*

User: *I need something to pour my coffee into.*

User: *I need something to cook my food in.*

Figure 4. **More experiment results from our model.** We use template-based, easy reasoning-based, and hard reasoning-based instructions, respectively.

Dataset	Subset	Annotation Tool and Categories
HANDAL	-	❶: strainer, fixed joint plier, hammer, ladle, whisk, measuring cup, locking plier, power drill, adjustable wrencher, mug, ratchet, utensil, combinational wrench, pots pan, spatula, screwdriver, slip joint plier
Open-X	RT-1	❸: redbull can, rxbar blueberry, green can, apple, orange can, 7up can, sponge, pepsi can, orange, paper bowl, green rice chip bag, banana, coke can, blue chip bag, water bottle, white bowl, rxbar chocolate, 7up can, brown chip bag, blue plastic bottle, green jalapeno chip bag, blue water bottle, ❺: right fridge door, bottom drawer, left fridge door, middle drawer, top drawer
	Bridge	❸: apple, apple slice, avocado, ball, banana, banana plush, baster, beet, beetroot, bell pepper, berry, blackberry, board, book, bot, bottle, bowel, bowl, bread, bread roll, broccoli, bunny, butter, cake, cake slice, can, cap, capsicum, carrot, cereal, cheese, cheese slice, cheese wedge, cherry, cake, cake slice, can, cap, capsicum, carrot, cauliflower, cereal, cheese slice, cherry, chicken drumstick, chicken leg, chicken piece, chili pepper, chocolate, croissant, cucumber, detergent, dishcloth, doll, dough, drumstick, egg, eggplant, eggroll, garlic, half bun, hot dog, hotdog, lime, lobster tail, mango, meat, mouse, plastic fish, plush animal, sausage, soap, stuffed animal, stuffed dog, stuffed mushroom, stuffed cheetah, stuffed duck, stuffed pig, strawberry, sushi, tube, turkey leg, yam ❹: knife ❺: brush, cutter, drawer of box, fork, gripper, hairbrush, ice cream scoop, kettle, laddle, microwave, mug, oven, pot, pan, saucepan, scissors, scrub brush, scrubber, spatula, spork, teapot, teal brush, wok
EgoObjects	-	❷: alarm clock, balloon, blanket, book, bottle, bowl, box, computer mouse, doll, envelope, eraser, flowerpot, flying disc, football, game controller/pad, glasses, glove, goggles, lipstick, necklace, paper, paper towel, pen, pencil, pencil case, perfume, phone charger, picture frame, pillow, plate, post-it, poster, pottery, remote control, ring, shirt, shorts, skateboard, soap, sock, stapler, sun hat, sunglasses, tablet computer, teddy bear, tennis ball, toothpaste, towel, umbrella, vase, wallet, watch ❹: spoon, mug, screwdriver, knife, wrench ❺: microwave oven, washing machine, wok, oven, drawer, teapot, toothbrush, wardrobe, door, jug, refrigerator, tap, tennis racket, spatula, fork, frying pan, scissors, hammer
GraspNet	-	❶: dish, cracker box, pear, camel, peach, tape, banana, head shoulders care, black mouse, tomato soup can, darlie toothpaste, rhinocero, baoke marker, hosjam, pantene, racquetball, cups, sum37 secret repair, gorilla, kispa cleanser, hippo, toy airplane, dabao wash soup, weiquan, strawberry, dabao facewash, head shoulders supreme, dabao sod, large elephant, darlie box, nzskincare mouth rinse, plum ❺: flat screwdriver, power drill, scissors, mug
HANDAL *Reasoning*	-	**Hard Instructions**: strainer, fixed joint plier, hammer, ladle, whisk, measuring cup, locking plier, power drill, adjustable wrencher, mug, ratchet, utensil, combinational wrench, pots pan, spatula, screwdriver, slip joint plier
EgoObjects *Reasoning*	-	**Easy Instructions**: alarm clock, balloon, blanket, book, bottle, bowl, box, computer mouse, doll, envelope, eraser, flowerpot, flying disc, football, game controller/pad, glasses, glove, goggles, lipstick, necklace, paper, paper towel, pen, pencil, pencil case, perfume, phone charger, picture frame, pillow, plate, post-it, poster, pottery, remote control, ring, shirt, shorts, skateboard, soap, sock, stapler, sun hat, sunglasses, tablet computer, teddy bear, tennis ball, toothpaste, towel, umbrella, vase, wallet, watch, spoon, mug, screwdriver, knife, wrench, microwave oven, washing machine, wok, oven, drawer, teapot, toothbrush, wardrobe, door, jug, refrigerator, tap, tennis racket, spatula, fork, frying pan, scissors, hammer **Hard Instructions**: spoon, mug, screwdriver, knife, wrench, microwave oven, washing machine, wok, oven, drawer, teapot, toothbrush, wardrobe, jug, refrigerator, tap, tennis racket, spatula, fork, frying pan, scissors, hammer

Table 2. **Annotation details of RAGNet on tool composition** for different subsets and categories.

messages = [‘role’: ‘system’, ‘content’: ‘You are a helpful assistant. Based on several words where the first is category name, please design an instruction < 1 > and instruction < 2 > in embodied scenes. The instruction < 1 > must include object category name itself. The instruction < 2 > must include object category name itself. The instruction < 2 > must belongs to embodied manipulation and give action if instruction < 1 > provides. The instruction < 2 >does not exceed 50 words.’,
‘role’: ‘user’, ‘content’: ‘mug’,
‘role’: ‘assistant’, ‘content’: ‘< 1 > I need a drink. Please find a mug to fill water. < 2 > The mug has a handle as affordance map. So the robot can hold its handle.’
‘role’: ‘user’, ‘content’: ‘knife’
‘role’: ‘assistant’, ‘content’: ‘< 1 > Please give me a knife to cut apple. < 2 > The knife has a handle, and you can use its handle to cut apple.’,
‘role’: ‘user’, ‘content’: ‘hammer’,
‘role’: ‘assistant’, ‘content’: ‘< 1 > What is the proper way to hold the hammer? < 2 > The correct method is to hold the hammer by its handle.’,
‘role’: ‘user’, ‘content’: ‘fork’,
‘role’: ‘assistant’, ‘content’: < 1 > Kindly pick up the fork. < 2 > You will be holding the fork handle.’,
‘role’: ‘user’, ‘content’: ‘screwdriver’,
‘role’: ‘assistant’, ‘content’: ‘< 1 > I need a tool to tighten or loosen screws. < 2 > The screwdriver is here, hold its handle to turn and control screws.’,
‘role’: ‘user’, ‘content’: ‘words’]

Table 3. Language prompt when **generating easy reasoning-based instructions on HANDAL via GPT-4.**

messages = [‘role’: ‘system’, ‘content’: ‘You are a helpful assistant. Based on several words where the first is category name, please design an instruction < 1 > and instruction < 2 > in embodied scenes. The instruction < 1 > must not include object category name itself. The instruction < 2 > must include object category name itself. The instruction < 2 > must belongs to embodied manipulation and give action if instruction < 1 > provides. The instruction < 2 >does not exceed 50 words.’,
‘role’: ‘user’, ‘content’: ‘microwave, open’,
‘role’: ‘assistant’, ‘content’: ‘< 1 > Heat up food quickly . < 2 > The microwave is closed, so it can be open to access the food inside.’
‘role’: ‘user’, ‘content’: ‘knife’
‘role’: ‘assistant’, ‘content’: ‘< 1 > I want to cut a bread. < 2 > The knife has a handle, you can use its handle to cut bread.’,
‘role’: ‘user’, ‘content’: ‘computer mouse’,
‘role’: ‘assistant’, ‘content’: ‘< 1 > Give me a tool to control the cursor on the screen. < 2 > The computer mouse is here. It has not handle, so you can grasp its whole body.’,
‘role’: ‘user’, ‘content’: ‘fork’,
‘role’: ‘assistant’, ‘content’: < 1 > Use to pierce and lift food. < 2 > The fork is here, and its handle can be grasped.’,
‘role’: ‘user’, ‘content’: ‘screwdriver’,
‘role’: ‘assistant’, ‘content’: ‘< 1 > I need a tool to tighten or loosen screws. < 2 > The screwdriver is here, hold its handle to turn and control screws.’,
‘role’: ‘user’, ‘content’: ‘words’]

Table 4. Language prompt when **generating hard reasoning-based instructions on HANDAL via GPT-4.**

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 1

[2] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *CVPR*, 2020. 1

[3] Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. HANDAL: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *IROS*, 2023. 1

[4] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 1

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 1

[6] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 1

[7] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *CVPR*, 2023. 1

[8] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[9] Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *ICCV*, 2023. 1

[10] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024. 1

[11] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 1

[13] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *ICCV*, 2023. 1