

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

Dongming Wu¹, Yanping Fu^{2*}, Saike Huang³, Yingfei Liu^{3†}, Fan Jia³, Nian Liu⁴, Feng Dai², Tiancai Wang³, Rao Muhammad Anwer⁴, Fahad Shahbaz Khan⁴, 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

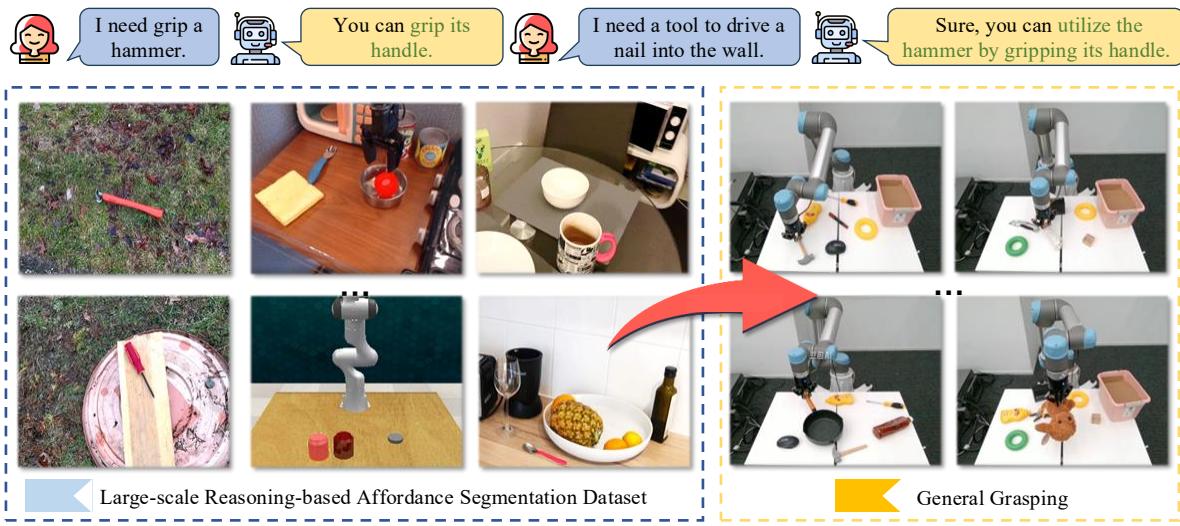


Figure 1. **Large-scale benchmark for reasoning-based affordance segmentation**, which sources from various embodied domains: wild, robot, ego-centric indoor, and simulation. By leveraging the extensive high-quality dataset for training, our model *AffordanceNet* exhibits remarkable open-world generalization capabilities, steering further towards robust general-purpose object grasping.

Abstract

General robotic grasping systems require accurate object affordance perception in diverse open-world scenarios following human instructions. However, current studies suffer from the problem of lacking reasoning-based large-scale affordance prediction data, leading to considerable concern about open-world effectiveness. To address this limitation, we build a large-scale grasping-oriented affordance segmentation benchmark with human-like instructions, named *RAGNet*. It contains 273k images, 180 categories, and 26k reasoning instructions. The images cover diverse embodied data domains, such as wild, robot, ego-centric, and even

simulation data. They are carefully annotated with an affordance map, while the difficulty of language instructions is largely increased by removing their category name and only providing functional descriptions. Furthermore, we propose a comprehensive affordance-based grasping framework, named *AffordanceNet*, which consists of a VLM pre-trained on our massive affordance data and a grasping network that conditions an affordance map to grasp the target. Extensive experiments on affordance segmentation benchmarks and real-robot manipulation tasks show that our model has a powerful open-world generalization ability. Our data and code is available at [this link](#).

1. Introduction

Affordance prediction is a foundational research topic that significantly contributes to diverse practical applications, including robotic manipulation [16, 18, 19, 39, 53] and human-object interaction [3, 21, 23, 52, 58, 59]. It requires

* The work is done during the internship at Dexmal. † Project lead.

‡ Corresponding author. This work was supported in part by the Science and Technology Development Fund of Macau SAR (FDCT) under grants 0102/2023/RIA2 and 0154/2022/A3 and 001/2024/SKL, the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC), and the University of Macau SRG2022-00023-IOTSC grant.

Dataset	Images	Categories	Wild	Robot	Ego-centric	Simulation	Reasoning	Output
UMD [38] ICRA2015	10k	17	✓				✓	Seg
AGD20k [36] CVPR2020	20k	50	✓					Seg
HANDAL [15] IROS2023	200k	17	✓					Seg/Box
AED [29] Arxiv2024	-	21		✓		✓		Seg
3DOI [43] ICCV2023	10k	-	✓		✓			Seg
AffordanceLLM [45] CVPR2024	20k	-	✓					Seg
ManipVQA [17] IROS2024	84k	-	✓		✓		✓	Seg
RAGNet (Ours)	273k	180	✓	✓	✓	✓	✓	Seg

Table 1. Comparisons between previous affordance data and our collection. “(11k)” represents the number of video clips. “-” means unavailable data. “Reasoning” refers to reasoning instructions.

comprehensively understanding the geometry and function of an object (*e.g.*, a wok affords to hold, a microwave door affords to pull) for further detecting graspable regions. Nonetheless, this task faces two additional primary difficulties in the realm of open-world generalization concerning vision and language instructions: 1) The capability to generalize in a broad range of unseen object categories and image domains. 2) The alignment of commands that mimic human-like high-level instructions.

Initial studies often specialize in particular domains, as depicted in Table 1. For instance, UMD [38] offers 10,000 RGB-D image pairs from three cluttered scenes with pixel-level affordance annotations. This dataset is categorized as *robot data*, characterized by a distinct operational table and a relatively stable background. Despite the abundance of public robot datasets [4, 42, 57], they often lack meticulous fine-grained labeling. *Ego-centric data*, which is captured from a first-person or egocentric viewpoint, is the most prevalent type of data [8, 14]. Nevertheless, this data is often collected in indoor kitchens and workshops, resulting in a lack of diversity in the data category distribution. *Wild data*, another prevalent type, is gathered in various situations, which usually cover diverse categories [36, 47, 54]. However, certain categories within this data, like bicycles, motorcycles, sofas and *etc.*, are not suitable for robotic manipulation. In addition, these data often fail to generalize to unseen domains and novel objects for affordance prediction.

To enhance open-world generalization, Vision Language Model (VLM) with massive image-text training [30, 33] has become an important increment in various visual prediction tasks [26, 48]. Inspired by this, VLM also has an increasing interest in affordance prediction along with complex reasoning, and several recent works are attempting to predict affordance regions [17, 28, 31, 45, 61, 64]. For example, ManipVQA [17] proposes tool detection, affordance recognition, and a broader understanding of physical concepts in a unified framework. However, these works employ a fixed format of language prompt and a limited data scale, which brings some concerns about open-world generalization and complex reasoning. Overall, current efforts are either constrained by insufficiently diverse datasets or by the absence of effective reasoning mechanisms, which hampers the ability to perform open-world grasping.

To address these challenges, we first build large-scale affordance segmentation data from various image sources, including wild, robot, ego-centric, and even simulated data, named **RAGNet**. It has 273k images and 180 categories. We design a set of affordance annotation tools for labeling the regions of objects that are amenable to grasping. Furthermore, we leverage Large Language Models (LLMs) to generate a vast array of reasoning-based instructions, totaling 26k distinct expressions. Here, we create two types of reasoning-based instructions beyond the template-based. One includes the name of the object and the other omits it. Take a knife as an example: “Please provide a knife” versus “I want something to slice the bread”. This approach closely mirrors real-life human interactions. In summary, we have constructed a massive-scale database with domains, object categories, and complex reasoning instructions.

Furthermore, we introduce an affordance-based grasping framework, titled **AffordanceNet**. It presents a deployable and general grasping pipeline, which stands out from prior MLLM-based affordance prediction methods [17, 45] that have not yet demonstrated real-robot deployment. Our model includes two crucial components, AffordanceVLM and Pose Generator. The AffordanceVLM transforms RGB images and human instruction into an accurate affordance map, while the pose generator uses the 2D affordance with a depth image to produce 3D grasper pose. In this work, we conduct extensive experiments to assess its open-world generalization and reasoning capabilities. First, we create two distinct validation datasets to assess open-world generalization: one for zero-shot category recognition and another for out-of-domain affordance prediction. Second, to test reasoning ability, we implement an affordance segmentation validation based on instructions that do not contain any target category names. Thirdly, we carry out a range of close-loop real-robot grasping tasks in an entirely out-of-domain setting. Last but not least, we test several representative simulation tasks from RLBench. All experiments show that our proposed method has great generalization and reasoning ability. In summary, our contributions are three-fold:

- We present a large-scale reasoning-based affordance segmentation benchmark, RAGNet, for general grasping. It is collected from diverse sources and carefully annotated with affordance mask and reasoning instructions.

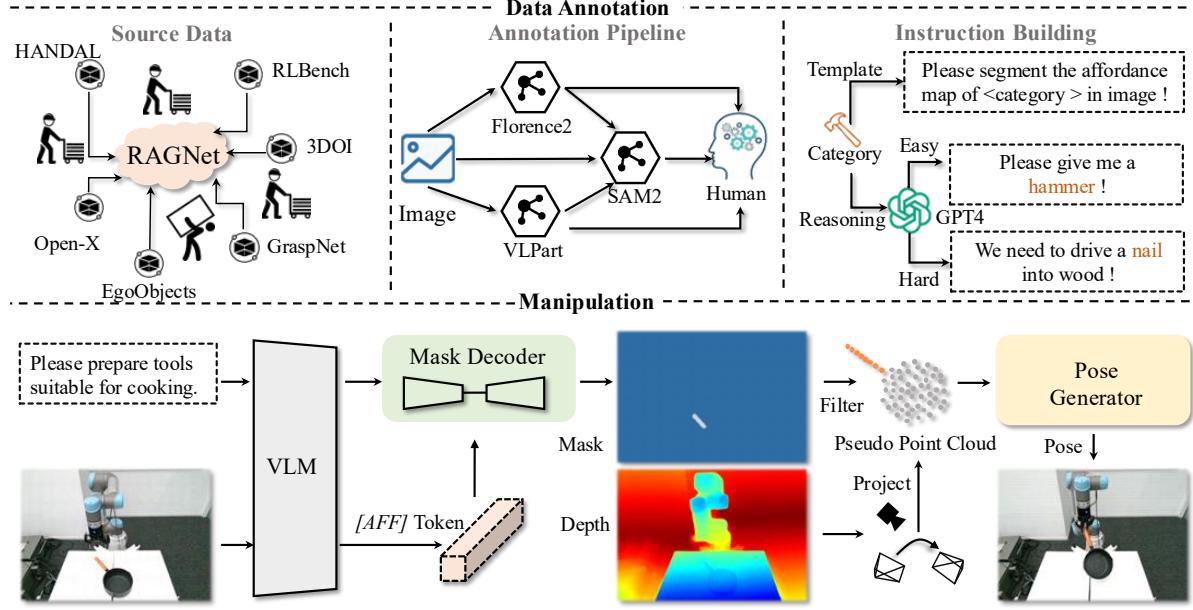


Figure 2. **Overview of our data annotation pipeline and manipulation model.** We collect data from several public datasets, including HANDAL, GraspNet, Open-X, etc. A variety of models and manual annotation are utilized to annotate affordance masks. Subsequently, we refine textual instructions using templates and GPT4 to emulate human-like commands. For grasp operations, the VLM model is employed to identify affordance regions, which are then converted into the required grasp poses by integrating depth information.

- We introduce an affordance-based grasping baseline, AffordanceNet, which bridges the gap between VLM-based affordance prediction and real-robot general grasping.
- We conduct extensive experiments, including zero-shot and out-of-domain affordance segmentation, real-robot grasping evaluation, which show great performance.

2. Related Work

The investigation of affordances has a rich history in the computer vision and robotics communities [1, 12, 13, 25]. For embodied agents to successfully interact with the functional components within a scene, they need the capability to comprehend visual affordances. To advance research in this area, previous studies have dedicated significant resources to two main areas: benchmarks and algorithms.

Benchmark. With the advances in deep learning, there has been a growing demand for the creation of large-scale affordance databases. UMD [38] stands out as a prior effort, which defines objects with effective affordances as those that an agent can grasp to produce an effect on another object. In addition, it offers a dataset of 10k RGB-D images, each accompanied by a pixel-level affordance segmentation mask. OPRA [11] proposes a different affordance learning pattern, which employs demonstration videos to guide the prediction of interaction region on a target image. It includes 20,612 video clips, each corresponding to an image and an annotation of interaction heatmap and action label. Despite the pioneering efforts in affordance learning,

the existing datasets [7, 11, 35, 38, 40, 51] still face constraints regarding affordance category diversity, image quality, and scene complexity. Afterward, Luo *et al.* establish a large-scale affordance grounding dataset, AGD20k [36], which contains 20k exocentric images from 36 affordance categories. Guo *et al.* build the HANDANL dataset [15] for real-world robot manipulation, which provides precise handle annotation for 17 hardware and kitchen tool categories. Recently, there are a variety of datasets on affordance prediction that have been introduced with different output formats, such as key points [22, 31, 43, 49], bounding box [55, 56], pixel-wise mask [28, 29, 32]. However, most current research tends to focus on specific domains, such as robotics or first-person perspectives, which limits their applicability to other areas. In contrast, our research aims to investigate the potential for open-world generalization in affordance prediction by leveraging extensive data.

Algorithm. In the deep learning era, the common methods use supervised learning for affordance prediction [7, 9, 15, 35, 40]. However, these approaches struggle to generalize to domains outside of their training environments. To tackle this issue, certain studies utilize transfer learning [27, 36, 37, 62] or self-supervised learning techniques [6] to enable affordance prediction from data originating in different domains. With the significant progress of VLM, attempting these foundation models in affordance prediction has attracted much interest in the research community [17, 28, 31, 45, 61, 64]. However, these methods

Data source	Domain	Annotation	Rea. Inst.	Categories
HANDAL [15]	Wild	①	8.5k	17
Open-X [42]	Robot	②④⑤	-	124
GraspNet [10]	Robot	①⑤	-	32
EgoObjects [68]	Ego	②④⑤	17.4k	74
RLBench [20]	Simulation	⑥	-	10

Table 2. **Details of training data annotation in RAGNet.** The affordance mask annotation are in §3.2 and the generation details of reasoning instructions (Rea. Inst.) can be found in §3.3.

usually learn from limited affordance demonstration data, which opens up opportunities for further research into how large-scale data can impact generalization in open-world scenarios and enhance knowledge reasoning.

3. Dataset

In this section, we will present how to build the large-scale reasoning-based affordance segmentation benchmark RAGNet. First, we gather extensive data from various sources, as detailed in §3.1. Then, we introduce five tools to annotate this data with grasping-oriented affordance masks according to the original dataset characteristics in §3.2. Meanwhile, we offer detailed instructions for reasoning-based affordance segmentation in §3.3. Lastly, we establish multiple validation benchmarks for testing grasping-oriented affordance segmentation in §3.4. The data annotation pipeline is illustrated in Fig. 2.

3.1. Data Source

Given that robotics can function across various embodiment realms, including real-world settings and robotic hardware, we collect a broad spectrum of data. The dataset comprises sources such as wild data (*i.e.*, HANDAL [15]), real robots (*i.e.*, Open-X [42], GraspNet [10]), and ego-centric data (*i.e.*, EgoObjects [68]). For a fair comparison environment, we consider a collection of simulation data (*i.e.*, RLBench [20]). We gather 273k images as shown in Table 2.

3.2. Affordance Map Annotation

To facilitate affordance segmentation with reduced manual input, we develop a suite of annotation tools. The annotation workflow is designed to be adaptable, taking into account the functional and physical characteristics of various object categories. For instance, when dealing with an object like a soda can, the robotic system typically grasps the entire object, necessitating full object annotation. On the other hand, an object such as a wok requires precise annotation of its handle for effective grasping. This tool suite comprises five annotation tools, each with its own level of priority.

① *Original mask*: Certain datasets specify affordance segmentation with precision (such as HANDAL dataset [15]), while others, like those featuring objects without handles (*e.g.*, computer mouse, soda can, and pen), do not require



User:
Can you find a *mug* for tea?

Assistant:
The *mug* has a *handle* which you can hold to pour the coffee. [AFF]

User:
I need something to drink coffee.

Assistant:
The *mug* is suitable for this, its *handle* can be held to drink coffee safely. [AFF]

Figure 3. **Easy (left) v.s. Hard (right) Reasoning Instruction.** The hard version has no category name itself.

detailed affordance masks for grasping. In these cases, we utilize the original masks as our affordance annotations without further refinement.

② *SAM2*: For objects that lack handles, SAM2 [50] can still be utilized to generate a mask, when only the ground-truth bounding box is available (like EgoObjects [68]).

③ *Florence2 + SAM2*: Due to the presence of language instructions (like Open-X [42]), Florence2 [60] which generates polygon boxes coupled with SAM2 can be used to produce an affordance map for objects that are handle-free.

④ *VLPart + SAM2*: VLPart [54] enables part-level recognition (*i.e.*, knife handle and mug handle), hence we leverage it and SAM2 for segmenting object handles if the corresponding category has been trained within VLPart.

⑤ *Human (+ SAM2)*: If the above four tools fail to complete the affordance segmentation accurately, we will consider manual affordance annotation. The use of SAM2 is optional, particularly when dealing with video sequences.

According to the original annotation information offered by data sources, we employ different compositions for affordance mask annotation. The composition details are listed in Table 2. We incorporate more detailed annotations, encompassing variations in tool arrangement across subsets and categories, within the supplementary. Besides, we provide several representative examples in Fig. 1 and more annotation examples can also be found in the supplementary.

3.3. Reasoning Instruction Annotation

As previously mentioned, the current VLM possesses compelling reasoning capabilities. To harness these capabilities for affordance reasoning, we construct a set of instructions. In this section, we introduce three types of instructions, including one template-based and two reasoning-based.

The first kind of instruction is *template-based*. For instance, a template is “Please segment the affordance map of <category_name> in this image”. This template can

Validation Set	Images	Zero-shot	Reason.	Anno.
HANDAL [15]	65k	-	N/A	1
HANDAL [†] [15]	1k	-	N/A	1
GraspNet seen [10]	1k	-	N/A	15
GraspNet novel [10]	1k	✓	N/A	15
3DOI [43]	1k	✓	N/A	5
HANDAL [†] [15]	1k	-	easy	1
HANDAL [†] [15]	1k	-	hard	1
3DOI [43]	1k	✓	easy	5

Table 3. **Details of validation set.** [†] means a HANDAL subset. Gray means zero-shot setting. As highly replicated images, we randomly select 1k images from the source dataset for validation.

be utilized in our entire dataset for affordance prediction. The second category consists of *easy reasoning-based* instructions that are based on straightforward reasoning. A key aspect of these instructions is the explicit mention of the object being referred to. In contrast to template-based approaches and previous studies [17, 26, 45], the third category comprises *hard reasoning-based* instructions, which do not include the category name. For precise identification of the target object, the hard mode instructions utilize a functional description. Fig. 3 presents a typical example. Towards grasping a mug, the easy instruction might be “Can you find a mug for tea”, while the hard instruction is “I need something to drink coffee”.

To produce these reasoning-based instructions with minimal human labor and computational resources, we make full use of the capabilities of GPT-4 [2]. The specific prompt utilized is detailed in the supplementary material. According to our data, we craft 8.5k hard instructions for the HANDAL dataset, 12.7k easy ones and 4.7k hard ones for the EgoObjects, totaling 26k reasoning-based instructions.

3.4. Evaluation Dataset

To assess the open-world generalization of our affordance prediction model, we contribute two distinct zero-shot affordance evaluation scenarios. The first scenario evaluates the affordance model by extending its predictions to object categories not encountered during training. The second scenario evaluates the model generalization ability across different data domains. For this purpose, we select the 3DOI dataset [43], which includes data from Articulation [44], EpicKitchen [8], and Taskonomy [65], ensuring that there is no overlap with the training data in the validation set.

We present four different validation sets in Table 3. Due to high-similarity images, we construct a subset from the HANDAL, which is marked with ‘[†]’. For the same reason, other validation sets also utilize 1k images from the source dataset. HANDAL and GraspNet seen represent the object categories and image domains that the model has been trained on. GraspNet novel represents the unseen object category, while 3DOI refers to the unseen image domain.

In addition, we provide three reasoning-based affordance segmentation validation sets, as illustrated in Table 3. The reasoning instruction version used in each subset is marked as “easy” or “hard”. We utilize the generalized Intersection over Union (gIoU) and complete Intersection over Union (cIoU) as our primary metrics.

4. AffordanceNet

To achieve the goal of open-world grasping, we propose a comprehensive framework, named AffordanceNet. This model consists of two key components: AffordanceVLM for predicting affordance segmentation mask and pose generation for transforming the mask into grasper position in 3D space. The overall framework is illustrated in Fig. 2.

4.1. AffordanceVLM

Our AffordanceVLM is based on the vision-language segmentation model LiSA [26] and incorporates two essential task-specific modifications to enhance affordance prediction: (1) developing a specialized system prompt, and (2) introducing a unique <AFF> token.

Specifically, we first process the input image using an image encoder (*i.e.*, ViT-CLIP [46]), which is then projected into the LLM’s embedding space via a linear layer projector. Meanwhile, the language prompt is tokenized by a text tokenizer, where each affordance instruction is enhanced by “You are an embodied robot.” The resulting image and text features are concatenated and fed into the LLM (*i.e.*, Vicuna-7B [66]). For general segmentation, the LLM generates a response that includes a special token <SEG>. However, since <SEG> is a token within the LLM’s vocabulary, its representation is confined to a fixed feature space. This limitation restricts its representation capacity, thereby affecting the quality of the decoded mask. To address this issue, we introduce another special token <AFF> to enrich the original mask embedding. The underlying motivation is to explicitly direct the final mask embedding to focus more on affordance-specific language expressions. Finally, SAM [24] is used as a mask decoder to convert this mask embedding into a pixel-wise mask.

Implementation Details. Beyond our reasoning-based affordance segmentation data, we also incorporate a variety of generic segmentation datasets into our training, which use <SEG> token. During inference, we first extract the <AFF> token, followed by the <SEG> token. More implementation details about data sampling and training settings can be found in the supplementary.

4.2. Pose Generator

Once obtained AffordanceVLM that delivers precise affordance segmentation prediction, we start to apply the model to robotic grasping tasks, aiming to bridge the gap for the

Method	HANDAL		HANDAL [†]		GraspNet seen		GraspNet novel		3DOI	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
<i>Foundation Model without LLMs</i>										
VLPart [54] + SAM2 [50]	40.9	28.9	40.7	27.6	-	-	-	-	-	-
Grounding DINO [34] + SAM2 [50]	34.7	26.8	34.9	26.9	-	-	-	-	-	-
Florence 2 [60] + SAM2 [50]	39.7	22.4	39.4	22.5	-	-	-	-	-	-
<i>Generalist MLLMs</i>										
LISA [26]	16.2	12.0	15.4	11.8	17.7	17.7	25.2	24.1	21.5	13.7
GLaMM [48]	24.9	17.2	25.1	17.0	21.6	10.5	19.2	8.6	19.7	14.1
AffordanceNet (Ours)	60.3	60.8	60.5	60.3	63.3	64.0	45.6	33.2	37.4	37.4

Table 4. **Quantitative results on affordance segmentation.** We use the fixed format of “<category_name> handle” in foundation models without LLMs, while utilize “affordance map of <category_name>” in generalist MLLMs. Gray means zero-shot benchmark.

Method	HANDAL [†]		Grasp. novel		3DOI	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
LISA	16.2	12.0	25.2	24.1	21.5	13.7
+ HANDAL	56.3	54.9	16.6	14.4	18.0	12.2
+ Open-X	59.3	56.7	19.2	18.4	24.5	16.0
+ EgoObjects	61.8	61.6	8.0	6.9	35.5	34.6
+ GraspNet	61.7	61.7	51.5	38.5	40.9	41.8
+ Reasoning	56.5	55.4	43.0	33.8	36.8	40.2
+ RLBench	56.7	55.0	42.8	33.2	36.5	39.9
Ours	60.5	60.3	45.6	33.2	37.4	37.4

Table 5. **Ablation study of data on affordance segmentation.** Each data is added one by one. Compared to “+ RLBench”, our final model is enhanced by task-specific modifications, including specialized system prompt and unique <AFF> token (see §4.1).

final steps in object grasping and manipulation. The impressive affordance segmentation results can be found in §5.1 and §5.2. In the following part, we will discuss more details about our grasp pose generator and the process of converting 2D affordance masks into the 3D pose estimates required for robotic arms.

As illustrated in Fig. 2, we project the depth maps into 3D space for precise grasp pose generation. Let P denotes the set of points on the 2D image. Firstly, we filter the affordance region by applying the affordance mask M to P through a binary multiplication \otimes , resulting in $\hat{P} = P \otimes M$. Subsequently, for each 2D position (u, v) within the point set \hat{P} , we have

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = T \cdot K^{-1} \begin{bmatrix} u \times d \\ v \times d \\ d \\ 1 \end{bmatrix}, \quad (1)$$

where d is the along the axis orthogonal to the image plane, (x, y, z) are the corresponding world coordinates. $K \in \mathbb{R}^{4 \times 4}$ and $T \in \mathbb{R}^{4 \times 4}$ represent the camera intrinsic and extrinsic parameters, respectively. After getting the 3D position of object affordance, we can use various grasping

Method	HANDAL (easy)		HANDAL (hard)		3DOI	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
G-DINO	3.6	3.0	3.4	3.1	4.1	3.9
LISA	15.5	11.9	12.3	8.1	12.3	8.1
GLaMM	4.7	3.5	5.0	3.5	4.4	2.9
Ours	58.3	58.1	58.2	57.8	38.1	39.4

Table 6. **Quantitative results on reasoning-based affordance segmentation.** ‘G-DINO’ is the short name of Grounding-DINO. They all use reasoning-based instructions as language prompts.

models to generate grasper position. Finally, the grasper arrive the requested position and grasp the target object.

5. Experiments on Visual Affordance

To minimize unnecessary source expenditure, we initially validate the quality of affordance segmentation (§4.1) prior to object grasping (§4.2).

5.1. Evaluation on Affordance Segmentation

Implementation Details. To evaluate the task of affordance segmentation, we implement various advanced open-sourced approaches with potential affordance segmentation ability. In specific, we choose the foundation models without LLMs (e.g., VLPart [54], Grounding DINO [34], Florence2 [60]), and the generalist MLLMs (e.g., LISA [26], GLaMM [48]). Since these foundation models only output bounding boxes or polygon boxes, we additionally employ SAM2 [50] for mask refinement. We load their official checkpoints for direct evaluation. Given the ambiguous nature of the “affordance map of <category_name>” concept, which poses a challenge to the foundation models without LLMs, we shift to a fixed format of instruction “<category_name> handle” and evaluate them on only HANDAL dataset as each object have a handle. In contrast, we employ “affordance map of <category_name>” when testing the generalist MLLMs.

Experiment Results. The main results on affordance segmentation are shown in Table 4. We can see our model out-



Figure 4. **Affordance segmentation from our AffordanceNet.** Even though they source from various data sources, such as wild, robot, ego-centric and simulation, our model can accurately capture their affordance region. More visualizations are included in supplementary.



Figure 5. **Reasoning-based affordance segmentation from our AffordanceNet.** The left examples represent easy reasoning-based instructions with referent name, while the right are hard instructions that include object function or intention rather than the name itself.

performs all other competitors across all datasets. Besides, the scores on the entire HANDAL dataset and its small subset are similar, guaranteeing the diversity and representative of the small subset. Several visualizations across various domains are shown in Fig. 4. Here, the wok handle is accurately segmented even if its scene has never been encountered, indicating remarkable open-world generalization.

Ablation Study. It is of interest to examine the impact of individual datasets on affordance segmentation tasks. In Table 5, we incrementally incorporate each dataset into the training process. We can see that the absence of HANDAL data significantly impairs the model performance on the HANDAL test set. Furthermore, the incorporation of reasoning data leads to a slight decline in performance metrics. However, the introduction of task-specific modifications, such as a specialized system prompt and a unique <AFF> token, enhances model performance. *Overall, our model shows powerful open-world affordance segmentation.*

5.2. Evaluation on Reasoning Affordance

We also test the foundation models and generalist MLLMs on reasoning-based affordance segmentation datasets. All the experiment settings are aligned with the above section.

Experiment Results. Despite this challenging task, the results in Table 6 show our model outperforms other methods by a large margin. We provide several reasoning-based qualitative results in Fig. 5. The last two examples show that our model can predict precise hammer and mug handles even if there is no target mentioned in the instructions. These results confirm the reasoning ability of our model.

Remark. The above two vision experiments demonstrate that only our model excels in both affordance perception and reasoning in zero-shot domain. This provides a solid foundation for the subsequent object grasping tasks.

6. Experiments on Object Grasping

6.1. Evaluation on Real Robot

Implementation Details. To evaluate the effectiveness of our model for open-world generalization in real-world environments, we introduce a series of manipulation experiments. Specifically, we deploy UR5 robot arm with a third-person RGB-D camera (Intel RealSense). We design 10 distinct grasping tasks, including grasping the can, pen, screwdriver, hammer, wok, mouse, circle, toy, spatula, scissors. Half of them require accurately localizing the affordance region, like the screwdriver handle. Each task is performed 10 times, and we report the average success rate. *Note that, we never provide any demonstration images or videos from this scene for our model training.* The well-trained AffordanceVLM is directly used for the zero-shot evaluation. As for the pose generation, we follow GraspNet [10] condition 3D affordance point cloud to generate a 3D grasp proposal for grasper operation. We compare our model with a popular grasping model, GraspNet [10]. Since GraspNet lacks the capability for language-conditioned grasping, *we ensure that only the target object remains on the table.*

Experiment Results. Table 7 displays the performance comparison between ours and GraspNet. Clearly, our model AffordanceNet has superior success rates, even in challenging and complicated environments. Fig. 6 highlights four sequences, which show accurate affordance perception and effective object grasping (*i.e.*, screwdriver, wok, circle, and mouse). More grasping examples are in our supplementary.

Ablation Study. Table 8 shows ablations of AffordanceNet on affordance prediction models (replaced by VLPart and LISA) and instruction types (replaced by easy and hard reasoning-based instructions) using five tasks. As demonstrated, AffordanceNet delivers superior affordance predic-

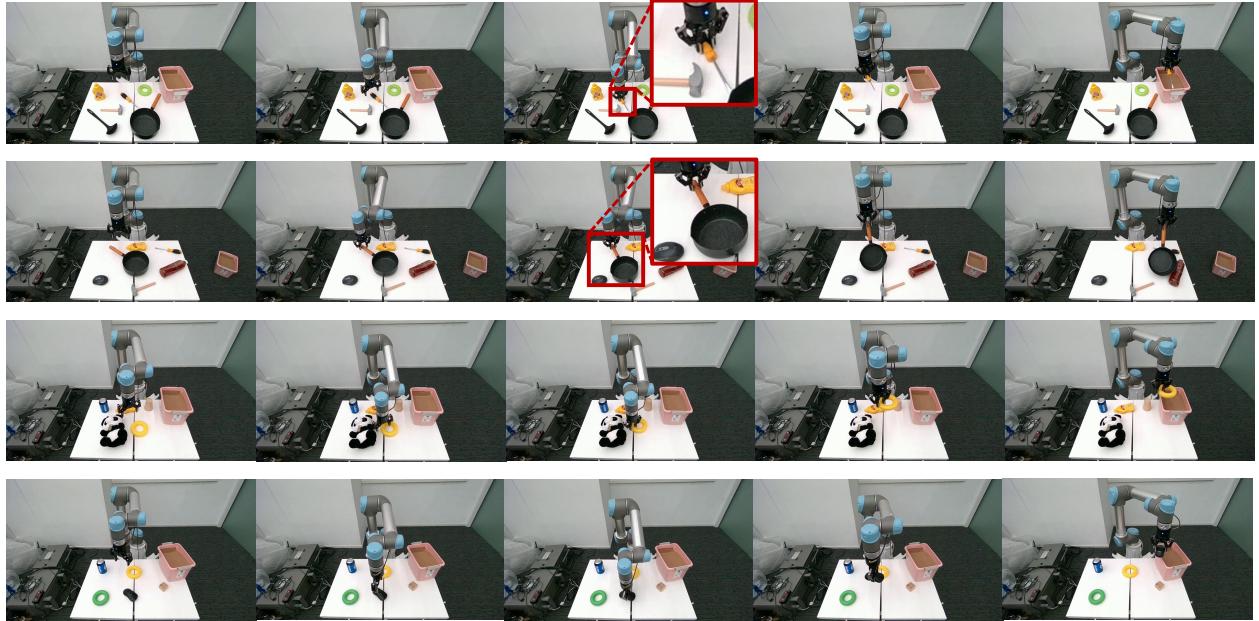


Figure 6. **Object grasping results from our AffordanceNet on robot arm UR5.** The instructions are “*I need a screwdriver for repairing*”, “*Can you hand me the wok, please?*”, “*Give me the circle*”, “*Please hand me a computer mouse*”, respectively.

Method	Can	Pen	Screwdriver	Hammer	Wok	Mouse	Circle	Toy	Spatula	Scissors	Average
GraspNet [10]	40%	10%	10%	20%	60%	50%	10%	60%	30%	30%	32%
AffordanceNet (Ours)	80%	60%	60%	80%	70%	80%	40%	90%	70%	70%	70%

Table 7. **Average success rates on robotic grasping.** GraspNet cannot support language (see §6.1). Each task is conducted by 10 times.

Method	Can	Pen	Screw.	Ham.	Wok	Average
VLPart [54]	70%	30%	40%	30%	0	34%
LISA [26]	80%	40%	0	10%	0	26%
Easy Reasoning	70%	50%	50%	80%	60%	62%
Hard Reasoning	60%	40%	40%	60%	40%	48%
AffordanceNet	80%	60%	60%	80%	70%	70%

Table 8. **Ablation studies on real-robot grasping** in terms of different affordance prediction models (*i.e.*, VLPart and LISA) and different instructions (*i.e.*, easy reasoning and hard reasoning).

tion while maintaining its reasoning capabilities.

6.2. Evaluation on Simulation

Implementation Details. We conduct simulations based on RLBench sub-task [20] to validate our approach, including open drawer, close jar, and slide block to target. In practice, we divide each task into several keyframes, as follows: open drawer (3 keyframes), close jar (4 keyframes), slide block to target (5 keyframes). More details are in the supplementary.

Experiment Results. We conduct 25 episodes for each task and calculate its average success rate as the primary evaluation metric. The quantitative results are in Table 9. Compared to another LLM-based method LLARVA [41] that is fine-tuned for a specific environment, our model focuses on

Task	LLARVA [41]	AffordanceNet (Ours)
Open drawer	60%	56%
Slide block to target	100%	64%
Close jar	28%	44%
Average	62%	54.7%

Table 9. **Success rates of our model on RLBench simulation.** Each task is conducted by 25 episodes.

stronger generalization. *Therefore, achieving comparable performance on RLBench is quite satisfying.*

7. Conclusion

In this work, we presented a new large-scale and diverse reasoning-based affordance segmentation dataset, named RAGNet, which aims to advance the capabilities of general robotic grasping systems in varied open-world scenarios. Further, we proposed the model AffordanceNet, a comprehensive framework including the AffordanceVLM and the grasping module, which uses our extensive affordance data to achieve open-world affordance capture and 3D grasper pose prediction, respectively. Through extensive experiments, we observed the superior performance and generalization ability of our AffordanceNet in zero-shot affordance segmentation, reasoning-based affordance segmentation, real-robot grasping evaluation, and simulation tasks.

References

- [1] Function-based generic recognition for multiple object categories. *CVGIP: Image Understanding*, 1994. 3
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5
- [3] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023. 1
- [4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2
- [5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 1
- [6] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *CVPR*, 2023. 3
- [7] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, 2018. 3
- [8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2022. 2, 5
- [9] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2018. 3
- [10] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *CVPR*, 2020. 4, 5, 7, 8, 1
- [11] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018. 3
- [12] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. *NeurIPS*, 2007. 3
- [13] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014. 3
- [14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2
- [15] 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. 2, 3, 4, 5, 1
- [16] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copo: General robotic manipulation through spatial constraints of parts with foundation models. *arXiv preprint arXiv:2403.08248*, 2024. 1
- [17] Siyuan Huang, Iaroslav Ponomarenko, Zhengkai Jiang, Xiaoli Li, Xiaobin Hu, Peng Gao, Hongsheng Li, and Hao Dong. Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models. In *IROS*, 2024. 2, 3, 5
- [18] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 1
- [19] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 1
- [20] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020. 4, 8
- [21] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *ICCV*, 2023. 1
- [22] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *ECCV*, 2024. 3
- [23] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 1
- [24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 5
- [25] Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 2011. 3
- [26] 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. 2, 5, 6, 8, 1
- [27] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *CVPR*, 2023. 3
- [28] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *CVPR*, 2024. 2, 3
- [29] Gen Li, Nikolaos Tsagkas, Jifei Song, Ruaridh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. Learning precise affordances from egocentric videos for robotic manipulation. *arXiv preprint arXiv:2408.10123*, 2024. 2, 3
- [30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2
- [31] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao

- Dong. Maniplmm: Embodied multimodal large language model for object-centric robotic manipulation. In *CVPR*, 2024. 2, 3
- [32] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *CVPR*, 2024. 3
- [33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 2, 1
- [34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6
- [35] Timo Ludecke and Florentin Worgotter. Learning to segment affordances. In *ICCV Workshops*, 2017. 3
- [36] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *CVPR*, 2022. 2, 3
- [37] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Leverage interactive affinity for affordance learning. In *CVPR*, 2023. 3
- [38] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, 2015. 2, 3
- [39] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. *arXiv preprint arXiv:2402.07872*, 2024. 1
- [40] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017. 3
- [41] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. *arXiv preprint arXiv:2406.11815*, 2024. 8
- [42] 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. 2, 4, 1
- [43] Shengyi Qian and David F Fouhey. Understanding 3d object interaction from a single image. In *ICCV*, 2023. 2, 3, 5
- [44] Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, and David F Fouhey. Understanding 3d object articulation in internet videos. In *CVPR*, 2022. 5
- [45] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *CVPR*, 2024. 2, 3, 5
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5
- [47] 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. 2, 1
- [48] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. 2, 6
- [49] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *CoRL*, 2023. 3
- [50] 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. 4, 6, 1
- [51] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *CVPR*, 2017. 3
- [52] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 1
- [53] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Clipor: What and where pathways for robotic manipulation. In *CoRL*, 2022. 1
- [54] 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. 2, 4, 6, 8, 1
- [55] An Dinh Vuong, Minh Nhat Vu, Hieu Le, Baoru Huang, Binh Huynh, Thieu Vo, Andreas Kugi, and Anh Nguyen. Grasp-anything: Large-scale grasp dataset from foundation models. *arXiv preprint arXiv:2309.09818*, 2023. 3
- [56] An Dinh Vuong, Minh Nhat Vu, Baoru Huang, Nghia Nguyen, Hieu Le, Thieu Vo, and Anh Nguyen. Language-driven grasp detection. In *CVPR*, 2024. 3
- [57] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *CoRL*, 2023. 2
- [58] Xiaohan Wang, Yuehu Liu, Xinhang Song, Yuyi Liu, Sixian Zhang, and Shuqiang Jiang. An interactive navigation method with effect-oriented affordance. In *CVPR*, 2024. 1
- [59] Zan Wang, Yixin Chen, Baoxiong Jia, Puha Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *CVPR*, 2024. 1
- [60] 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. 4, 6, 1
- [61] Ran Xu, Yan Shen, Xiaoqi Li, Ruihai Wu, and Hao Dong. Naturalvlm: Leveraging fine-grained natural language for affordance-guided visual manipulation. *arXiv preprint arXiv:2403.08355*, 2024. 2, 3

- [62] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *ICCV*, 2023. 3
- [63] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1
- [64] Qiaojun Yu, Siyuan Huang, Xibin Yuan, Zhengkai Jiang, Ce Hao, Xin Li, Haonan Chang, Junbo Wang, Liu Liu, Hongsheng Li, et al. Uniaff: A unified representation of affordances for tool usage and articulation with vision-language models. *arXiv preprint arXiv:2409.20551*, 2024. 2, 3
- [65] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 5
- [66] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023. 5
- [67] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 1
- [68] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaeei, 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. 4, 1

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

Supplementary Material

8. Details of Data Annotation

As the original data sources, such as HANDAL [15], Open-X [42], EgoObjects [68], GraspNet [10], 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: ①: Original mask, ②: SAM2 [50], ③: Florence2 [60] + SAM2, ④: VLPart [54] + SAM2, ⑤: Human (+ SAM2). The details of how to compose these tools for one specific dataset are shown in Table 11. In addition, Table 11 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.*

9. 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. 8. 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 [42] is segmented using the combination of Florence2 [60] and SAM2 [50] according to its original grasping instruction. The knife handle from EgoObjects [68] can be accurately grounded using VLPart [54], and its output box can be further transformed into a pixel-wise mask using SAM2 [50]. Regarding the microwave handle in the EgoObjects dataset [68], 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.

10. Reasoning-based Affordance Examples

More reasoning-based affordance segmentation examples are shown in Fig. 10. 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 Ta-

ble 12 and Table 13. 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”.

11. 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 [67], COCO-Stuff [5], PACO [47]), referring segmentation (*e.g.*, RefCOCO [63]), VQA (*e.g.*, LLaVA-150k [33]) and reasoning-based segmentation (*e.g.*, ReasonSeg [26]). The data sampling ratios are presented in Table 10. We deploy eight NVIDIA A100 GPUs (80GB) to train our model, with a learning rate of 2e-5. The training loss follows [26], 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.

12. More Results on Visual Affordance

We provide more visualization results of affordance segmentation from our AffordanceVLM model in Fig. 10. The testing images are selected from multiple validation 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.

13. 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

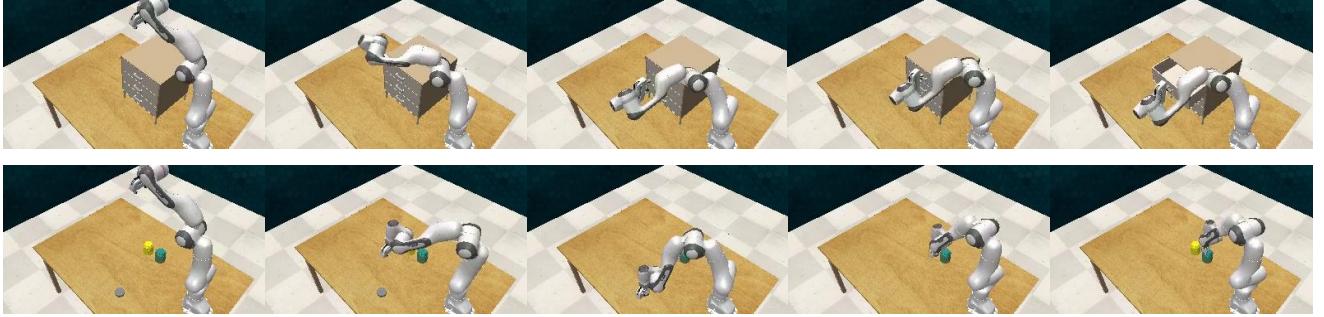


Figure 7. **Object grasping results from our AffordanceNet on RLBench.** The instruction of the top video is “*Open the top drawer*”, and the bottom one refers to “*Close the green jar*”.

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

Table 10. **Data sampling ratios during training.**

object perception and grasping proficiency.

14. More Results on RLBench

We present several visualization results from the simulation task RLBench in Figure 7. 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 8. More affordance segmentation examples from our RAGNet. It covers various data sources, like robot, wild, and ego-centric domains.

Easy Instructions



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.

Hard Instructions



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

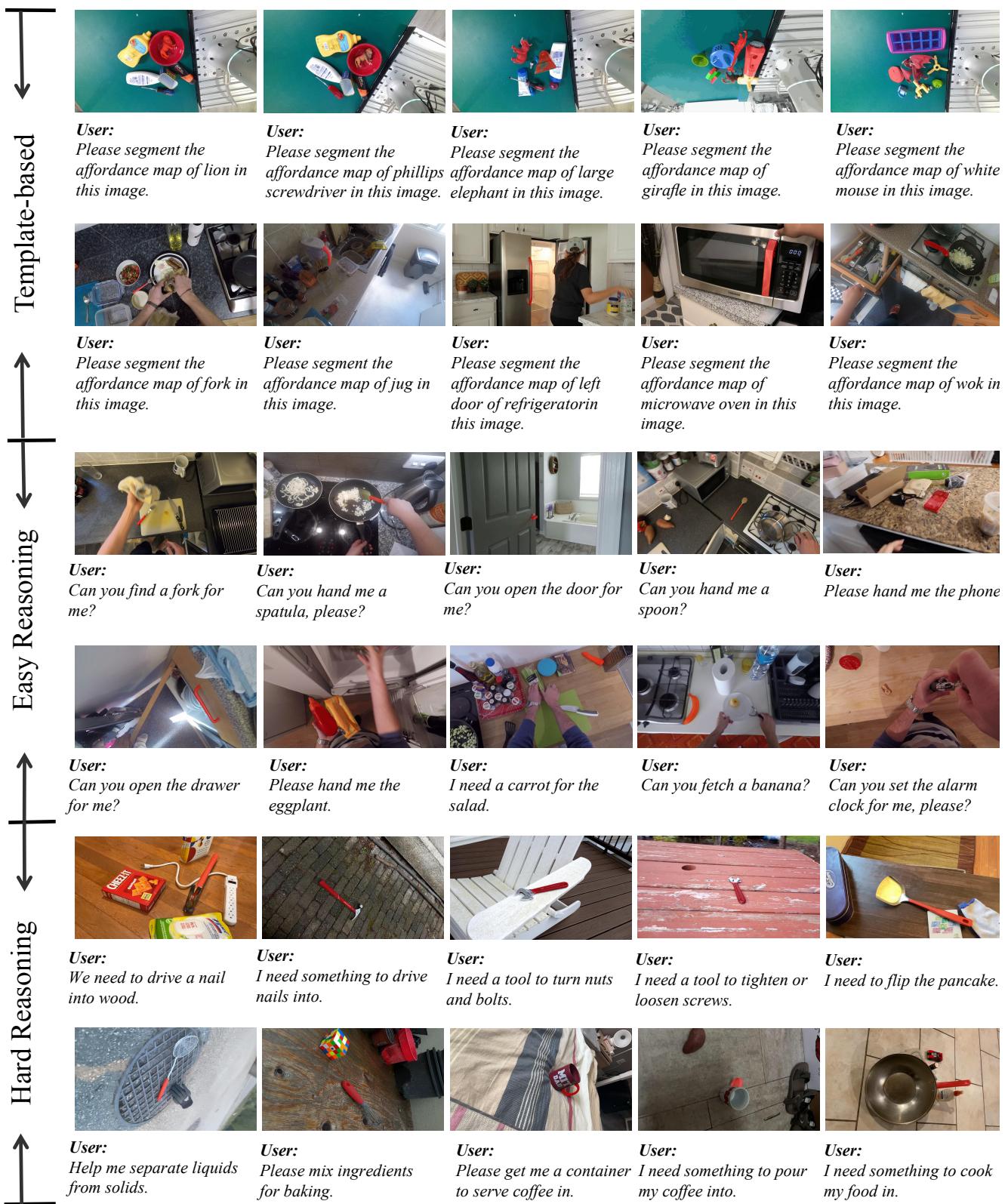


Figure 10. More experiment results from our model. We use template-based, easy reasoning-based, and hard reasoning-based instructions, respectively.

Dataset	Subset	Annotation Tool and Categories
HANDAL	-	①: 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	③: 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, ⑤: right fridge door, bottom drawer, left fridge door, middle drawer, top drawer
	Bridge	③: 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 ④: knife ⑥: 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	-	②: 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 ④: spoon, mug, screwdriver, knife, wrench ⑤: microwave oven, washing machine, wok, oven, drawer, teapot, toothbrush, wardrobe, door, jug, refrigerator, tap, tennis racket, spatula, fork, frying pan, scissors, hammer
GraspNet	-	①: 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 ⑤: flat screwdriver, power drill, scissors, mug
HANDAL <i>Reasoning</i>	-	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 <i>Reasoning</i>	-	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 11. 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 12. 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 13. Language prompt when **generating hard reasoning-based instructions on HANDAL via GPT-4**.