InstructPart: Affordance-based Part Segmentation from Language Instruction

Zifu Wan Yaqi Xie Ce Zhang Zhiqiu Lin Zihan Wang
Simon Stepputtis Deva Ramanan Katia Sycara

Robotics Institute
Carnegie Mellon University
{zifuw, yaqix, cezhang, zhiqiul, zihanwa3, sstepput, deva, sycara}@andrew.cmu.edu

Abstract

Recent advancements in Vision-Language Models (VLMs) have led to their increased application in robotic tasks. While the implementation of VLMs is primarily at the object level, the distinct affordances of an object's various parts — such as a knife's blade for cutting versus its handle for grasping — remain a challenge for current state-of-the-art models. Our investigations reveal that these models often fail to accurately segment parts based on task instructions, a capability crucial for precise robotic interactions. Addressing the lack of real-world datasets to evaluate these fine-grained tasks, we introduce a comprehensive dataset that includes image observations, task descriptions, and precise annotations for object-part interactions, complemented by part segmentation masks. We present an evaluation of common pre-trained VLMs using this benchmark, shedding light on the models' performance in understanding and executing part-level tasks within everyday contexts.

Introduction

Robots play an increasingly significant role in our daily life (Matheson et al. 2019; Kaiser et al. 2021). However, before moving on to the next generation filled with more advanced agents, two critical challenges remain. First, robots must *comprehend natural language instructions in the context*, i.e., translating commands into actionable tasks. This ability is essential for seamless interactions with humans, especially in unstructured environments such as homes. Another key challenge for robots is to *perceive their environments and ground to specific areas*. This involves not only recognizing objects but also fine-grained *parts* that the robot should interact with. Specifically, given the instruction to cut an onion and a visual observation depicted in Fig. 1, an intelligent agent needs to first understand that the handle of the knife can be held and then refer to the part mask before performing subsequent manipulation and control tasks. To rigorously evaluate the agents' capability in addressing these challenges, we introduce a comprehensive dataset comprising image observations with task instructions, accompanied by part segmentation masks as ground truth.

Despite significant progress in language comprehension and visual perception, current foundational vision-language models (VLMs) still struggle with fine-grained *multimodal* reasoning, e.g., accurately localizing specific parts of an object based on user instruction and functional reasoning (Zhou et al. 2022; Mogadala, Kalimuthu, and Klakow 2021). We attribute this primarily to the scarcity of training data, as annotating part-level labels is excessively time-consuming and expensive. For instance, most large-scale vision datasets focus on object-level understanding (Liu et al. 2023a; Zou et al. 2023b,a; Xu et al. 2023; Liang et al. 2023; Sun et al. 2023) and existing part-level recognition datasets are either too limited in part categories (Nguyen et al. 2017; Myers et al. 2015; Roy and Todorovic 2016), collected in a controlled environment (Myers et al. 2015), or obtained from simulator (Geng et al. 2023; Deng et al. 2021; Xiang et al. 2020; Mo et al. 2019). Moreover, these datasets are not specifically designed for robotics tasks and thus do not include language instructions. Given the recent surge of interest in building next-generation embodied agents (Yang et al. 2023a; Huang et al. 2023; Ahn et al. 2022) capable of perceiving multimodal inputs and following language instructions (Touvron et al. 2023; Liu et al. 2023b), we are motivated to comprehensively assess existing VLMs by building a dataset including both language and part annotations.

To this end, we introduce a novel dataset, **InstructPart**, consisting of 700 images, 54 object classes, and 48 part classes. Each image is accompanied by human-annotated and GPT-polished instructions for common household tasks and detailed part segmentation masks. Thorough evaluations of current visual language models on our dataset reveal a notable deficiency in their capacity to comprehend natural language and accurately ground it across diverse objects and

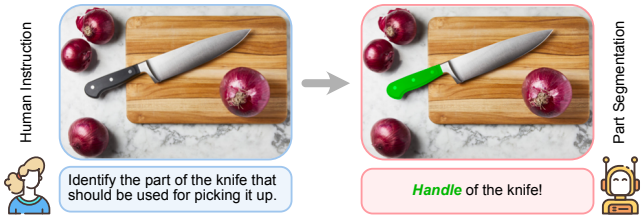

Figure 1: Task Description for InstructPart: Presented with an image observation (left) and a corresponding instruction (in the blue box on the left), the model is required to identify and output the specific part segment (highlighted in the green mask on the right) referenced in the instruction.

parts. This finding underscores the necessity to resolve a critical shortfall in vision-language models for robotics.

To enhance our understanding of our dataset's capabilities, we fine-tune a state-of-the-art model using one-third of the dataset. This approach results in a significant 20% improvement in performance, demonstrating the exceptional quality and value of our data for advanced training purposes. With our proposed benchmark, we emphasize the essentiality of advancing vision-language models to excel not only at object-level understanding but also at discerning parts-level details. This advancement is crucial in varied contexts, from home service robots aiding in domestic tasks to industrial robotic arms, all playing a vital role in improving our daily experiences. Our contributions are as follows:

- To the best of our knowledge, we present the first dataset that bridges instruction-based interactions with part segmentation for common household tasks, a critical step towards more intuitive and versatile robotic systems.
- We rigorously evaluate various vision-language models on the introduced dataset, demonstrating their applicability and potential for the advancement of the community.
- We boost the state-of-the-art model's performance by 20% via fine-tuning with one-third of our dataset, highlighting our data's quality and training potential.

Related Work

Part Segmentation

Object segmentation aims to find semantically meaningful pixels of each object in an image. Unlike traditional object segmentation, part segmentation provides a more fine-grained understanding of each object, assigning different semantic labels to prominent parts within an object (Wang et al. 2015). Previous methods are mainly based on a fully supervised manner which needs to be trained on numerous data, and various part annotation datasets have been collected to support the process (Sun et al. 2023), such as PartImageNet (He et al. 2022), Pascal-Part (Chen et al. 2014), ADE20K (Zhou et al. 2019), and PACO (Ramanathan et al. 2023). However, such methods are constrained to the domain of the training data and focused on specific classes, such as humans (Gong et al. 2017), birds (Wah et al. 2011), fashion (Jia et al. 2020), and cars (Song et al. 2019). This limits their usage in daily scenarios, especially for the interaction of robots and the environment.

In robotic tasks, understanding the parts of articulated objects is a popular topic, and the affordances of parts are used for further manipulation (Gadre, Ehsani, and Song 2021; Yi et al. 2018). Related datasets include PartNet (Mo et al. 2019), PartNet-Mobility (Xiang et al. 2020), 3D AffordanceNet (Deng et al. 2021), GAPartNet (Geng et al. 2023) etc. However, these datasets are all generated from simulators and have potential sim-to-real gaps when transferring to real-world scenarios. On the other hand, several works collect real-world images curated for affordance understanding, such as UMD-Affordance (Myers et al. 2015), NYUv2-Affordance (Roy and Todorovic 2016), and IIT-AFF (Nguyen et al. 2017). However, these datasets cover a limited set of scenes and contain less than 10 classes of affordance. Besides, using a simple word or phrase can be insufficient to represent affordance sometimes. For example, the affordance of the light switch can be "turn on", while a more precise description can be "press" or "twist" according to the switch's type. In real-world situations, people tend to refer to a part using an instructional sentence instead of a single word. Motivated by this, we construct a comprehensive dataset with instruction-part pairs, object-part classes, affordances, and actions. A comparison of relevant real-world part segmentation datasets is shown in Tab. 1. We hope the dataset can provide more insights into future work about the interaction between robots and the environment.

Table 1: Comparison of relevant part segmentation datasets. We show the number of object classes (#Object), part classes (#Part), affordances (#Affordance), actions (#Action), and whether instructions are included (Instruction). N/A means there is no such type of data, while – means the data exists while no relevant information is provided. 11/158 indicates the super-class and sub-class numbers in PartImageNet.

	#Object	#Part	#Affordance	#Action	Instruction
UMD	17	N/A	7	N/A	✗
NYUv2	40	N/A	5	N/A	✗
IIT-AFF	10	N/A	9	N/A	✗
PartImageNet	11/158	13	N/A	N/A	✗
Pascal-Part	20	–	N/A	N/A	✗
PACO	75	–	N/A	N/A	✗
InstructPart	54	48	32	38	✓

Open-Vocabulary Segmentation

Traditional fully-supervised recognition methods have limited generalization ability, leading to more attention to recognizing open-world classes beyond the training phase. Recently, the pre-trained vision language model, CLIP (Radford et al. 2021), which was trained on 400 million image-caption pairs, aligns the gap between embedding space of vision and language and demonstrates valuable potential in open-vocabulary segmentation. Most related methods use existing image encoders as the backbone, e.g., MaskFormer (Cheng et al. 2022) or SAM (Kirillov et al. 2023), and apply CLIP for classification. The key to these methods is to align the embedding space of the image encoder with that of CLIP. For example, OVSeg (Liang et al. 2023) proposes to crop the region proposals and finetune CLIP using a mask prompt tuning mechanism. FC-CLIP (Yu et al. 2023) uses a frozen convolutional CLIP backbone and aligns the region features from the visual backbone with CLIP. SAN (Xu et al. 2023) applies a side adapter network to a frozen CLIP to get the class of masks.

Going beyond object-level segmentation, more recent works propose the open-vocabulary part segmentation task. VLPart (Sun et al. 2023) parses the novel object into parts using its semantic correspondence with the base object and classifies it with CLIP. OPS (Pan et al. 2023) generates pseudo labels for unlabeled data with a trained object detector and clusters between different parts in a self-supervised procedure. However, OPS cannot predict semantic labels for the segments and is not suitable for our task.

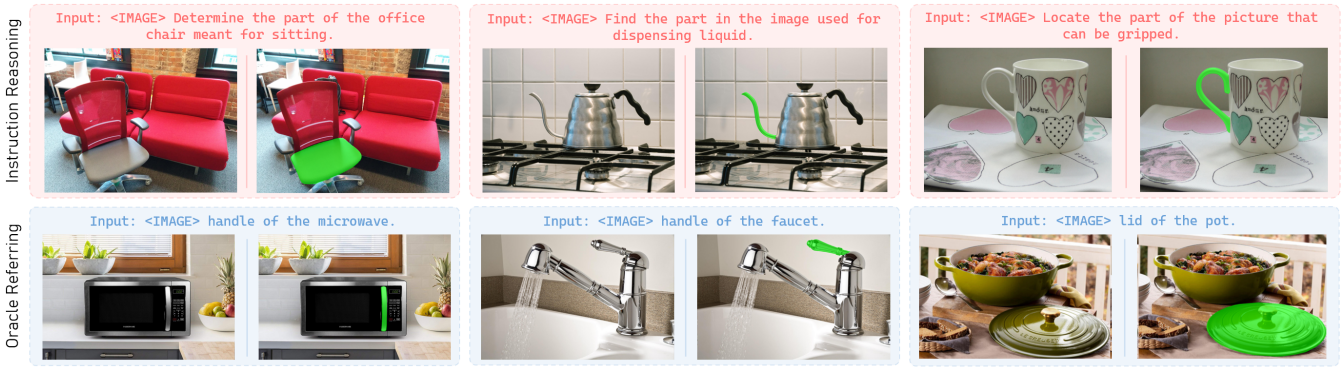

Figure 2: Examples from our InstructPart dataset are illustrated as follows: instructions are denoted in red text, while object and part names are indicated in blue. Each example includes a input observation image (left), with the corresponding ground truth part segments (right), highlighted with a green mask.

Although these open-world recognition methods have demonstrated potential in recognizing out-of-distribution classes, they have limited reasoning ability to understand complex instructional sentences, thus prohibiting their wide usage in real-world robotic tasks.

Referring Segmentation

The aforementioned open-world recognition methods recognize objects and parts by their names while failing to refer to a specific instance using a detailed description. On the other hand, referring expression segmentation aims to generate a segmentation mask from a given language expression (Hu, Rohrbach, and Darrell 2016), and various datasets such as ReferIt (Kazemzadeh et al. 2014), CLEVR-Ref+ (Liu et al. 2019), refCOCO (Yu et al. 2016), refCOCOg (Mao et al. 2016), gRefCOCO (Liu, Ding, and Jiang 2023) are collected to support the task. These datasets contain image-expression pairs and the masks of objects being referred to. Popular referring segmentation methods use a visual and a language encoder to extract features from the two modalities respectively, and design attention mechanisms to incorporate the features and assemble classes for region masks (Yang et al. 2022; Liu, Ding, and Jiang 2023; Ouyang et al. 2023; Liu et al. 2023a). Recently, more works have applied pre-trained foundation models, e.g., SAM(Kirillov et al. 2023) and CLIP (Radford et al. 2021) as the encoder and focused on the design of the decoder, such as X-Decoder (Zou et al. 2023a) and SEEM (Zou et al. 2023b). However, the referring expression task only takes short phrases as input and does not consider complex reasoning, for example, when the target name does not directly appear in the expression.

Reasoning Segmentation

On the other hand, remarkable advances have been made in large language models (LLMs), which possess the ability to understand complex language inputs and have the potential for more complex referring segmentation. Models such as BLIP-2 (Li et al. 2023), LLaVA-1.5 (Liu et al. 2023b), MiniGPT-4 (Zhu et al. 2023), Flamingo (Alayrac et al. 2022), and GPT-4V (Yang et al. 2023c) have explored the design of multi-modal LLMs for visual understanding and demonstrate their ability through tasks such as image captioning, visual question answering (VQA), etc. To enable the grounding ability of multimodal LLMs, VisionLLM (Wang et al. 2023) proposes an open-ended task decoder with LLM and returns the coordinates of object polygons. Similarly, Shikra(Chen et al. 2023b) and MiniGPT-v2 (Chen et al. 2023a) process object coordinates as input and enable the localization ability by returning coordinates. However, both these methods cannot produce segmentation masks and can only implicitly generate texts using LLMs rather than using a visual decoder for localization directly, which can be counterintuitive for image segmentation or detection.

Recently, LISA (Lai et al. 2023) incorporates a multimodal LLM(Liu et al. 2023b) with a vision backbone and jointly trains a decoder to produce the segmentation masks. They propose a new task – reasoning segmentation, which requires the model to ground an area after comprehending a complex input sentence. Intuitively, we are curious about whether LISA has the ability to reason from instructions in the context of robotic applications and refer to specific parts that is essential for effective interaction.

To quantitatively evaluate the aforementioned methodologies in these scenarios, we propose the InstructPart dataset, which contains instruction-part pairs, object-part names, high-level affordance, low-level action, and the part segmentation mask. It aims to rigorously test and potentially improve the ability of current models to accurately identify specific parts based on instructions in robotic applications.

Dataset and Task Settings

InstructPart Dataset

Motivated by scenarios where an agent is required to follow instructions, we create the InstructPart dataset. This dataset is designed to measure the effectiveness of current models in understanding natural language and their ability to perform reasoning and grounding. The dataset consists of 700 images sourced from Flickr and the internet, carefully chosen to align with everyday household tasks. We ensure a uniform collection of object classes and thoroughly annotate all relevant parts within each image.

We design tailored instructions for each image to aid intelligent agents in better understanding their surroundings and performing an action, as illustrated in the first row of Fig. 2.

In the three examples shown in the first row, the agents need to understand the parts for sitting, dispensing, and gripping respectively, where the parts are highlighted in the green masks in the image. For each image-instruction pair, we annotate all the fine-grained segmentation masks that satisfy the instruction and treat them as the ground truth to evaluate whether a system successfully localizes the regions.

We purposefully avoid specific references to part names in our instructions to enhance their practicality in real-world situations. For instance, commonly used expressions such as *"Flush the toilet"* or *"Turn on the faucet"* are preferred over more detailed directives like *"Press the toilet handle"* or *"Lift the faucet handle"*. We had 6 human experts create free-form natural language instructions, which were then refined using GPT-4 to ensure grammatical precision and diverse sentence construction. This process was followed by a thorough human verification of the refined instructions.

In addition to the instruction-image pairs, we provide the names of objects and parts relevant to the image, such as *seat of the chair*, *spout of the kettle*, *handle of the cup*. We also include a corresponding affordance and action for each instruction. Specifically, affordances refer to low-level actions performed to a specific part, like *"pull"*, *"push"*, or *"twist"*, while actions refer to the high-level function to be achieved, such as *"turn on"*, *"pick up"*, or *"open"*. Note that the affordance and action could be identical sometimes, e.g., *"pour"*, *"cut"*, etc. In the examples shown in the first row of Fig. 2, the affordances are *"support"*, *"pour"*, *"grip"*, and the actions are *"sit"*, *"pour"* and *"pick up"*. This allows us to categorize affordances into two levels, addressing the ambiguity in definitions as noted in previous studies (Nguyen et al. 2017; Roy and Todorovic 2016; Myers et al. 2015).

In summary, the components for an image of our dataset can be represented as:

$$(I_{\text{text}}, I_{\text{image}}, O, P, M, A_{\text{affordance}}, A_{\text{action}}),$$

where these items refer to text instruction, image observation, object name, part name, segmentation mask, affordance name, and action name, respectively. Note that $I_{\text{text}} \in \{I_{\text{human}}, I_{\text{GPT}}\}$, which means the text instruction is either directly annotated by humans or rewritten by GPT-4.

Task Definition

The task of localizing an area based on a given instruction necessitates the model's capability to comprehend complex directives and refer to them appropriately. Accordingly, we propose the **Instruction Reasoning Part Segmentation (IRPS)** task, challenging the model to develop proficiency in both linguistic reasoning and visual grounding. Furthermore, to exclusively evaluate the visual grounding capability of existing models, we introduce the **Oracle Referring Part Segmentation (ORPS)** task, which utilizes oracle information about the designated object and part.

Instruction Reasoning Part Segmentation (IRPS). To explore the reasoning and part grounding ability of current models, we propose the IRPS task, which is shown in the first row of Fig. 2. The models should only take an instruction-image pair as the input, and find the part segmentation masks being referred to, which are shown in green masks in Fig. 2. This requires the model to possess the ability to understand the instruction, analyze the image, and refer to the corresponding part area. The task can be formulated as:

$$\mathcal{F}(I_{\text{text}}, I_{\text{image}}) \Rightarrow M, \tag{1}$$

where $\mathcal{F}$ is the vision-language model, and $I_{\text{text}} \in \{I_{\text{human}}, I_{\text{GPT}}\}$, indicating that the instruction can be either human-annotated or GPT-4 rewritten.

Oracle Referring Part Segmentation (ORPS). In the ORPS setting, we consider using the part name to directly refer to the part, which ensures the model has a correct text input. The task is shown in the second row of Fig. 2. We formulate the ORPS task in two formats:

1. We use a template to connect the part name and the object name with the word *"of"*, e.g., *the handle of the faucet*:

$$\mathcal{F}(P \text{ of } O, I_{\text{image}}) \Rightarrow M. \tag{2}$$

2. Besides, considering that affordance could potentially help the model refer to a part, we add the affordance name to the former format:

$$\mathcal{F}(P \text{ of } O \text{ that } A_{\text{a}}, I_{\text{image}}) \Rightarrow M, \tag{3}$$

where A_{a} incorporates the affordance, e.g., *the handle of the cup that can be held*.

Metrics

We follow LISA (Lai et al. 2023) to use two metrics, gIoU and cIoU. gIoU is the average of all per-image Intersection-over-Unions (IoUs), and cIoU is defined by the cumulative intersection over the cumulative union.

Besides, to evaluate the precision of the models, we adopt Precision@50 (P@50) metric as the previous referring segmentation works (Liu et al. 2023a; Mao et al. 2016) and develop a Precision@50:95 (P@50:95) metric according to COCO (Lin et al. 2014). The P@50 metric simply considers a mask to be a true positive when the IoU ratio exceeds 0.5, and P@50:95 calculates across a range of IoU thresholds from 0.50 to 0.95 with increments of 0.05, then averages across all the thresholds. The P@50:95 metric requires a higher least IoU for the prediction hence it is always lower than the P@50 metric.

For the two metric types, IoU and Precision, the latter metric only counts those results greater than a threshold, hence can pose more challenges to the model than the former one and fairly evaluate the results with a high recall rate.

Experiments

Evaluated Methods

Open-vocabulary Segmentation Models. The open-vocabulary part segmentation model, i.e., VLPart (Sun et al. 2023), is intuitively suitable for our tasks since plentiful part segments were used for training. As a result, we would like to know whether they can perform well in our dataset collected especially for robotic daily tasks. We also choose OVSeg (Liang et al. 2023) and SAN (Xu et al. 2023) to discover the performance of the open-vocabulary object segmentation methods on our task. We select the best-reported models for the three methods.

Refering Segmentation Models. We conduct experiments with off-the-shelf models including X-Decoder (Zou et al. 2023a), SEEM (Zou et al. 2023b), and TRIS (Liu et al. 2023a). Besides, we also evaluate Grounding-DINO (Liu et al. 2023c), which has witnessed a great open-vocabulary referring detection ability and been integrated with SAM (Kirillov et al. 2023) to a project, Grounded-SAM[1]. We adopt the best models for these methods.

Reasoning Segmentation Models. For our tasks, LISA (Lai et al. 2023) can naturally be a good choice since it can return masks and has been trained on several part segmentation datasets. As a result, it is interesting to explore whether it possesses the ability to understand instructions and find part segments. Other multi-modal LLMs, including VisionLLM (Wang et al. 2023), Shikra (Chen et al. 2023b), and MiniGPT-v2 (Chen et al. 2023a) also have localization ability. Since they can only return bounding box outputs, we use the results as box prompts for SAM (Kirillov et al. 2023) to get a mask output for fair comparison. However, we cannot test VisionLLM since its code has not been released. We adopt LISA-7B-v1 (Lai et al. 2023) model that has been fine-tuned on both training and validation data of LISA's dataset. Besides, we select the Shikra-7B-delta-v1-0708 for Shikra and the stage-3 model for MiniGPT-v2.

Grid-based GPT-4V. The recent release of GPT-4V has demonstrated remarkable advancements in complex visio-linguistic reasoning (Yang et al. 2023c), outperforming its predecessors in previous challenges, such as Winoground (Thrush et al. 2022). As a natural thought, we wonder if GPT-4V can also succeed in understanding and grounding object parts. However, GPT-4V API cannot return segmentation mask output directly, and our preliminary experiments showed that GPT-4V performs poorly when it is asked to generate text coordinates. As a result, we first use Grounding-DINO (Liu et al. 2023c) to find the bounding box of the entire object and crop it, then ask GPT-4V to virtually divide the box to 7×7 grids and identify the grids including the desirable parts. Afterward, the coordinates of the grids are used as a prompt for SAM (Kirillov et al. 2023) to obtain the segmentation mask.

SoM-based GPT-4V. SoM (Yang et al. 2023b) proposes to label the masks obtained by SAM (Kirillov et al. 2023) with numbers in the center of each object. As it proves that precise referring can boost the performance of GPT-4V, we apply a similar manner for our part segmentation task. Although SAM (Kirillov et al. 2023) can be a superior choice to obtain masks at multiple granularities (Zou et al. 2023b), it is prone to failing in part segmentation for small objects. As a result, we add Grounding-DINO (Liu et al. 2023c) to detect the object first, then apply SoM to the object patch instead of entire image.

Quantitative Results

The left part of Tab. 2 shows the result of oracle referring part segmentation, where object and part names are explicitly embedded into a template, mitigating the need for models' reasoning ability. The *Object-Part* column stands for the results using the template mentioned in Eq. 2, while the *Object-Part-Affordance* column incorporates the affordance according to Eq. 3. The right part of Tab. 2 shows the result of instruction reasoning part segmentation, where part names are not present in the instruction and require more reasoning ability to understand the implicit meaning.

Comparing the left and right parts of Tab. 2 we can find that the performance of oracle referring task is generally better than that of instruction reasoning. This demonstrates that current models lack the reasoning ability to infer from an instruction-image pair to the correct interactive part. For the oracle referring segmentation task, incorporating the affordance in the instruction leads to no apparent increase in the average performance. While LISA and X-Decoder achieve some increase, other models are impacted by the affordance. This indicates that most models may not possess the common sense to relate a part to an affordance. Besides, from the average results shown in the right part of Tab. 2, we can find that GPT-4 rewritten instructions lead to overall better performances. We infer this may derive from GPT-4's potential reasoning ability. For example, given an instruction *"If I want to pick up the knife, which part in the picture can be used?"*, GPT-4 would explicitly point out the specific part that the model should refer to, such as *"Indicate which part of the picture represents the handle of the knife."*. This performance reveals that current vision-language models (VLMs) still lag behind LLMs in the realm of reasoning, and more work needs to be done to leverage LLMs' ability to VLMs.

As the oracle referring segmentation task is the optimal condition of instruction reasoning segmentation task, we will first analyze it to explore the visual grounding ability of the models.

Oracle Referring Results. The methods in Tab. 2 are divided into 3 categories, namely, open vocabulary segmentation (OVS), referring expression segmentation (RES), and reasoning segmentation (RS). VLPart, although has been trained on numerous part segmentation data, still fails to handle the oracle referring task, indicating that it cannot generalize to our robot-oriented data distribution. The OVS methods, OVSeg and SAN, outperform VLPart in two IoU metrics. However, it obtains low P@50 and P@50-95 scores, which can be attributed to the models generating object segments rather than parts, leading to high recall rates while suffering from reduced accuracy.

Out of expectation, the RES methods, X-Decoder, SEEM, and TRIS perform poorly even in the oracle referring part segmentation task. On the other hand, Grounded-SAM is relatively better than the aforementioned methods. This indicates that its base model, Grounding-DINO (Liu et al. 2023c), possesses a better ability to detect parts than other segmentation-based methods. This may be explained by the fact that detection data, with the format of only four coordinates, is easier to obtain and Grounding-DINO has seen more data with various distributions.

LISA, as a reasoning segmentation method trained on sentence-mask pairs, is indeed better than previous methods. Remarkably, even though Shikra and MiniGPT-v2 are

[1]https://github.com/IDEA-Research/Grounded-Segment-Anything

Table 2: Results on oracle referring part segmentation task (left) and instruction referring part segmentation task (right). We divide the methods into three categories, namely, open-vocabulary segmentation (OVS), referring expression segmentation (RES), and reasoning segmentation (RS).

	Methods	Oracle referring part segmentation results								Instruction reasoning part segmentation results							
		Object-Part				Object-Part-Affordance				Human-Annotated				GPT-4-Rewritten			
		gIoU	cIoU	$P_{50\text{-}95}$	P_{50}	gIoU	cIoU	$P_{50\text{-}95}$	P_{50}	gIoU	cIoU	$P_{50\text{-}95}$	P_{50}	gIoU	cIoU	$P_{50\text{-}95}$	P_{50}
OVS	VLPart	15.11	17.93	11.44	15.38	13.56	12.15	10.86	13.64	0.32	0.94	0.06	0.15	0.46	0.97	0.20	0.44
OVS	OVSeg	20.86	17.78	7.21	15.67	20.61	17.59	7.07	15.53	16.44	13.11	5.34	10.74	16.93	14.21	4.67	10.89
OVS	SAN	13.45	19.70	5.36	10.74	14.72	20.17	5.86	12.19	9.03	13.99	2.70	6.10	9.02	15.87	2.79	6.39
RES	X-Decoder	8.76	10.90	2.79	5.66	9.64	11.47	3.11	6.53	9.92	10.62	3.32	6.97	9.51	9.74	2.99	6.68
RES	SEEM	8.29	11.66	2.53	5.37	8.29	11.66	2.53	5.37	7.80	9.88	2.00	4.50	9.39	11.67	2.55	6.53
RES	TRIS	17.04	17.28	5.12	11.76	17.05	17.28	5.01	12.19	14.74	14.85	3.53	9.58	15.47	15.56	3.95	10.89
RES	G-SAM	26.21	22.47	10.09	19.30	25.71	22.43	9.78	18.87	22.11	19.79	7.16	15.67	22.38	19.60	7.30	15.53
RS	LISA	35.91	41.36	20.55	34.98	36.67	42.91	20.87	35.56	26.06	28.63	14.15	23.08	26.25	30.14	13.72	24.38
RS	Shikra	37.74	33.89	24.37	36.87	35.44	32.64	21.94	33.24	2.76	4.33	1.26	2.18	8.59	11.38	4.38	6.82
RS	MiniGPT-v2	37.64	40.80	22.44	34.11	36.32	38.88	21.38	32.66	21.81	20.95	10.39	17.56	23.01	22.18	12.38	20.46
	Average	22.10	23.38	11.19	18.98	21.80	22.72	10.84	18.58	13.10	13.71	4.99	9.65	14.10	15.13	5.49	10.90

Table 3: GPT-4's performance in the object-part oracle referring part segmentation task, as applied to a subset of InstructPart.

Methods	Object-Part			
	gIoU	cIoU	$P_{50\text{-}95}$	P_{50}
Grid-based GPT-4V	14.14	17.15	5.67	12.37
SoM-based GPT-4V	25.41	26.82	17.90	25.81

not specifically designed for segmentation tasks, simple integration with SAM (Kirillov et al. 2023) yields results that surpass those of LISA. This demonstrates again that detection data with a wide range of distribution is easier to collect and could cover more part categories.

Instruction Reasoning Results. In the right part of Tab. 2, we report the result using our hand-collected data and GPT-4 rewritten data respectively. As the task poses more challenges to understanding sentences, all the results undergo a decrease, especially in VLPart, which drops significantly. The results of OVSeg and SAN have a smaller decrease, which can be explained that these models recognize keywords to find the entire object. Since our instructions do not intentionally avoid object names, these OVS methods can still find the entire object. Interestingly, the performance of X-Decoder experiences a small increase in the human-annotated data. Not surprisingly, LISA performs best among all the models since it has been trained on similar data. However, Shikra experiences an unusual decrease of nearly 20 times. This indicates that although trained on various visiolinguistic data, Shikra still lacks enough reasoning ability to handle our instruction reasoning task.

GPT-4V Based Methods Results. Tab. 3 shows the results of two GPT-4V segmentation methods. However, due to the quota restriction, we are not able to frequently call GPT-4V API. As a result, we first test the two methods on the oracle referring task to explore GPT-4V's localization ability. Besides, we select a subset consisting of 226 samples from the dataset according to the original category distribution. Although the results cannot be fairly compared with other methods in Tab. 2, it still reveals the poor performance of GPT-4V. For the two methods, the SoM-based method performs better, demonstrating that GPT-4V can hardly directly localize the part targets and has to choose from a set. However, compared to the satisfactory results in (Yang et al. 2023b), SoM fails to handle our dataset. This may be explained by two reasons: 1) While GPT-4V can localize objects (Yang et al. 2023b), we hypothesize that it is not directly trained on fine-grained part data. 2) Labeling numbers in the center of fine-grained parts may lead to overlapping and ambiguity in referring.

Qualitative Results

Fig. 3 shows the visualization results on the instruction reasoning segmentation task. The first column depicts the ground truth labels, and the remaining columns include the results of three categories, namely, OVS: OVSeg (Liang et al. 2023) , SAN (Xu et al. 2023) , RES: X-Decoder (Zou et al. 2023a) , SEEM (Zou et al. 2023b) , TRIS (Liu et al. 2023a) , Grounded-SAM, and RS: MiniGPT-v2 (Chen et al. 2023a) , LISA. (Lai et al. 2023) . We do not include results on VLPart (Sun et al. 2023) and Shikra (Chen et al. 2023b) because they barely obtain any output in these samples.

The first five rows show *handle of the refrigerator*, *bowl of the spoon*, *handle of the microwave*, *seat of the chair*, *stem of the wine glass*, respectively. We can find that LISA can correctly distinguish a specific part from the entire object while other methods tend to predict a larger area. However, in the remaining three samples, including *handle of the knife*, *handle of the pliers*, and *handle of the cup*, although the parts are distinguished from the background and seem to be easy for human judgment, all the models struggle to find the correct part. This indicates that the models lack understanding of the feature of parts and more work needs to be performed to solve the challenge.

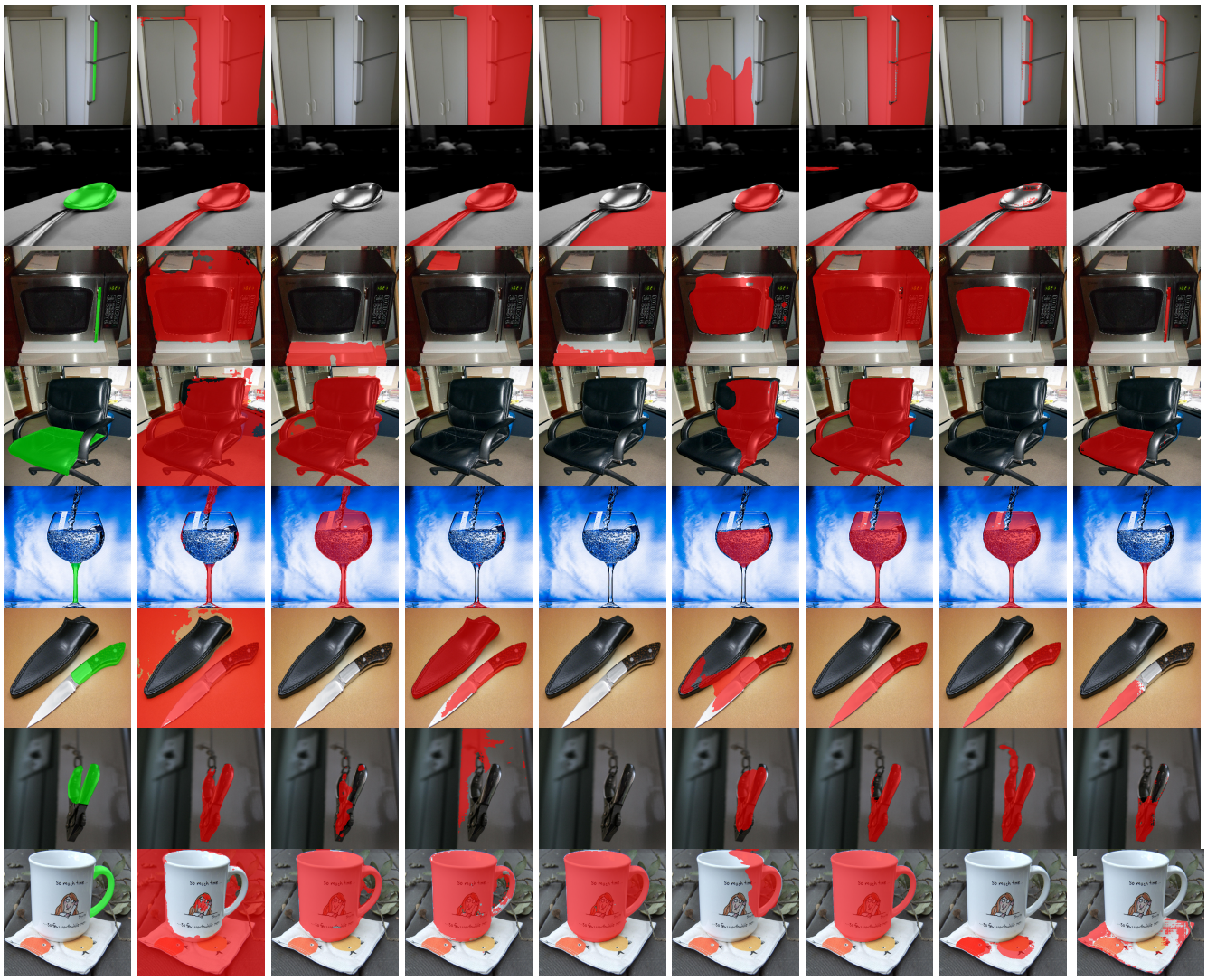

Figure 3: Visualization results on the instruction reasoning part segmentation task. The instructions are human-annotated. Green masks stand for the ground truth and red masks represent the predictions.

Discussion

Our findings reveal that reasoning segmentation techniques typically outperform other methods. This highlights the effectiveness of end-to-end multi-modal foundation models in vision-language grounding, surpassing traditional models that are limited to specific, narrow tasks. Additionally, the impressive results achieved by detection-based models suggest that incorporating a mix of detection and segmentation data during the training phase could be beneficial.

Moreover, we show that a minimal set of our InstrutPart samples can significantly enhance instruction reasoning capabilities in existing models. Specifically, fine-tuning the LISA-7B-v1 model (Lai et al. 2023) with only 226 samples for 600 iterations yielded notable improvements. This was evidenced by the substantial increases in gIoU (26.18% to 42.01%) and cIoU (28.10% to 48.75%) on the test set comprising the remaining 474 samples. This demonstrates that even a small subset of our high-quality data can significantly enhance the model's ability to comprehend instructions and segment parts, affirming the exceptional quality and utility of our data for further training.

Conclusion

In this work, we have introduced InstructPart, a novel dataset containing part annotations for common household tasks instructions. We showed that even the most advanced vision-language models struggle with instructions that link specific affordances to the corresponding parts of an object. This highlights a significant gap in foundation models for robotics. With our dataset, we hope to enable future research that can pave the way for more natural human-robot interactions by allowing laypeople to effectively interact with assistive robots in their in-home environment through grounding high-level instructions in objects and, most importantly, object parts that can fulfill their needs.

Acknowledgments

This work has been funded in part by the Army Research Laboratory (ARL) under grants W911NF-23-2-0007 and W911NF-19-2-0146, and the Air Force Office of Scientific Research (AFOSR) under grants FA9550-18-1-0097 and FA9550-18-1-0251.

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023a. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023b. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1971–1978.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Deng, S.; Xu, X.; Wu, C.; Chen, K.; and Jia, K. 2021. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1778–1787.

Gadre, S. Y.; Ehsani, K.; and Song, S. 2021. Act the part: Learning interaction strategies for articulated object part discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15752–15761.

Geng, H.; Xu, H.; Zhao, C.; Xu, C.; Yi, L.; Huang, S.; and Wang, H. 2023. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7081–7091.

Gong, K.; Liang, X.; Zhang, D.; Shen, X.; and Lin, L. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 932–940.

He, J.; Yang, S.; Yang, S.; Kortylewski, A.; Yuan, X.; Chen, J.-N.; Liu, S.; Yang, C.; Yu, Q.; and Yuille, A. 2022. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, 128–145. Springer.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 108–124. Springer.

Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; and Fei-Fei, L. 2023. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*.

Jia, M.; Shi, M.; Sirotenko, M.; Cui, Y.; Cardie, C.; Hariharan, B.; Adam, H.; and Belongie, S. 2020. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 316–332. Springer.

Kaiser, M. S.; Al Mamun, S.; Mahmud, M.; and Tania, M. H. 2021. Healthcare robots to combat COVID-19. *COVID-19: Prediction, decision-making, and its impacts*, 83–97.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, C.; Ding, H.; and Jiang, X. 2023. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23592–23601.

Liu, F.; Liu, Y.; Kong, Y.; Xu, K.; Zhang, L.; Yin, B.; Hancke, G.; and Lau, R. 2023a. Referring image segmentation using text supervision. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, 22124–22134.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, R.; Liu, C.; Bai, Y.; and Yuille, A. L. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4185–4194.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Matheson, E.; Minto, R.; Zampieri, E. G.; Faccio, M.; and Rosati, G. 2019. Human–robot collaboration in manufacturing applications: A review. *Robotics*, 8(4): 100.

Mo, K.; Zhu, S.; Chang, A. X.; Yi, L.; Tripathi, S.; Guibas, L. J.; and Su, H. 2019. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 909–918.

Mogadala, A.; Kalimuthu, M.; and Klakow, D. 2021. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71: 1183–1317.

Myers, A.; Teo, C. L.; Fermüller, C.; and Aloimonos, Y. 2015. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 1374–1381. IEEE.

Nguyen, A.; Kanoulas, D.; Caldwell, D. G.; and Tsagarakis, N. G. 2017. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5908–5915. IEEE.

Ouyang, S.; Wang, H.; Xie, S.; Niu, Z.; Tong, R.; Chen, Y.-W.; and Lin, L. 2023. Slvit: Scale-wise language-guided vision transformer for referring image segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 1294–1302.

Pan, T.-Y.; Liu, Q.; Chao, W.-L.; and Price, B. 2023. Towards Open-World Segmentation of Parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15392–15401.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramanathan, V.; Kalia, A.; Petrovic, V.; Wen, Y.; Zheng, B.; Guo, B.; Wang, R.; Marquez, A.; Kovvuri, R.; Kadian, A.; et al. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7141–7151.

Roy, A.; and Todorovic, S. 2016. A multi-scale cnn for affordance segmentation in rgb images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 186–201. Springer.

Song, X.; Wang, P.; Zhou, D.; Zhu, R.; Guan, C.; Dai, Y.; Su, H.; Li, H.; and Yang, R. 2019. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5452–5462.

Sun, P.; Chen, S.; Zhu, C.; Xiao, F.; Luo, P.; Xie, S.; and Yan, Z. 2023. Going Denser with Open-Vocabulary Part Segmentation. *arXiv preprint arXiv:2305.11173*.

Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; and Ross, C. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, P.; Shen, X.; Lin, Z.; Cohen, S.; Price, B.; and Yuille, A. L. 2015. Joint object and part segmentation using deep learned potentials. In *Proceedings of the IEEE International Conference on Computer Vision*, 1573–1581.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*.

Xiang, F.; Qin, Y.; Mo, K.; Xia, Y.; Zhu, H.; Liu, F.; Liu, M.; Jiang, H.; Yuan, Y.; Wang, H.; et al. 2020. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11097–11107.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Yang, J.; Dong, Y.; Liu, S.; Li, B.; Wang, Z.; Jiang, C.; Tan, H.; Kang, J.; Zhang, Y.; Zhou, K.; et al. 2023a. Octopus: Embodied Vision-Language Programmer from Environmental Feedback. *arXiv preprint arXiv:2310.08588*.

Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.; and Gao, J. 2023b. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023c. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18155–18165.

Yi, L.; Huang, H.; Liu, D.; Kalogerakis, E.; Su, H.; and Guibas, L. 2018. Deep part induction from articulated object pairs. *arXiv preprint arXiv:1809.07417*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 69–85. Springer.

Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2023. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *arXiv preprint arXiv:2308.02487*.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. 2023a. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15116–15127.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Gao, J.; and Lee, Y. J. 2023b. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.