This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Grounding 3D Object Affordance with Language Instructions, Visual Observations and Interactions

He Zhu[1,2] Quyu Kong[2] Kechun Xu[1,2] Xunlong Xia[2] Bing Deng[2]
Jieping Ye[2] Rong Xiong[1*] Yue Wang[1]
[1]Zhejiang University [2]Alibaba Cloud

Abstract

Grounding 3D object affordance is a task that locates objects in 3D space where they can be manipulated, which links perception and action for embodied intelligence. For example, for an intelligent robot, it is necessary to accurately ground the affordance of an object and grasp it according to human instructions. In this paper, we introduce a novel task that grounds 3D object affordance based on language instructions, visual observations and interactions, which is inspired by cognitive science. We collect an Affordance Grounding dataset with Points, Images and Language instructions (AGPIL) to support the proposed task. In the 3D physical world, due to observation orientation, object rotation, or spatial occlusion, we can only get a partial observation of the object. So this dataset includes affordance estimations of objects from full-view, partial-view, and rotation-view perspectives. To accomplish this task, we propose LMAffordance3D, the first multi-modal, language-guided 3D affordance grounding network, which applies a vision-language model to fuse 2D and 3D spatial features with semantic features. Comprehensive experiments on AGPIL demonstrate the effectiveness and superiority of our method on this task, even in unseen experimental settings. Our project is available at https://sites.google.com/view/lmaffordance3d.

1. Introduction

Affordance is the property of an object that suggest how it can be used [10]. Grounding 3D object affordance aims to classify, detect, or segment the object's affordances in a scene with the input of images, videos, and point clouds, which plays an important role in robotics, human-object interactions (HOI), scene understanding, functionality understanding, and other fields [3, 14].

Although there are numerous studies [6, 8, 27, 30] focusing on grounding object affordances from 2D images or videos, the knowledge of affordances derived from 2D data remains challenging to apply in 3D space. Some works [5, 44] have been dedicated to using 3D data to ground object affordances. They learn from geometry to segment out affordance regions from point clouds. However, they exhibit limited generalization for grounding unseen object affordances and tend to show confusion when encountering similar objects. [47]. Actually, humans learn the affordances of a new object from demonstrations in images or videos, as well as from verbal instructions. It is a complex and comprehensive task that involves understanding of language instructions, geometric relationships, and scene information. Cognitive science research [12] also indicates that visual information and auditory language cues humans perceive can enhance their ability to use or manipulate new objects. Inspired by this, we introduce a novel task that grounds 3D object affordance based on language instructions, visual observations, and interactions.

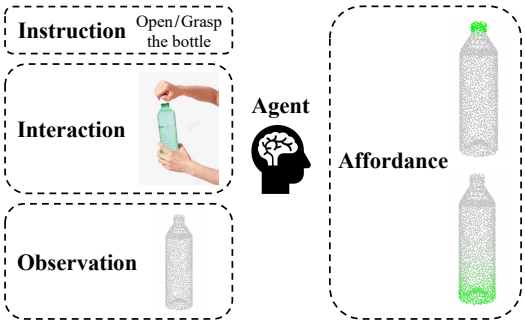


Figure 1. **Illustration for the affordance grounding task.** Inspired by cognitive science, when humans encounter a new object, they learn its affordance through language instructions, vision information from eyes, and human-machine interactions, thus obtaining its affordance.

In this paper, we present AGPIL, which is the first multi-modal and multi-view dataset for grounding 3D object affordance. For the full-view dataset, each frame of data includes a language instruction, an image of human-object interactions, and point clouds of the object's entire external surface. In the 3D physical world, due to observation ori-

*Corresponding author (rxiong@zju.edu.cn).

Dataset	Year	Input			Output	View	Number
		Image	Point cloud	Language			
PAD [26]	2021	✓	✗	✗	2D Segmentation	-	4002
CAD-120++ [11]	2022	✓	✗	✗	2D Box	-	32327
AGD20K [27]	2022	✓	✗	✗	2D Heatmap	-	20061
3D AffordanceNet [5]	2021	✗	✓	✗	3D Heatmap	F, P, R	56307
GAPartNet [9]	2023	✗	✓	✗	3D Segmentation & Pose	P, R	8489
PIAD [47]	2023	✓	✓	✗	3D Heatmap	F	7012
LASO [19]	2024	✗	✓	✓	3D Heatmap	F	19751
AGPIL (Ours)	2024	✓	✓	✓	3D Heatmap	F, P, R	30972

Table 1. **Comparion of affordance grounding dataset.** Here we mainly compare the input, output, view, and dataset scale for different datasets. F, P, and R respectively refer to full-view, partial-view, and rotation-view.

entation, object rotation, or spatial occlusion, we can only acquire partial scans of the object. So we also build the partial-view and rotation-view datasets. The partial-view dataset includes only one side or part of the object's point clouds, while the rotation-view dataset contains part of the object's point clouds with random rotation along the x, y, and z axes. In addition, each view of datasets has two settings: seen and unseen, which can comprehensively test the model's generalization performance.

To address the task of grounding 3D object affordance, we propose LMAffordance3D, which is an end-to-end multi-modal, language-guided 3D affordance grounding framework based on the vision-language model (VLM). And This framework operates as a one-stage method. The key idea is to integrate 2D and 3D spatial features and project them into a text space via multilayer perceptron (MLP). We then concatenate all the features as input for the vision-language model. Using a decoder, we fuse the output from the vision-language model with the 2D and 3D spatial features, subsequently employing a segmentation head to delineate affordance regions.

In summary, our contributions are mainly two-fold:

- We introduce a novel task that grounds 3D object affordance based on language instructions, visual observations, and interactions. And we present AGPIL which is composed of 30972 point-image-text pairs, covering 17 affordance categories and 23 object classes. It has full-view, partial-view, and rotation-view in seen and unseen settings. To the best of our knowledge, this is the first multi-modal and multi-view 3D affordance dataset with well-defined probabilistic affordance score annotations.

- We establish a comprehensive benchmark based on AGPIL for three different views in two settings. And we propose LMAffordance3D as a baseline, which is a novel end-to-end multi-modal, language-guided 3D affordance grounding framework based on VLM. The results show the effectiveness and better generalization capabilities of our fusion framework.

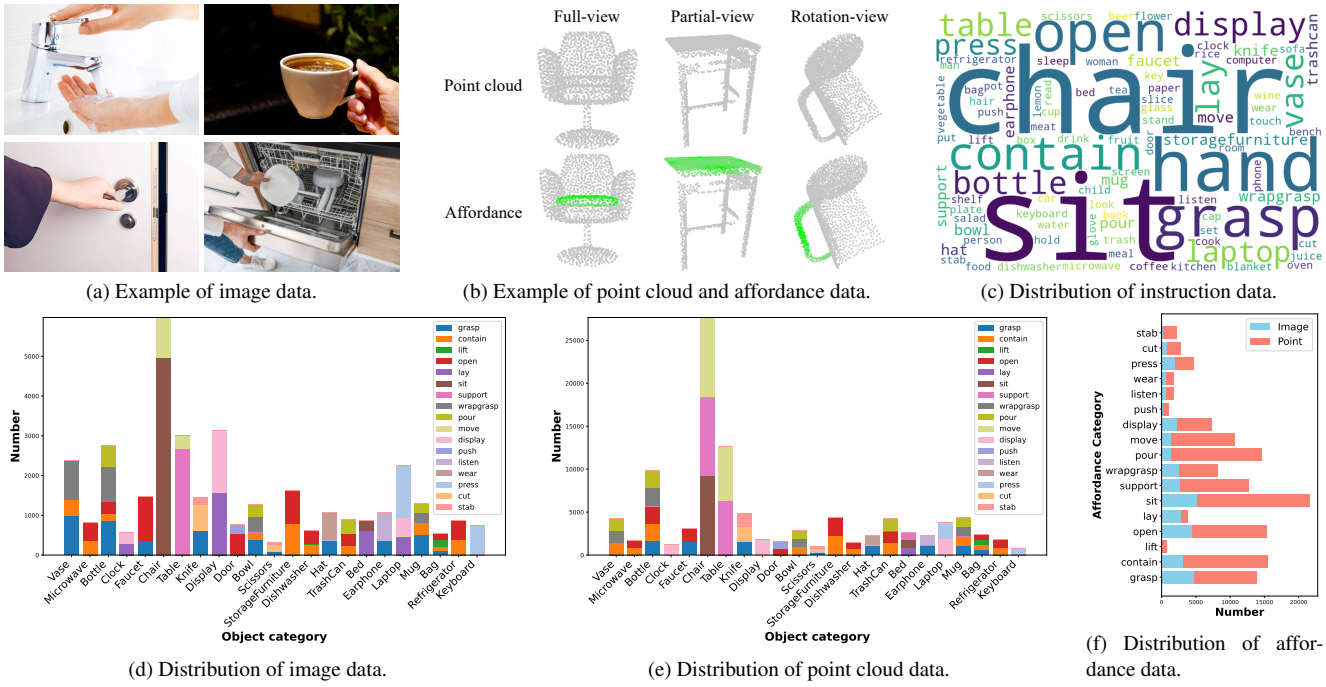
2. Related Works

2.1. Vision-Language Models

Due to the remarkable performance of large language models (LLMs) such as LlaMA [7, 40, 41], Qwen [1] and GPT [36] in natural language processing tasks, many scholars have extended them to the visual domain, giving rise to a series of large vision-language models (VLMs), such as LLaVA [20–22], Qwen-VL [2], GPT-4o [15], and so on. Leveraging the prior knowledge learned by these large models, they have been applied in 3D visual grounding task. LLM-Grounder [46] utilizes an LLM to decompose complex natural language queries into semantic parts and employs a visual grounding tool to identify objects in a 3D scene. VLM-Grounder [45] utilizes a VLM based on image sequences and language to achieve zero-shot 3D visual grounding. Unlike previous work on object-level grounding tasks, we define our task as utilizing multi-modal data from point clouds and images to ground 3D object affordances based on human language instructions, which is a part-level 3D grounding task.

2.2. Affordance Grounding

Affordance grounding is to identify the locations where interactions are possible, which is important for human-object interaction. AffordanceLLM [34] uses LLaVA [21] for 2D visual affordance grounding based and it demonstrates the importance of 3D information in affordance grounding. However, it obtains a 2D heatmap of affordance based on the image and depth map, which is not convenient to map it to 3D space for human or robot operation. Several studies have focused on 3D affordance. For instance, 3D AffordanceNet [5] predicts a 3D heatmap based on point clouds. And some methods leverage multi-modal data to predict 3D affordance, such as IAG [47], OpenAD [32], and LASO [19]. However, IAG relies on the 2D detection bounding boxes of interactive objects, which limits the task's scalability. OpenAD uses simple labels for text input, lacking support for context-rich language instructions. Moreover, LASO overlooks challenges posed by incom-



(a) Example of image data. (b) Example of point cloud and affordance data. (c) Distribution of instruction data.

(d) Distribution of image data. (e) Distribution of point cloud data. (f) Distribution of affordance data.

Figure 2. **Examples and statistics of the AGPIL dataset.** Figures (a) and (b) are examples of images, point clouds, and a certain affordance that we randomly selected from the dataset. Figure (c) shows a word cloud generated according to the frequency of each word appearing in the language instructions. Figures (d) and (e) respectively show the distribution of affordances corresponding to different objects in image and point cloud data. The horizontal axis represents the types of objects, and the vertical axis represents the quantity. Different colors indicate different affordances. Figure (f) illustrates the distribution of image and point cloud data corresponding to each affordance. It indicates that images and point clouds are not a one-to-one match, as a single image may correspond to multiple objects.

plete point clouds and object rotations in the real world.

In order to improve the generalization of grounding 3D affordance of unseen objects, various approaches can be adopted, such as contrastive learning [43], zero-shot learning [16], reinforcement learning [31], supervised learning based on vision-language models [34] and so on. For zero-shot learning, there are issues such as incorrect HOI prompt generation and low granularity. And contrastive learning is sensitive to data augmentation strategies and is limited by similarity metric of comparing different things. Reinforcement learning typically requires a large amount of interaction data, which is time-consuming for training. In this paper, we use a vision-language model to ground 3D object affordance in full-view, partial-view and rotation-view, which can be easily extended to other new scenarios in the real world.

3. Dataset

3.1. Collection and Generation

We collect the 3D affordance grounding dataset (AGPIL), which contains point clouds, images and language instructions. The point clouds are mainly collected from 3D AffordanceNet [5]. According to the differences in the point

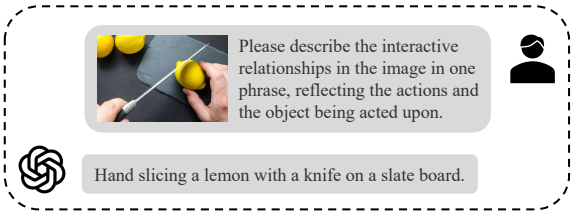


Figure 3. **Language instructions generation.** In this example, the affordance is "cut" and the object category is "knife".

clouds, the dataset has three parts: **full-view**, **partial-view**, and **rotation-view**. The term "full-view" indicates that the data include the point clouds of all the surfaces of an object. Both "partial-view" and "rotation-view" signify that the data include only part of the point clouds of an object's surfaces, with the distinction being whether or not there is rotation. And the rotation of the object along the x, y, and z axes is random. The images are mainly collected from AGD20K [28] and PIAD [47]. The image should depict the interaction between objects and humans, corresponding to the affordance of the object in the point clouds. Unlike PIAD, our dataset does not contain 2D bounding boxes of the interactive subject and object for each image. To improve efficiency, we use GPT-4o [15] combined with an im-

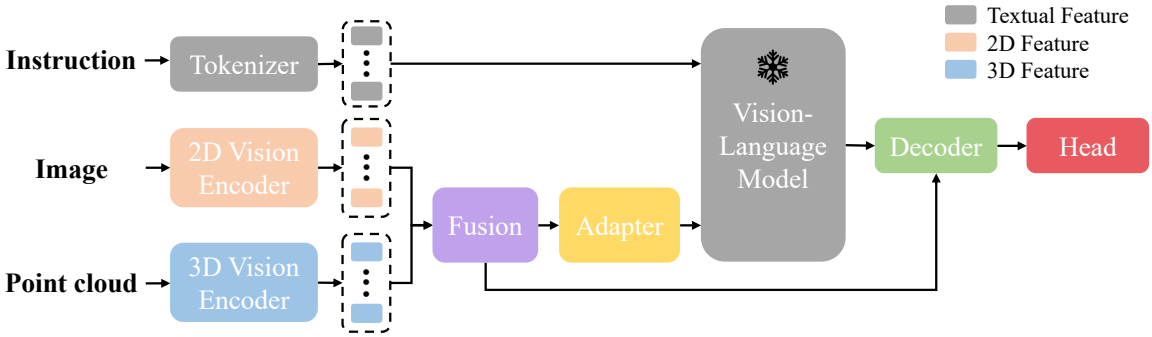


Figure 4. **Method.** The structure of the proposed LMAffordance3D model, which consists of four major components: 1) a vision encoder that processes multi-modal data, including images and point clouds, to encode and fuse the 2D and 3D features; 2) a vision-language model and its associated component (tokenizer and adapter) that takes in the instruction token, 2D and 3D vision token for fusion; 3) a decoder that uses 2D and 3D spatial features as query, instructional features as key and semantic feature as value to predict the affordance feature; 4) a head for segmenting and grounding 3D object affordance.

age and a carefully designed prompt to generate each language instruction. The prompt is "Please describe the interactive relationships in the image in one phrase, reflecting the actions and the object being acted upon". To ensure accuracy, we carefully score the generated texts based on consistency, word choice accuracy, semantic completeness and conciseness, and sentence structure diversity. For texts with lower scores, we regenerate or manually refine them. The diagram which illustrates the generation of language instructions is shown in the Fig. 3.

3.2. Annotation

For the affordance of objects, we use the segmented annotations from 3D AffordanceNet [5], which enables part-level prediction. And we render a 3D heatmap based on confidence scores with different colors. Each annotation is a (2048, 17) matrix, where 2048 represents the number of points in the point cloud, and 17 represents the number of affordance types. Each element in the matrix indicates the probability of a point providing a specific affordance, with a range from 0 to 1. If all elements in a column are 0.0, it indicates that the object does not have the corresponding affordance. In our task, we only use the 3D heatmap of a single type of affordance corresponding to the image and language instruction in the point cloud annotation for training. The prediction is a (2048, 1) matrix during inference. In this way, we can predict different affordances in the point clouds with different images and language instructions.

3.3. Statistic Analysis

As Tab. 1 shows, we compare the open affordance dataset with AGPIL. Our total dataset consists of 41,628 point clouds, 30,972 images, and 30,972 language instructions, encompassing 23 object categories and 17 types of affordances. Each object has an average of 2.4 functional parts. In particular, objects in the images and point clouds are not captured from the same scene and they are paired based on object categories. Our dataset has two settings: **Seen** and **Unseen**. In the seen dataset, the categories of objects and affordances in the training and test sets are consistent, while in the unseen dataset, they are inconsistent. More examples and statistics about the dataset are illustrated in Fig. 2.

4. Method

4.1. Overview

In this section, we propose LMAffordance3D, a framework that learns from language instructions, visual observations and interactions for grounding 3D object affordance. As illustrated in Fig. 4, LMAffordance3D is composed of four major components: 1) a vision encoder that processes multi-modal data, including images and point clouds, to encode and fuse the 2D and 3D features, where an RGB image contains information about the colors, scenes, and interactions between objects, while point clouds contain different information about the shape, size, and geometry of objects. 2) a vision-language model and its associated components (tokenizer and adapter) that take in the instruction token, 2D and 3D vision token for fusion; 3) a decoder that uses 2D and 3D spatial features as query, instructional features as key and semantic feature as value to predict the affordance feature; 4) a head for segmenting and grounding 3D object affordance. We will introduce the components of LMAffordance3D in Sec. 4.2. And we present the loss function in Sec. 4.3.

4.2. LMAffordance3D

Vision encoder. In the vision-language community [17, 18, 20–22, 48], the most common method for vision-language alignment is to use a pre-trained CLIP [35] model and project image features into the semantic space. How-

Setting	Metrics	Full-view					Partial-view					Rotation-view				
		3D A.N.	IAG	O.AD	P.R.	Ours	3D A.N.	IAG	O.AD	P.R.	Ours	3D A.N.	IAG	O.AD	P.R.	Ours
Seen	AUC ↑	0.8067	0.8485	0.8584	0.8766	0.8895	0.7609	0.8085	0.8145	0.8212	0.8478	0.5954	0.6791	0.7325	0.7564	0.7823
	IOU ↑	0.1341	0.2051	0.1993	0.2002	0.2123	0.1077	0.1454	0.1562	0.1613	0.1755	0.0452	0.0804	0.0911	0.1012	0.1161
	SIM ↑	0.4833	0.5450	0.5866	0.5953	0.6102	0.4785	0.5362	0.5652	0.5796	0.5928	0.3892	0.4359	0.4783	0.4984	0.5191
	MAE ↓	0.1141	0.0980	0.0944	0.0905	0.0816	0.1262	0.1042	0.1006	0.0987	0.0921	0.1648	0.1345	0.1236	0.1302	0.1182
Unseen	AUC ↑	0.6652	0.7184	0.7375	0.7545	0.7741	0.6419	0.6718	0.7014	0.7211	0.7602	0.5821	0.5994	0.5989	0.6088	0.6303
	IOU ↑	0.0522	0.0795	0.0781	0.0813	0.0903	0.0425	0.0612	0.0671	0.0682	0.0724	0.0277	0.0311	0.0356	0.0379	0.0415
	SIM ↑	0.3012	0.3520	0.3704	0.3849	0.4089	0.3347	0.3568	0.3766	0.3894	0.4144	0.3552	0.3623	0.3622	0.3631	0.3842
	MAE ↓	0.1556	0.1270	0.1193	0.1178	0.0945	0.1601	0.1518	0.1428	0.1262	0.1183	0.1744	0.1709	0.1506	0.1503	0.1398

Table 2. **Benchmark.** We present the overall results of all comparative methods on different views and settings. "3D A.N." refers to "3D AffordanceNet", "O.AD" refers to "OpenAD", and "P.R." refers to "PointRefer"

ever, we need to process images $I_I \in \mathbb{R}^{B\times 3\times H\times W}$ and point clouds $I_P \in \mathbb{R}^{B\times 3\times 2048}$ data for fusion in this task. And the large flops and parameter size of the CLIP models pose challenges for deployment in robotics. So we design a multi-modal vision encoder, in which the 2D vision encoder uses ResNet18 [13] to extract 2D features $F_{2D} \in \mathbb{R}^{B\times C_I\times H\times W}$, and the 3D vision encoder uses PointNet++ [33] to extract 3D features $F_{3D} \in \mathbb{R}^{B\times C_P\times N_P}$. Then fusion module uses MLP and self-attention to obtain multi-modal spatial features $F_S \in \mathbb{R}^{B\times N_S\times C_S}$.

Vision-language model. Vision-language model as the brain of agent is used for comprehending natural language instructions, integrating semantic features and spatial features to obtain affordance features. Specifically, we choose LLaVA-7B as the backbone networks for the vision-language model. Given the action instruction, we apply the tokenizer to convert the instruction into textual feature $F_T \in \mathbb{R}^{B\times N_L\times C_L}$. We use an adapter, which consists of two linear layers and an activation layer. It takes spatial features $F_S \in \mathbb{R}^{B\times N_S\times C_S}$ as input and maps them to a semantic space while making an increase in the channel dimension to obtain projected features $F_{SP} \in \mathbb{R}^{B\times N_S\times C_L}$. These projected features F_{SP} are then concatenated with textual features F_T as the input of vision-language model. And the output of vision-language model is its hidden state.

Decoder. In order to fuse all kinds of feature to get the affordance features, we design a decoder based on cross-attention. We split the output of vision-language model into instructional features and semantic features. Then we use spatial features as the query, instructional features as the key, and semantic features as the value, and obtain the affordance feature $F_A \in \mathbb{R}^{B\times N_A\times C_A}$ with the decoder.

Head. Here we upsample the affordance feature F_A for propagation and use a head composed of two linear layers, a activation layer, a batch normalization layer and a sigmoid layer to obtain 3D object affordance $O \in \mathbb{R}^{B\times 2048\times 1}$.

4.3. Loss Function

We predict the probability that each point cloud belongs to the corresponding affordance, ranging from 0 to 1. In order to supervise the point-wise 3D heatmap on point clouds, we use a weighted sum of focal loss [38] and dice loss [29] as the final loss function:

$$Loss = \omega_f L_f + \omega_d L_d \tag{1}$$

where ω_f and ω_d represent the weight of the focal loss and dice loss respectively.

5. Experiments

5.1. Benchmark

Metrics. In accordance with PIAD [47], we use four metrics to build the AGPIL benchmark: AUC [23], aIoU [37], SIM [39], and MAE [42]. The AUC computes the area under the ROC curve to assess the significance map on a 3D point cloud heatmap. The aIOU is the average intersection over union, which is used to evaluate the proximity between the affordance prediction and the ground truth. SIM measures the similarity between the predicted map and the ground truth map. MAE is the mean absolute error between the predicted probabilities and the ground truth. For AUC, aIoU, and SIM, higher is better, while for MAE, lower is better.

Baselines. Since there are no prior works on grounding 3D object affordance based on language, images, and point clouds, we choose 3D AffordanceNet [33], IAG [47], OpenAD [32], and PointRefer [19] as baselines. We evaluate them in both seen and unseen settings, as well as under different perspectives including full-view, partial-view, and rotation-view. This allows us to comprehensively understand the model's generalization ability and performance across different object categories and viewpoints.

Training and testing setup. Our model is implemented using PyTorch and all training and testing experiments are conducted on a single V100 GPU. We train the model with the AdamW optimizer [24] and a cosine learning rate scheduler [25]. The initial learning rate is 1e-4 and the weight decay is set to 0.06. We train the models with the first 2000 iteration steps for warm-up [13]. The number of training epochs is set to 20 and the training batch size is set to 6. The image 2D vision encoder ResNet18 uses parameters pre-trained on ImageNet [4], while the point cloud 3D vision

F	Metrics	grasp	press	stab	open	lay	sit	cut	cont.	disp.	wrap.	supp.	push	listen	wear	move	lift	pour
Seen	AUC↑	0.859	0.898	0.997	0.932	0.935	0.957	0.945	0.871	0.907	0.689	0.846	0.801	0.910	0.733	0.776	0.973	0.945
	aIOU↑	0.203	0.136	0.354	0.239	0.294	0.358	0.156	0.199	0.293	0.060	0.111	0.029	0.160	0.052	0.098	0.407	0.232
	SIM↑	0.650	0.401	0.651	0.398	0.693	0.719	0.681	0.567	0.655	0.645	0.716	0.570	0.683	0.564	0.612	0.497	0.597
	MAE↓	0.098	0.059	0.017	0.050	0.086	0.065	0.056	0.089	0.095	0.128	0.097	0.082	0.087	0.138	0.135	0.020	0.075
Unseen	AUC↑	0.363	0.854	0.932	0.894	0.788	0.719	0.921	0.866	0.681	0.556	-	-	-	-	-	-	-
	aIOU↑	0.008	0.038	0.053	0.103	0.122	0.109	0.104	0.150	0.047	0.025	-	-	-	-	-	-	-
	SIM↑	0.319	0.304	0.356	0.215	0.427	0.376	0.533	0.522	0.277	0.562	-	-	-	-	-	-	-
	MAE↓	0.205	0.065	0.053	0.058	0.127	0.110	0.059	0.098	0.111	0.134	-	-	-	-	-	-	-
P	**Metrics**	**grasp**	**press**	**stab**	**open**	**lay**	**sit**	**cut**	**cont.**	**disp.**	**wrap.**	**supp.**	**push**	**listen**	**wear**	**move**	**lift**	**pour**
Seen	AUC↑	0.818	0.836	0.987	0.863	0.791	0.952	0.909	0.806	0.857	0.647	0.832	0.781	0.858	0.709	0.811	0.963	0.907
	aIOU↑	0.151	0.127	0.377	0.212	0.134	0.313	0.155	0.128	0.264	0.041	0.086	0.029	0.103	0.036	0.095	0.393	0.214
	SIM↑	0.608	0.411	0.522	0.404	0.465	0.702	0.751	0.498	0.602	0.744	0.731	0.549	0.510	0.657	0.622	0.500	0.637
	MAE↓	0.115	0.078	0.035	0.074	0.101	0.058	0.065	0.103	0.117	0.137	0.102	0.110	0.099	0.135	0.119	0.063	0.089
Unseen	AUC↑	0.358	0.748	0.939	0.880	0.801	0.742	0.879	0.811	0.577	0.604	-	-	-	-	-	-	-
	aIOU↑	0.004	0.028	0.057	0.100	0.141	0.093	0.047	0.092	0.043	0.043	-	-	-	-	-	-	-
	SIM↑	0.358	0.254	0.429	0.271	0.451	0.362	0.450	0.514	0.232	0.621	-	-	-	-	-	-	-
	MAE↓	0.092	0.106	0.088	0.093	0.126	0.120	0.092	0.116	0.188	0.153	-	-	-	-	-	-	-
R	**Metrics**	**grasp**	**press**	**stab**	**open**	**lay**	**sit**	**cut**	**cont.**	**disp.**	**wrap.**	**supp.**	**push**	**listen**	**wear**	**move**	**lift**	**pour**
Seen	AUC↑	0.758	0.726	0.837	0.785	0.820	0.868	0.835	0.683	0.778	0.608	0.752	0.861	0.776	0.586	0.765	0.777	0.829
	aIOU↑	0.125	0.042	0.178	0.108	0.106	0.190	0.102	0.041	0.190	0.032	0.069	0.022	0.097	0.018	0.055	0.246	0.145
	SIM↑	0.594	0.259	0.562	0.294	0.461	0.537	0.650	0.432	0.567	0.706	0.659	0.594	0.543	0.648	0.599	0.466	0.541
	MAE↓	0.116	0.117	0.065	0.106	0.161	0.098	0.103	0.135	0.165	0.138	0.136	0.081	0.103	0.128	0.136	0.057	0.119
Unseen	AUC↑	0.788	0.643	0.992	0.565	0.760	0.602	0.925	0.576	0.533	0.583	-	-	-	-	-	-	-
	aIOU↑	0.029	0.024	0.422	0.006	0.110	0.045	0.123	0.027	0.055	0.031	-	-	-	-	-	-	-
	SIM↑	0.589	0.208	0.587	0.168	0.431	0.337	0.628	0.485	0.249	0.656	-	-	-	-	-	-	-
	MAE↓	0.101	0.144	0.028	0.082	0.173	0.175	0.079	0.147	0.233	0.152	-	-	-	-	-	-	-

Table 3. **Evaluation metrics for different affordance with LMAffordance3D.** We compare on seen and unseen setting in different views, where F represents full-view, P represents partial-view, and R represents rotation-view. "cont." refers to "contain", "disp." refers to "display", "wrap." refers to "wrapgrasp", and "supp." refers to "support". "-" indicates that there is no corresponding affordance category in the test set, hence no metrics.

encoder is trained from scratch. The vision-language model uses llava-v1.6-vicuna-7b, and its parameters are frozen. During training, automatic mixed precision (AMP) is introduced to accelerate training and reduce memory usage. Furthermore, given that images and point clouds do not need to be strictly one-to-one paired, we pair images and point clouds online during training. In one training step, an image can be paired with two point clouds, effectively augmenting the training samples. During training, the loss of one image and all paired point clouds is accumulated, and gradients are calculated based on the accumulated loss. During inference, an image is paired with only one point cloud to predict the result of affordance.

5.2. Results and Analysis

Quantitative results. We build a benchmark based on the AGPIL dataset, as shown in the Tab. 2. It can be observed that our method outperforms the compared baselines, and there is a significant performance improvement in the unseen setting compared to the seen setting. This improvement can be attributed to the extensive world knowledge and prior information in the large language model. In the same experimental setup, the metrics for the full view are greater than those for the partial view, which in turn are greater than those for the rotation view. In the Tab. 3, we present the evaluation metrics for different affordances using LMAffordance3D in detail. In the seen setting, both the training and testing datasets include all affordance categories. However, in the unseen setting, the training dataset is limited to seven categories: support, push, listen, wear, move, lift, and pour. The test dataset, on the other hand, consists of ten categories: grasp, press, stab, open, lay, sit, cut, contain, display, and wrapgrasp. The results indicate that our model demonstrates good generalization performance.

Qualitative analysis. As shown in the Fig. 5, we compare the results of affordance prediction using different methods across various views and settings. When in full view, the model is able to observe the object from various angles, maximizing its grasp of the geometric and morphological information of the object. This setting can provide the model with the most complete information, allowing for comprehensive affordance analysis. In partial view, the model can only observe a partial view of the object, simulating real-world situations where full information about the object is inaccessible due to occlusion or viewpoint limitations. This experiment tests the model's robustness when information is incomplete. In rotation view, the object is rotated, and only partial views of the object are observed. It evaluates the model's robustness in handling object pose

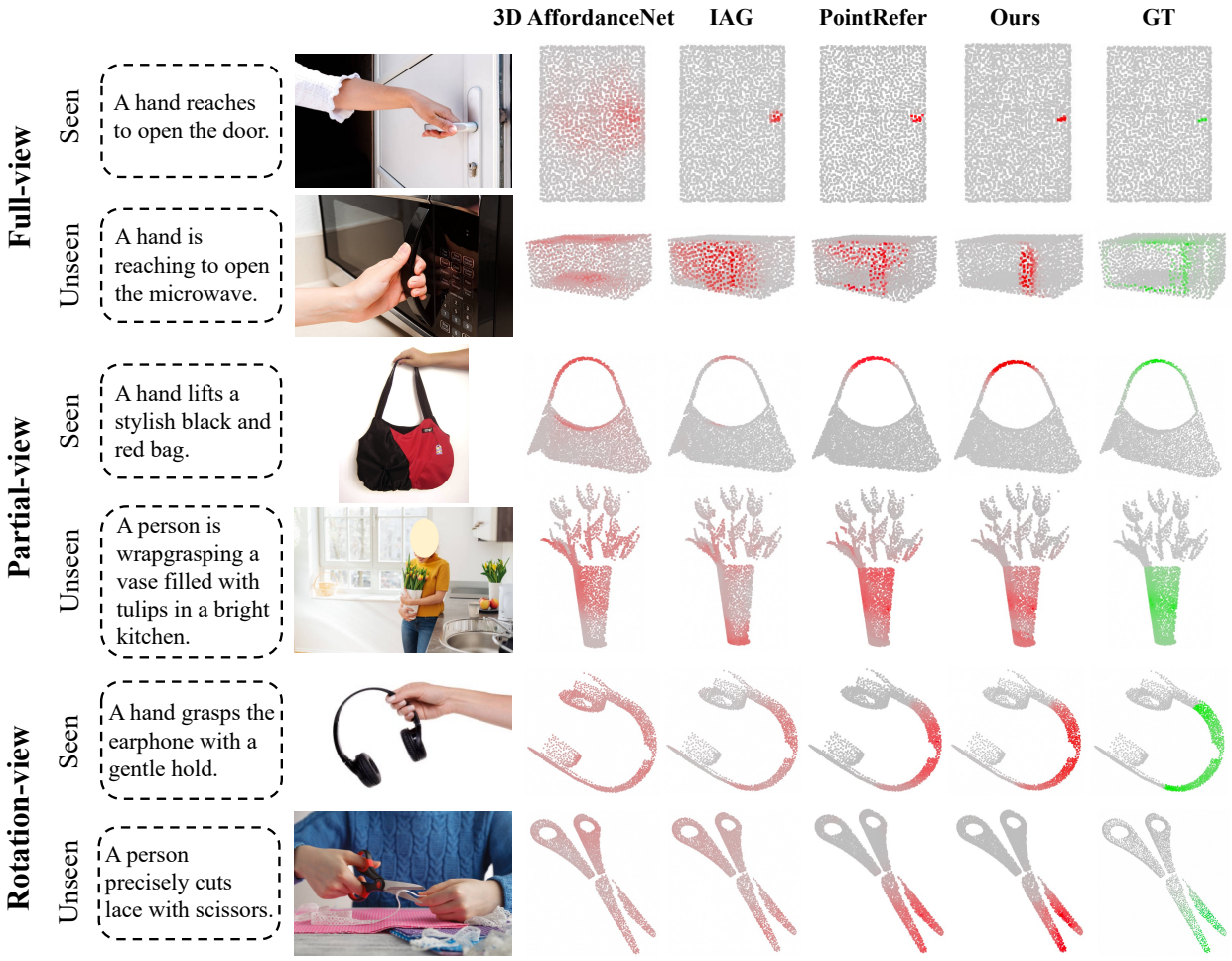


Figure 5. **Visualization.** We select several examples from the test set under different views and experimental settings, showcasing the model's inputs and outputs, and comparing them with the ground truth (GT).

rotations or changes in observer perspective. The figure clearly shows that our method achieves significantly better results, confirming the validity of our setting and demonstrating the superiority of our method.

5.3. Ablation Study

We conduct ablation experiments on the AGPIL dataset to evaluate the influence of interaction images and language instructions on the model. The results are shown in Tab. 4.

To assess whether the visual cues from RGB images affect affordance accuracy, we can find that the performance without image input decreases in both seen and unseen settings, with a significant drop in the unseen setting. This demonstrates the necessity of interactive image input for this task.

To explore the impact of different language instructions on the experiment, we select four different settings:

- **Full**: We use the content generated by GPT-4o as the instruction.
- **Action & Object**: We use an action and an object name, that is, a verb and a noun as an instruction, such as "grasp bottle".
- **Action**: The instruction includes only an action, for example, "grasp".
- **None**: An empty instruction.

We can observe that full instructions yield a higher performance compared to other simpler prompts. This demonstrates that full instructions contain more knowledge and information, which are more beneficial for the pre-trained large language model to understand the intent.

5.4. Generalization to Real World

To evaluate the generalization ability of our model in the real world, we collect some data from everyday life for testing. And the language instructions and affordance categories are unseen to the model, which poses a greater challenge. As illustrated in the Fig. 6, the inference results

View	Setting	Image	Instruction	AUC↑	aIOU↑	SIM↑	MAE↓
Full-view	Seen	✓	Full	**0.8895**	**0.2123**	**0.6102**	**0.0816**
		✗	Full	0.8827	0.2037	0.6045	0.0881
		✓	Action & Object	0.8829	0.2057	0.6058	0.0829
		✓	Action	0.8813	0.2046	0.6034	0.0842
		✓	None	0.8263	0.1705	0.5394	0.1018
	Unseen	✓	Full	**0.7741**	**0.0903**	**0.4089**	**0.0945**
		✗	Full	0.7604	0.0856	0.3990	0.1213
		✓	Action & Object	0.7712	0.0783	0.4008	0.0968
		✓	Action	0.7667	0.0774	0.3980	0.0997
		✓	None	0.6741	0.0371	0.3036	0.1282
Partial-view	Seen	✓	Full	**0.8478**	**0.1755**	**0.5928**	**0.0921**
		✗	Full	0.8383	0.1642	0.5854	0.0972
		✓	Action & Object	0.8457	0.1721	0.5882	0.0924
		✓	Action	0.8425	0.1706	0.5866	0.0937
		✓	None	0.8281	0.1512	0.5702	0.0996
	Unseen	✓	Full	**0.7602**	**0.0724**	**0.4144**	**0.1183**
		✗	Full	0.7426	0.0701	0.4022	0.1311
		✓	Action & Object	0.7600	0.0692	0.4096	0.1216
		✓	Action	0.7587	0.0598	0.4070	0.1198
		✓	None	0.6513	0.0475	0.3412	0.1566
Rotation-view	Seen	✓	Full	**0.7823**	0.1161	**0.5191**	**0.1182**
		✗	Full	0.7732	0.1063	0.5067	0.1214
		✓	Action & Object	0.7742	**0.1163**	0.5188	0.1238
		✓	Action	0.7731	0.1159	0.5155	0.1189
		✓	None	0.7052	0.085	0.4657	0.1394
	Unseen	✓	Full	**0.6303**	**0.0415**	**0.3842**	0.1398
		✗	Full	0.6126	0.0401	0.3709	0.1512
		✓	Action & Object	0.6283	0.0413	0.3794	**0.1390**
		✓	Action	0.6277	0.0412	0.3782	0.1403
		✓	None	0.5899	0.0369	0.3702	0.1489

Table 4. **Ablation Study.** We investigate the impact of interaction images and language instructions on model performance relative to the baseline, with the best results highlighted in **bold**.

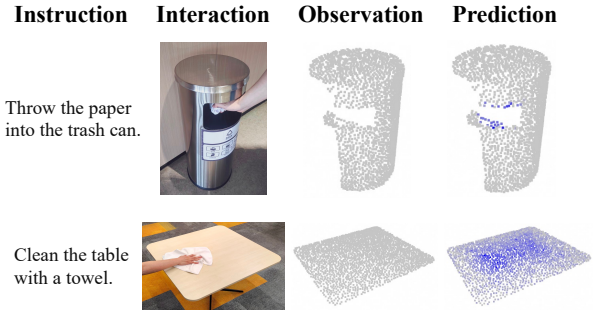


Figure 6. **Real world.** We provide two examples in which the verbs "throw" and "clean" in the language instructions are not in our dataset and are unseen by the model.

indicate that the model effectively leverages the strengths of vision-language models. By utilizing the prior world knowledge learned during the pre-training process and fine-tuning it on our dataset, the model can be easily transferred to other new scenarios in real world, achieving a complex and comprehensive understanding of semantics and spatial relationships.

6. Conclusion

In this paper, we introduce the task of grounding 3D object affordance based on language instructions, visual observations, and interactions. We present the first 3D affordance dataset combining images, point clouds, and language instructions, covering full-view, partial-view, and rotation-view. And inspired by cognitive science, we propose LMAffordance3D which leverages a vision-language framework to fuse 2D, 3D spatial features with semantic features, achieving part-level affordance estimation. We conduct experiments in both seen and unseen settings, and the results demonstrate the effectiveness and better generalization capabilities of our approach.

In future work, it will be interesting to explore model quantization and compression, and deploy them on robots to achieve real-time inference, combined with planning and control modules to accomplish tasks such as manipulation and operation. Moreover, since the objects in our dataset are all rigid, investigating affordance prediction for flexible objects is another promising direction for future research.

Acknowledgment

This work was supported in part by the Joint Funds of the National Natural Science Foundation of China under Grant U24A20128, and in part by the National Nature Science Foundation of China under Grant 62173293.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2

[3] Dongpan Chen, Dehui Kong, Jinghua Li, Shaofan Wang, and Baocai Yin. A survey of visual affordance recognition based on deep learning. *IEEE Transactions on Big Data*, 2023. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[5] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1778–1787, 2021. 1, 2, 3, 4

[6] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018. 1

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[8] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018. 1

[9] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. 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*, pages 7081–7091, 2023. 2

[10] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977. 1

[11] Yamin Han, Tao Zhuo, Peng Zhang, Wei Huang, Yufei Zha, Yanning Zhang, and Mohan Kankanhalli. One-shot video graph generation for explainable action reasoning. *Neurocomputing*, 488:212–225, 2022. 2

[12] Rex Hartson and Pardha S Pyla. *The UX book: Agile UX design for a quality user experience*. Morgan Kaufmann, 2018. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 1

[15] Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*, 2024. 2, 3

[16] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Zero-shot learning for the primitives of 3d affordance in general objects. *arXiv preprint arXiv:2401.12978*, 2024. 3

[17] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 4

[19] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024. 2, 5

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 4

[21] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 2, 4

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 4

[23] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008. 5

[24] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[25] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[26] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot affordance detection. *arXiv preprint arXiv:2106.14747*, 2021. 2

[27] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022. 1, 2

[28] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *International Journal of Computer Vision*, 132(6):1945–1969, 2024. 3

[29] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 5

[30] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 1

[31] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. *Advances in Neural Information Processing Systems*, 33:2005–2015, 2020. 3

[32] Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5692–5698. IEEE, 2023. 2, 5

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5

[34] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. *arXiv preprint arXiv:2401.06341*, 2024. 2, 3

[35] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[37] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016. 5

[38] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 5

[39] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991. 5

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[42] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005. 5

[43] Ruihai Wu, Kai Cheng, Yan Zhao, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[44] Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. *arXiv preprint arXiv:2202.13519*, 2022. 1

[45] Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024. 2

[46] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701. IEEE, 2024. 2

[47] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10905–10915, 2023. 1, 2, 3, 5

[48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 4