

GREAT: Geometry-Intention Collaborative Inference for Open-Vocabulary 3D Object Affordance Grounding

Yawen Shao¹, Wei Zhai^{1,†}, Yuhang Yang¹, Hongchen Luo², Yang Cao¹, Zheng-Jun Zha¹

¹ MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China ² Northeastern University

{shaoyawen@mail., wzhai056@, yyuhang@mail.}ustc.edu.cn
luohongchen@ise.neu.edu.cn {forrest@, zhazj@}ustc.edu.cn

Abstract

*Open-Vocabulary 3D object affordance grounding aims to anticipate “action possibilities” regions on 3D objects with arbitrary instructions, which is crucial for robots to generically perceive real scenarios and respond to operational changes. Existing methods focus on combining images or languages that depict interactions with 3D geometries to introduce external interaction priors. However, they are still vulnerable to a limited semantic space by failing to leverage implied invariant geometries and potential interaction intentions. Normally, humans address complex tasks through multi-step reasoning and respond to diverse situations by leveraging associative and analogical thinking. In light of this, we propose **GREAT** (**GeometRy-intEntion collAboraTive inference**) for Open-Vocabulary 3D Object Affordance Grounding, a novel framework that mines the object invariant geometry attributes and performs analogically reason in potential interaction scenarios to form affordance knowledge, fully combining the knowledge with both geometries and visual contents to ground 3D object affordance. Besides, we introduce the **Point Image Affordance Dataset v2 (PIAdv2)**, the largest 3D object affordance dataset at present to support the task. Extensive experiments demonstrate the effectiveness and superiority of GREAT. The code and dataset are available at <https://yawen-shao.github.io/GREAT/>.*

1. Introduction

Open-Vocabulary 3D object affordance grounding aims to locate “action possibilities” on objects [8, 14], both for seen and unseen scenarios, identifying specific regions on objects that support certain interactions. This bridges visual perception and physical manipulation for embodied agents and possesses bountiful application scenarios, e.g. robot

†Corresponding Author.

manipulation [12, 19, 40], scene understanding [11, 27], action anticipation [35, 66, 67] and imitation learning [13, 23].

Recently, most existing methods [2, 6, 17] establish explicit mappings between semantic affordance categories and geometric structures, restricted to predefined seen categories and fail to ground object affordance out of the training categories. Thus, some studies [25, 41, 50, 56, 58, 64] explore grounding object affordance through additional instructions, encompassing combining images or languages that depict interactions with 3D geometries to introduce external interaction priors, and mitigate the generalization gap lead by affordance diversity. Despite their remarkable progress, they are still vulnerable to a limited semantic space by failing to leverage implied invariant geometries among objects with the same affordance, as well as potential correlations among distinct interactions of the same object. As shown in Fig. 1 (a), current paradigms ground the 3D object affordance by aligning the object’s geometric features and instruction modalities, working well in the seen partition. However, such a paradigm relies excessively on the data fed into the model, when testing unseen affordance like “pour” that does not appear in the training set, the model endeavors to categorize it as the seen category “grasp”, shown in Fig. 1 (b).

In cognitive science, studies [7, 16, 62] have shown that humans tackle complex tasks through multi-step reasoning and respond to diverse situations by employing associative and analogical thinking. As demonstrated in Fig. 1 (c), when observing a water pouring scenario, humans employ multi-step reasoning to infer multiple potential interaction procedures and identify geometric properties of objects, indexing relevant knowledge from their brains. Analogous to this procedure, we leverage Multi-modal Large Language Models (MLLMs) [4, 22, 36, 49] to simulate prior knowledge, encompassing visual, linguistic, and other modalities, lifting their superior reasoning and generalization abilities to visual tasks [24, 42]. However, the dynamic and di-

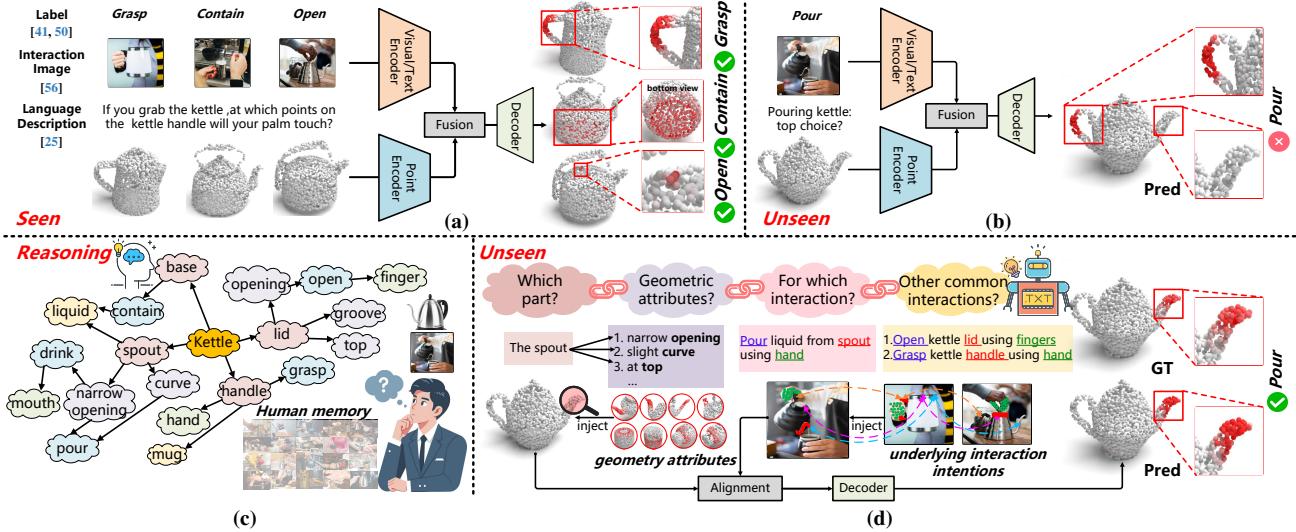


Figure 1. **Difference and Motivation.** (a) Object affordance grounding on seen setting. (b) Open-Vocabulary Affordance Grounding (OVAG) with previous paradigms. (c) Observing interaction images, people engage in brainstorming through memory representations, drawing on prior interaction experiences to perform analogical reasoning and infer appropriate actions. (d) OVAG with our geometry-intention collaborative inference with chain-of-thought, step-by-step identifies the interaction part, extracts geometric attributes, reasons about corresponding interaction and brainstorms underlying interaction intentions, jointly grounding the 3D object affordance.

verse nature of affordances makes it challenging to solely derive complex reasoning outcomes from MLLMs, to address this, a Chain-of-Thought (CoT) inference strategy that mirrors human reasoning processes is designed (Fig. 1 (d)). The strategy step-by-step identifies the interaction primitives, extracts geometric attributes, and reasons about the corresponding interaction actions and intentions, which enables the model to analyze interaction images from multiple perspectives, thereby eliminating the hallucination and ambiguities raised by MLLMs when reasoning interactions.

To achieve this, we present the **GREAT**, a novel framework that excavates implied affordance knowledge from interaction images and effectively integrates it with both point cloud and image representations to jointly ground object 3D affordance. Specifically, we first devise a Multi-Head Affordance Chain-of-Thought (MHACoT) reasoning strategy to infer implied invariant geometries and underlying interaction intentions from fine-tuning MLLM, then GREAT employs attention mechanisms to model the correlation between these primitives to form the affordance knowledge dictionary. Following this, we design the Cross-Modal Adaptive Fusion Module (CMAFM) to integrate knowledge into the point cloud features and directly fuse it with image features, leveraging the combined representation to accurately ground 3D object affordance.

Moreover, to further unleash the model’s capability, we extend the Point Image Affordance Dataset (**PIAD**) [56] to **PIADv2**, including 24 common affordances, 43 different object categories, over 15K interaction images from diverse

Table 1. **Comparison of related datasets.** Img.: interaction images. # Img.: number of images. # 3D.: number of 3D object instances. # Obj.: number of object categories. # Aff.: number of affordance categories.

Dataset	Img.	# Img.	# 3D.	# Obj.	# Aff.
3D AffordanceNet [6]	X	—	22949	23	18
PIAD [56]	✓	5162	7012	23	17
PIADv2 (Ours)	✓	15213	38889	43	24

scenes and 38K 3D objects with affordance annotations, triple interaction images and over **5 times** 3D instances compared to the PIAD (Tab. 1).

The contributions are summarized as follows:

- 1) We propose grounding 3D object affordance in an Open-Vocabulary fashion, which further reasons from interaction images, extrapolating from predefined sample space and generalize to unseen scenarios.
- 2) We present GREAT, a novel framework that designs a MHACoT fine-tuning and inference strategy that excavates geometric attributes and underlying interaction intention to support the object affordance reasoning.
- 3) We introduce the large-scale PIADv2 dataset, including 24 affordance and 43 object categories, 15K interaction images from diverse scenes and over 38K 3D objects with annotations. Extensive experiments on it demonstrate the effectiveness and superiority of GREAT.

2. Related Work

Affordance Grounding. Affordance grounding aims to locate the region of “action possibilities”, which is a link between robot perception and manipulation. Some works ground the object affordance from the 2D data *e.g.* images and videos [21, 31, 33, 34, 61, 63], while some works leverage natural language understanding to ground affordance regions in 2D data [3, 21, 30, 44]. However, robot manipulation usually requires 3D information of objects, and the 2D affordance grounding obtained from the above works make it difficult to infer the interaction position of 3D objects. With the presentation of several 3D object datasets [5, 38, 46], some works focus on the 3D object affordance grounding [6, 38, 39], which map semantic affordance to 3D object structure and fail to handle the open-vocabulary scenario.

Open-Vocabulary 3D Affordance Grounding. OVAG presents a substantial challenge, aiming to locate object affordance region in arbitrary scenario. Recently, some methods explore the possibility of OVAG. IAGNet [56] utilizes the 2D interaction semantics to guide the grounding of 3D object affordance. LASO [25] employs textual-conditioned affordance queries to isolate afforded segments and injects text clues to point features. OpenAD [41] and OpenKD [50] propose a text-point correlation method for affordance synonym substitutions by utilizing the power of large language encoder clip. Although above methods have made remarkable progress, they are still vulnerable to a limited training semantic space due to the presence of critical learnable parts in the framework. GREAT mitigates this limitation by leveraging geometry-intention collaborative inference with CoT to ground object affordance.

Chain-of-Thought Prompting with MLLMs. Chain-of-Thought (CoT) [51, 65] and its variants [20, 59, 60] are proposed to enhance the reasoning capabilities of Multi-modal Large Language Models (MLLMs), which guide the model through multiple logical steps. With the rapid development of MLLMs [4, 15, 68], vision-related [42, 53, 55] methods have made significant progress in collaborating CoT and MLLMs to get the desired results. Some methods explore object affordance at task driven object detection [48], robot manipulation [24] and 2D-level object detection [3], while these methods only describe the egocentric images and then reason through the text or image encoding results as inputs, that can only obtain limited and static knowledge. However, despite the considerable achievements of CoT enhanced MLLMs reasoning, it is challenging to reason about 3D object affordance from interaction images due to the complex and dynamic properties of affordance. To bridge this gap, we fine-tune the InternVL [4] and directly input interaction images with prompts to reveal the geometric attributes and underlying interaction intentions.

3. Method

3.1. Overview

Given the inputs $\{P, I\}$, where $P \in \mathbb{R}^{N \times 4}$ denotes a point cloud of the object comprising the coordinate $P_c \in \mathbb{R}^{N \times 3}$ and the affordance annotation $P_{label} \in \mathbb{R}^{N \times 1}$, $I \in \mathbb{R}^{3 \times H \times W}$ denotes an image. N is the number of points, H, W are the height and width of an image. The goal is to optimize the model f_θ that outputs 3D object affordance ϕ , expressed as $\phi = f_\theta(P_c, I)$. As shown in Fig. 2, initially, the inputs are sent to ResNet [9] and PointNet++ [43], obtain specific features $\mathbf{F}_i \in \mathbb{R}^{C \times H_1 \times W_1}$, $\mathbf{F}_p \in \mathbb{R}^{C \times N_p}$, and reshape \mathbf{F}_i to $\mathbb{R}^{C \times N_i}$ ($N_i = H_1 \times W_1$). Then, GREAT captures object structure attributes and affordance interaction procedure by fine-tuning MLLM [4] with the multi-head affordance chain-of-thought to reason about interaction images. Next, the features encoded by Roberta [28] are fused through cross-attention mechanism, calculating object geometric feature $\bar{\mathbf{T}}_o$ and affordance intention feature $\bar{\mathbf{T}}_a$ (Sec. 3.2). Afterwards, with $\bar{\mathbf{T}}_o$, $\bar{\mathbf{T}}_a$ as knowledge dictionaries, GREAT leverages Cross-Modal Adaptive Fusion Module to inject knowledge clues into point features and directly fuse knowledge within image features to obtain fusion features $\mathbf{F}_{tp}, \mathbf{F}_{ti}$ (Sec. 3.3). Eventually, \mathbf{F}_{tp} and \mathbf{F}_{ti} are sent to the decoder to ground 3D object affordance ϕ , the whole process is optimized by a combined loss (Sec. 3.4).

3.2. Multi-Head Affordance Chain-of-Thought

Fine-Tuning MLLM. We adopt InternVL [4] and the injected learnable adapters [10] to fine-tune the MLLM, due to the MLLM primarily focuses on object recognition and description without sufficient understanding of what objects are actually used for and how they interact with humans. As shown in Fig. 2 (c), given an interaction image $I \in \mathbb{R}^{3 \times H \times W}$ and text prompts T , InternVL is desired to perform multi-modal understanding and give correct answers. During the training process, we only fine-tune the injected adapters for 10 epochs with a learning rate of 4e-5 and a LoRA rank of 16, while freezing the main parameters, to preserve the power of InternVL and further empower the model with capabilities in affordance understanding.

Object-Head Reasoning for Geometry. It consists of *Object Interaction Perception* and *Geometric Structure Reasoning*. First, the model needs to focus on understanding the interactive components of an object within the image, refining its perception of the object’s key parts rather than the entire object. As illustrated in Fig. 2 (a), we design the first prompt as “*Point out which part of the object in the image interacts with the person.*” Next, since similar regions of different objects can often perform the same interaction based on shared geometric attributes, the model needs to reason about these features. This allows it to move beyond the constraints of object categories and focus more

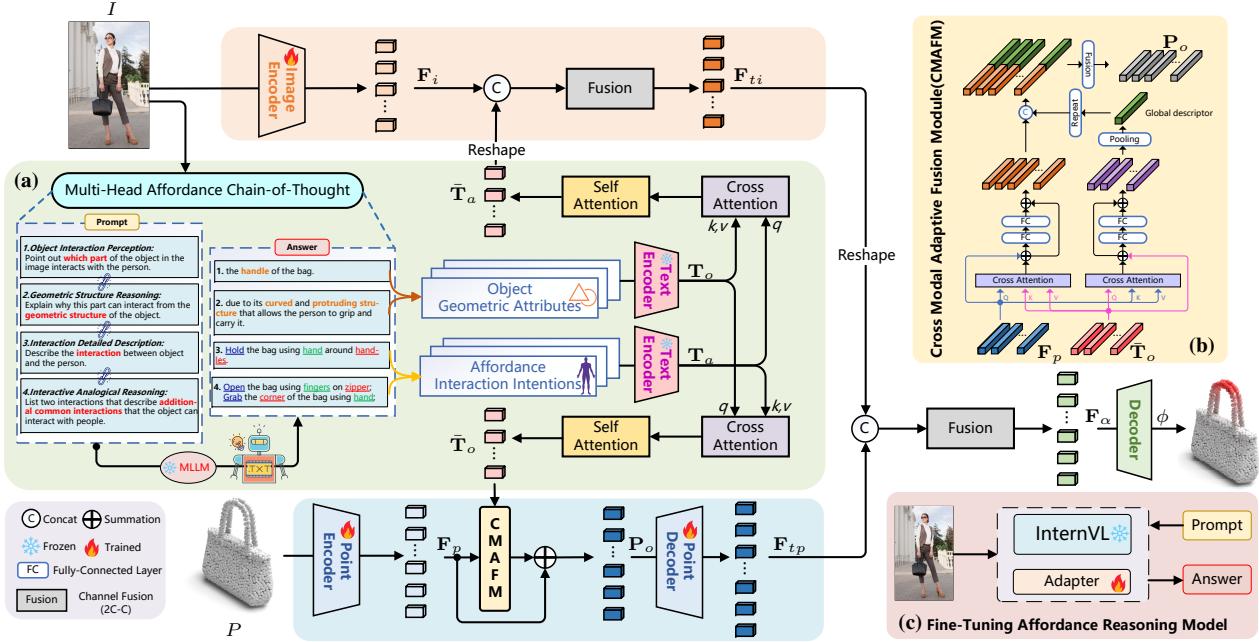


Figure 2. **GREAT pipeline.** Initially, it extracts the respective features $\mathbf{F}_i, \mathbf{F}_p$ through modality-specific backbones, then results of MHACoT inference are encoded and aggregated to form object/affordance knowledge features $\bar{\mathbf{T}}_o, \bar{\mathbf{T}}_a$ (Sec. 3.2). Next, GREAT utilizes CMAFM to inject knowledge into \mathbf{F}_p and \mathbf{F}_i is directly fused to obtain fusion features $\mathbf{F}_{tp}, \mathbf{F}_{ti}$ (Sec. 3.3). Eventually, \mathbf{F}_{tp} and \mathbf{F}_{ti} are sent to the decoder to ground 3D object affordance ϕ (Sec. 3.4).

on the relationship between structure and affordance. Thus, we design the second prompt as “*Explain why this part can interact from the geometric structure of the object.*”

Affordance-Head Reasoning as Brainstorming. It consists of *Interaction Detailed Description* and *Interactive Analogical Reasoning*. First, the model needs to identify the entire interaction process between the object and the person in the image, including the interaction parts on both the object and the person, as well as the type of interaction. This allows the model to generate a fine-grained feature representation and capture the physical structure constraints of the interaction between the person and the object. As shown in Fig. 2 (a), we design the third prompt as “*Describe the interaction between object and the person.*” Next, in the human mind, observing an object is typically followed by associating it with various potential ways of interaction. Inspired by this, the MLLM’s world knowledge repository is leveraged to explore other possible interaction intentions of an object, reducing reliance on specific affordance instances and enhancing analogical reasoning. We design the fourth prompt as “*List two interactions that describe additional common interactions that the object can interact with people.*” Due to the complexity of interaction images, object in the prompts is filled with the object category. Only the key part of the prompts is provided here, for the full prompts, please refer to the appendix.

Knowledge Encoding and Integration. For each interaction image, we concatenate the textual descriptions of the interaction components and the object’s geometric attributes inferred from the object-head into a textual sequence. Additionally, we combine the actual interaction and two potential interactions inferred from the affordance-head. The text encoder Roberta [28] is used to obtain object geometric knowledge feature $\mathbf{T}_o \in \mathbb{R}^{N_o \times C}$ and affordance intention knowledge feature $\mathbf{T}_a \in \mathbb{R}^{N_a \times C}$, where N_o, N_a denote as the number of interaction objects and the number of interaction ways. By applying the cross-attention layer f_m to correlate information from the two knowledge repositories and the self-attention layer f_δ to enrich the contextual information, the process aligns the object’s geometric structure attributes with potential interaction intentions, as formulated below:

$$\bar{\mathbf{T}}_o = f_\delta(f_m(\mathbf{T}_o, \mathbf{T}_a)), \bar{\mathbf{T}}_a = f_\delta(f_m(\mathbf{T}_a, \mathbf{T}_o)), \quad (1)$$

where $\bar{\mathbf{T}}_o \in \mathbb{R}^{N_o \times C}, \bar{\mathbf{T}}_a \in \mathbb{R}^{N_a \times C}$.

3.3. Cross-Modal Adaptive Fusion Module

To better facilitate the cross-modal fusion of the geometric attributes of interaction regions and point cloud features, we propose Cross-Modal Adaptive Fusion Module (CMAFM) that integrates the geometric attributes into the deepest encoder layer of PointNet++ [43] to refine the point feature

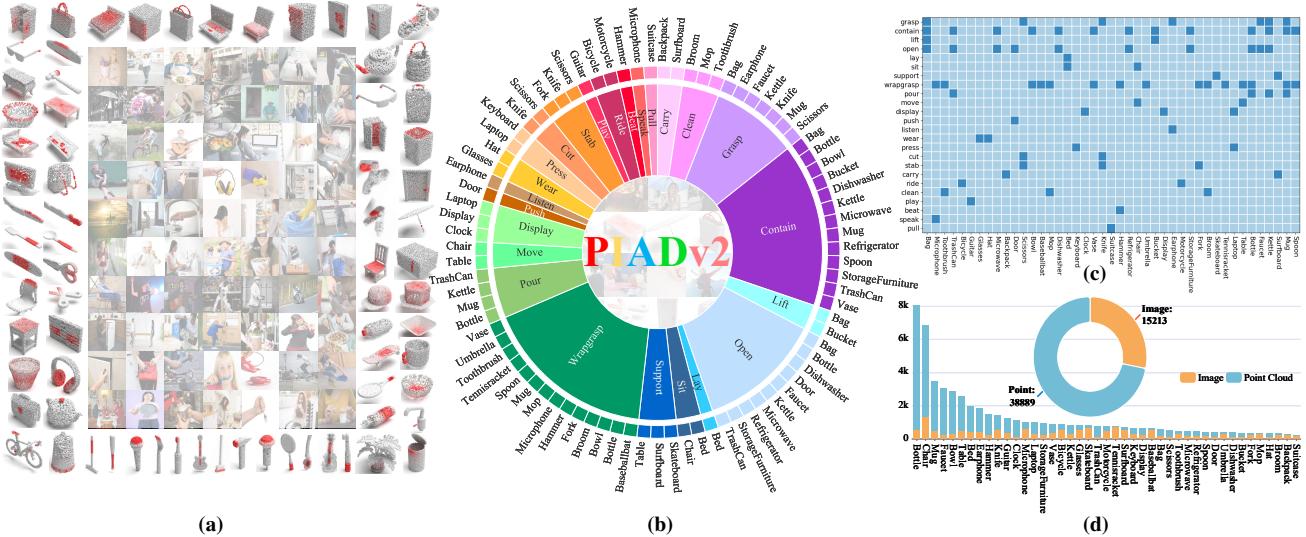


Figure 3. **PIADv2 Dataset.** (a) Extensive data examples from PIADv2, the red region in point clouds is the affordance annotation. (b) Category distribution in PIADv2. (c) Confusion matrix between affordance and object categories, where the horizontal axis represents object category and the vertical axis represents affordance category. (d) Ratio of images and point clouds in each object category.

map, enabling effective cross-modal feature alignment and fusion, as shown in Fig. 2 (b).

Specifically, CMAFM re-represents $\mathbf{F}_p \in \mathbb{R}^{C \times N_p}$ and $\bar{\mathbf{T}}_o \in \mathbb{R}^{N_o \times C}$ in the same feature space to further align the local features. \mathbf{F}_p is projected to form the query $\mathbf{Q} = \mathbf{F}_p \mathbf{W}_1$, $\bar{\mathbf{T}}_o$ is projected to form the key $\mathbf{K} = \bar{\mathbf{T}}_o \mathbf{W}_2$ and value $\mathbf{V} = \bar{\mathbf{T}}_o \mathbf{W}_3$, where $\mathbf{W}_{1\sim 3}$ are projection weights. The cross-attention mechanism integrates these features to extract the interaction context information, formulated as:

$$\mathbf{F}'_p = (\text{softmax}(\mathbf{Q}^T \cdot \mathbf{K} / \sqrt{d}) \cdot \mathbf{V}^T)^T, \quad (2)$$

where $\mathbf{Q} \in \mathbb{R}^{d \times N_p}$, $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{d \times N_o}$, $\mathbf{F}'_p \in \mathbb{R}^{C \times N_p}$, d is the dimension of projection. Similarly $\bar{\mathbf{T}}'_o \in \mathbb{R}^{C \times N_o}$ can be re-represented. Then, CMAFM injects object geometric structure attributes into each point in the point cloud features to obtain the fused point features \mathbf{P}_o , formulated as:

$$\mathbf{P}_o = f[\mathbf{F}'_p + f_\varphi(\mathbf{F}'_p), \Theta(\bar{\mathbf{T}}'_o + f_\varphi(\bar{\mathbf{T}}'_o))], \quad (3)$$

where f_φ denotes two Fully-Connected (FC) layers, Θ denotes pooling and expand it to $\mathbb{R}^{C \times N_p}$, $[.]$ denotes the concatenation, f indicates convolution layers with 1×1 kernel. Finally, \mathbf{P}_o is upsampled to $\mathbb{R}^{C \times N}$ by Feature Propagation layers (FP) [43], formulated as: $\mathbf{F}_{tp} = \text{FP}(\mathbf{P}_o)$.

To understand the multiple affordances of objects and provide rich contextual information for analogical reasoning, underlying interaction intention textual features are fused with the image features:

$$\mathbf{F}_{ti} = f[\Gamma(\bar{\mathbf{T}}_a), \mathbf{F}_i], \mathbf{F}_{ti} \in \mathbb{R}^{C \times N_i}, \quad (4)$$

where Γ denotes reshape $\bar{\mathbf{T}}_a$ to $\mathbb{R}^{C \times N_i}$.

3.4. Decoder and Loss Functions

Image features with interaction intention and point features with geometric structure are fed into the decoder, which jointly reveals the 3D affordance region, formulated as:

$$\mathbf{F}_\alpha = f[\Gamma(\mathbf{F}_{ti}), \mathbf{F}_{tp}], \phi = \sigma(f_\phi(\mathbf{F}_\alpha)), \quad (5)$$

where Γ denotes reshape \mathbf{F}_{ti} to $\mathbb{R}^{C \times N}$, f_ϕ denotes an output head, σ denotes the sigmoid function, $\mathbf{F}_\alpha \in \mathbb{R}^{C \times N}$ is affordance feature representation and $\phi \in \mathbb{R}^{N \times 1}$ represents the 3D object affordance.

Unconstrained by the affordance category labels, we focus on the differences between 3D object affordance ϕ and affordance ground truth annotation P_{label} , enabling the model to directly link 3D object affordances with interaction images through reasoning. Therefore, the total loss consists of a focal loss [26] and a dice loss [37], which supervises point-wise heatmaps, formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{focal} + \mathcal{L}_{dice}. \quad (6)$$

4. Dataset

Collection. We construct the **Point Image Affordance Dataset v2 (PIADv2)**, which comprises paired 2D interaction images and 3D object point clouds. Points are mainly collected from 3DIR [57], 3D-AffordanceNet [6], objaverse [5], etc. Images are mainly collected from AGD20k [32], OpenImage [18] and websites with free licenses. In total, PIADv2 contains 15213 images and 38889 point clouds, covering 43 object and 24 affordance categories, which is the largest scale 3D object affordance dataset so far. Some

Table 2. **Comparison on the PIADv2.** Evaluation metrics of comparison methods on the benchmark, the best results are in **bold**. **Seen**, **Unseen Object** and **Unseen Affordance** are three partitions of the dataset. AUC and aIOU are shown in percentage. \diamond denotes the relative improvement of our method over other methods.

	Seen			Unseen Object				Unseen Affordance				
Methods	AUC ↑	aIOU ↑	SIM ↑	MAE ↓	AUC ↑	aIOU ↑	SIM ↑	MAE ↓	AUC ↑	aIOU ↑	SIM ↑	MAE ↓
Baseline	87.04	34.18	0.594	0.079	72.74 \diamond 9.4%	16.34 \diamond 23.4%	0.336 \diamond 19.6%	0.156 \diamond 30.1%	58.09 \diamond 20.2%	7.88 \diamond 52.9%	0.208 \diamond 39.4%	0.160 \diamond 20.6%
FRCNN [54]	87.05	33.55	0.600	0.082	72.20 \diamond 10.2%	18.08 \diamond 11.5%	0.362 \diamond 11.0%	0.152 \diamond 28.3%	59.08 \diamond 18.2%	7.96 \diamond 51.4%	0.210 \diamond 38.1%	0.156 \diamond 18.6%
XMF [1]	87.39	33.91	0.604	0.078	74.61 \diamond 6.6%	17.40 \diamond 15.9%	0.361 \diamond 11.4%	0.126 \diamond 13.5%	60.99 \diamond 14.5%	8.11 \diamond 48.6%	0.225 \diamond 28.9%	0.152 \diamond 16.4%
IAG [56]	89.03	34.29	0.623	0.076	73.03 \diamond 9.0%	16.78 \diamond 20.1%	0.351 \diamond 14.5%	0.123 \diamond 11.4%	62.29 \diamond 12.1%	8.99 \diamond 34.0%	0.251 \diamond 15.5%	0.141 \diamond 9.9%
LASO [25]	90.34	34.88	0.627	0.077	73.32 \diamond 8.5%	16.05 \diamond 25.6%	0.354 \diamond 13.6%	0.123 \diamond 11.4%	64.07 \diamond 9.0%	8.37 \diamond 44.0%	0.228 \diamond 27.2%	0.140 \diamond 9.3%
Ours	91.99	38.03	0.676	0.067	79.57	20.16	0.402	0.109	69.81	12.05	0.290	0.127

data samples are shown in Fig. 3 (a). The affordance and the object categories are shown in Fig. 3 (b). Both figures show that the dataset covers numerous affordances, supporting various interaction scenarios and diverse object categories.

Annotation. We annotate the affordance of each point cloud instance by affordance category. For instance, an annotation is a matrix of (2048, 4), 2048 is the number of points, and 4 indicates 3D coordinates with the corresponding affordance heatmap, each affordance category possesses such annotation of an instance. For images, we classify images according to the affordance category.

Statistical Analysis. In our training setting, images and point clouds do not require a fixed one-by-one pairing, as they are sampled from different instances to ensure the generalization to distinct instances. Fig. 3 (c) shows the confusion matrix of affordance and object categories, revealing a multi-to-multi relationship, which poses a significant challenge to the accuracy and generalization of the 3D object affordance grounding. Fig. 3 (d) shows the ratio of images and point clouds for each object category, offering insights into the balance between interaction images and 3D object point clouds, further highlighting its comprehensiveness.

Data Partitions. Our dataset provides three partitions, two of which follow the PIAD [56]. **Seen**: the training and test sets share the same objects and affordances. **Unseen Object**: affordances are consistent between the training and test sets, but some objects in the test set do not appear in the training set. **Unseen Affordance**: affordances in the test set are not present in the training set, and so does certain objects. Please refer to the appendix for more details.

5. Experiment

5.1. Benchmark Setting

Evaluation Metrics. For a thorough assessment, we use previous evaluation metrics from advanced works [25, 56] to benchmark the model on our PIADv2 dataset, which include Area Under Curve [29], average Intersection Over Union [45], SIMilarity [47], Mean Absolute Error [52].

Compare Baselines. For a comprehensive comparison, we

select two leading works (IAG [56] and LASO [25]) on 3D affordance grounding and two leading image-point cloud cross-modal works (FRCNN [54] and XMF [1]) mentioned in the IAG. Following IAG, the baseline directly concatenates the features from modality-specific extractors.

Implementation Details. We take PointNet++ [43] and ResNet18 [9] as the default 3D and 2D backbones, respectively. To ensure a fair comparison, the same feature extractor and settings are used to reproduce the baselines. We train the GREAT for 65 epochs with a batch size of 16, utilizing the Adam optimizer with a learning rate of 1e-4.

5.2. Comparison Results

The comparison results of evaluation metrics are presented in Tab. 2, demonstrating **GREAT** significantly outperforms the compared baselines across all metrics in three partitions and achieves the state-of-the-art performance. Furthermore, the results are visualized in Fig. 4.

Seen vs. Unseen. Quantitatively analyzing the results of the Tab. 2, all baselines demonstrate a stepwise metrics decrease in all partitions. This trend emphasizes the difficulty of generalizing unseen objects and affordances. Compared to other baselines, the superior performance demonstrated by GREAT in the open-vocabulary proves the necessity and rationality of our task setting for open-vocabulary affordance grounding. Qualitative analysis of the visualization results in Fig. 4 shows little difference in the Seen setting, but in the Unseen setting, GREAT significantly outperforms the other methods. For example, in the case of kettle, only IAG detects the small affordance region for pouring, while methods that directly link object structure with textual descriptions fail to predict and can only identify the trained affordance: grasp. In contrast, our method uncovers interaction details in 2D interaction images by leveraging the MLLM embedded with world knowledge for MHACoT reasoning, leading to more precise predictions.

5.3. Ablation Study

We conduct a thorough ablation study to validate the effectiveness of the model design, as shown in Tab. 3. First,

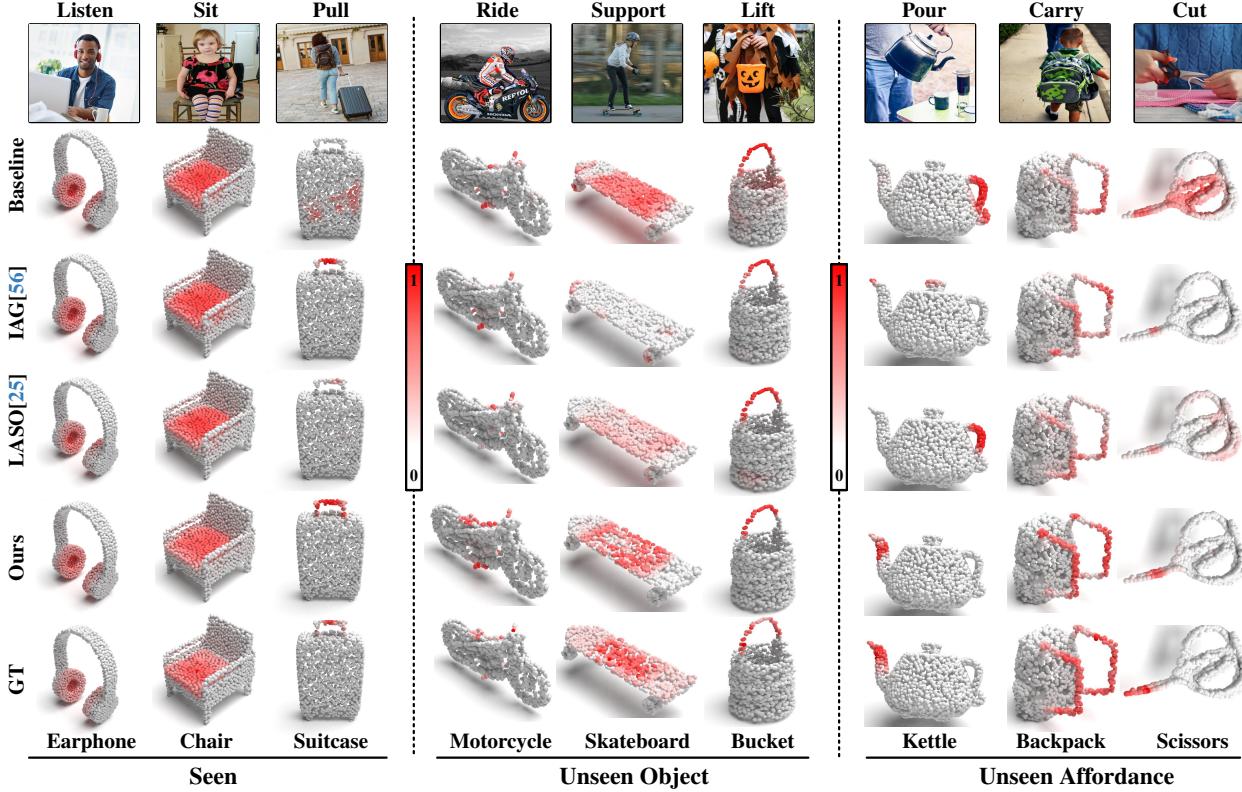


Figure 4. **Visualization Results.** The first row is the interaction image and the last row is the ground truth of 3D object affordance in point cloud. The left-middle-right partitions correspond to the visual comparison results for different 3D object affordance in the Seen, Unseen Object, and Unseen Affordance partitions, respectively. The depth of red represents the affordance probability.

we ablate the object head and affordance head in MHACoT separately to demonstrate the importance of acquiring geometric attributes of objects and underlying interaction intentions. Additionally, removing the CMAFM module impairs the alignment between geometric attributes and the point cloud, while omitting MLLM fine-tuning reduces the accuracy of reasoning, resulting in decreased performance.

To further demonstrate the effectiveness of AffCoT and ObjCoT, we visualize attention maps on object geometries and interaction images when one of them is removed. As shown in Fig. 5 (a), relying solely on object geometric attributes without AffCoT results in a model that only focuses on affordance seen in the training set and fails to reason analogically for unseen affordance. The same issue arises with partial object point clouds. Additionally, as shown in Fig. 5 (b), we visualize features from the interaction images, as the algorithm is not limited by additional classification heads. When AffCoT guides the model to identify the approximate affordance region in the interaction image, ObjCoT inference further localizes the key interaction part of the object rather than the entire object, *e.g.* knife-cut perceives the blade of the knife and kettle-pour perceives the spout of the kettle in the interaction image.

Table 3. **Ablation studies.** Performance when not modeling MHACoT with affordance-head CoT (AffCoT.) and object-head CoT (ObjCoT.), CMAFM and not introducing MLLM fine-tuning (FT.). \times means without.

	Metrics	Ours	\times AffCoT.	\times ObjCoT.	\times CMAFM	\times FT.
Seen	AUC	91.99	90.88	90.13	89.52	88.75
	aIOU	38.03	36.94	36.55	29.48	35.19
	SIM	0.676	0.659	0.651	0.590	0.625
	MAE	0.067	0.069	0.071	0.078	0.075
Unseen Object	AUC	79.57	74.58	75.87	78.42	77.83
	aIOU	20.16	18.50	18.56	16.62	17.07
	SIM	0.402	0.390	0.383	0.349	0.374
	MAE	0.109	0.111	0.120	0.119	0.118
Unseen Affordance	AUC	69.81	67.18	64.69	63.00	66.49
	aIOU	12.05	10.93	8.81	6.24	10.06
	SIM	0.290	0.287	0.254	0.235	0.256
	MAE	0.127	0.164	0.133	0.128	0.134

5.4. Performance Analysis

We conducted experiments separately for Seen, Unseen object, and Unseen affordance partitions, enabling a deeper analysis of the model’s performance in various contexts.

Multiple Objects. In the case of humans interacting with

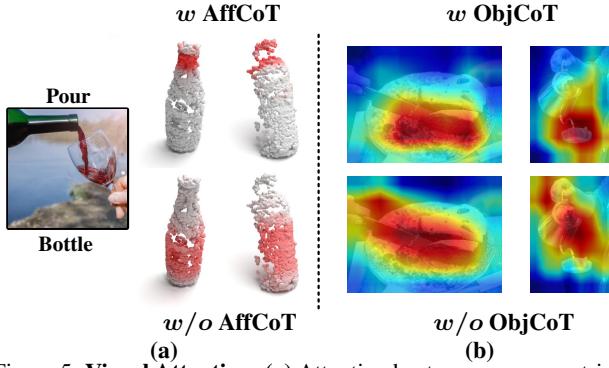


Figure 5. **Visual Attention.** (a) Attention heatmaps on geometries with (w) and without (w/o) the AffCoT. (b) Feature maps on interaction images with (w) and without (w/o) the ObjCoT.

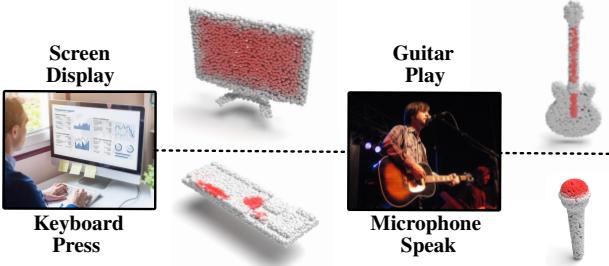


Figure 6. **Multiple Objects.** The anticipations for multiple objects with the same interaction image that contain different interactions.

distinct objects at the same time, the model needs to have the ability to understand interactions with different objects. As shown in Fig. 6, when reasoning about different objects of the same interaction image, the model can pinpoint the object affordance region.

Multiple Affordances. To verify whether the model reasons about the 3D object affordance regions based on the understanding of interaction images, we use the model to infer distinct affordances for the same objects, as shown in Fig. 7. The results demonstrate that, for the same object, the model outputs different results depending on the interaction, and the localized 3D object affordance regions are consistent with the interactions depicted in the 2D images.

Multiple Instances. To assess the generalization and robustness of the model, we conduct an experiment to validate its understanding of different point cloud instances with the same object category, as shown in Fig. 8. The results demonstrate that the model can not only accurately locate point clouds that are highly similar to the shapes of the interacting objects in the images (Fig. 8 (a)), but also can effectively locate the same affordance regions in point clouds of the same category, even with significant geometric variations (Fig. 8 (b)), demonstrating its ability to generalize affordance grounding and maintain robustness under geometric variations.



Figure 7. **Multiple Affordances.** The anticipations for multiple affordances with the same point cloud for different interactions.

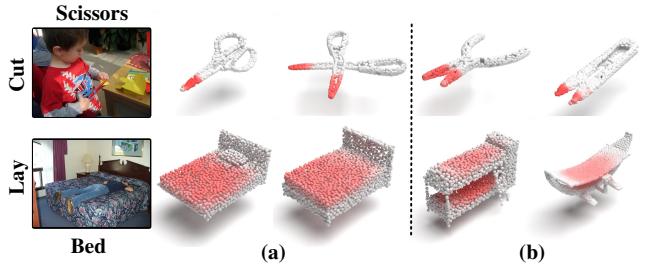


Figure 8. **Multiple Instances.** (a) Similar geometric instances. (b) Different geometric instances.

6. Conclusion

We present grounding 3D object affordance in an open-vocabulary fashion, which reasons from interaction images, extrapolating from predefined sample space and generalize to unseen scenarios. To achieve so, we propose a novel framework to utilize multi-head affordance chain-of-thought reasoning, excavating invariant geometric properties and analogous reasoning about potential interactions, with the alignment of cross-modal features to localize 3D object affordance region. Furthermore, We introduce the largest 3D object affordance dataset PIADv2, which contains 15K interaction images and over 38K 3D objects with annotations. Extensive experiments demonstrate the effectiveness and superiority of GREAT. It supports affordance understanding in open scenarios, potentially enhancing robots’ autonomous interaction in unknown environments. We believe it could offer fresh insights and promote research in the area of visual affordance understanding.

Limitations and Future Work. The limitation of GREAT lies in the high computational complexity of its multi-step reasoning, which can become a bottleneck in large-scale or real-time applications. In future work, we aim to create inference-specific datasets and use them to distill multi-modal models into specialized knowledge domains, enabling faster and more efficient performance in real-world.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 62306295, 62225207 and 62436008.

References

- [1] Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. In *Advances in Neural Information Processing Systems*, 2022. 6
- [2] Xu Chao, Yixin Chen, He Wang, songchun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. *arXiv preprint arXiv:2202.13519*, 2022. 1
- [3] Changmao Chen, Yuren Cong, and Zhen Kan. Worldafford: Affordance grounding based on natural language instructions. *arXiv preprint arXiv:2405.12461*, 2024. 3
- [4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1, 3
- [5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 3, 5
- [6] 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 Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 5
- [7] Mary L. Gick and Keith J. Holyoak. Analogical problem solving. *Cognitive Psychology*, 12(3):306–355, 1980. 1
- [8] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding. *ACM Computing Surveys (CSUR)*, 54:1 – 35, 2018. 1
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 6
- [10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3
- [11] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1
- [12] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 1
- [13] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Comput. Surv.*, 50(2), 2017. 1
- [14] James J Gibson. The ecological approach to visual perception: classic edition. *Psychology press*, 2014. 1
- [15] Wu Jiannan, Zhong Muyan, Xing Sen, Lai Zeqiang, Liu Zhaoyang, Chen Zhe, Wang Wenhui, Zhu Xizhou, Lu Lewei, Lu Tong, Luo Ping, Qiao Yu, and Dai Jifeng. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint arXiv:2406.08394*, 2024. 3
- [16] Philip Johnson-Laird. *How We Reason*. Oxford University Press, 2008. 1
- [17] David Inkyu Kim and Gaurav S. Sukhatme. Semantic labeling of 3d point clouds with object affordance for robot manipulation. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5578–5584, 2014. 1
- [18] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 5
- [19] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation, 2024. 1
- [20] Bin Lei, pei Hung Lin, Chunhua Liao, and Caiwen Ding. Boosting logical reasoning in large language models through a new framework: The graph of thought, 2023. 3
- [21] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1
- [23] Puhao Li, Tengyu Liu, Yuyang Li, Muzhi Han, Haoran Geng, Shu Wang, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations. *arXiv preprint arXiv:2404.17521*, 2024. 1
- [24] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *CVPR*, 2024. 1, 3
- [25] Y. Li, N. Zhao, J. Xiao, C. Feng, X. Wang, and T. Chua. Laso: Language-guided affordance segmentation on 3d object. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 6, 7
- [26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 5
- [27] Cuiyu Liu, Wei Zhai, Yuhang Yang, Hongchen Luo, Sen Liang, Yang Cao, and Zheng-Jun Zha. Grounding 3d scene affordance from egocentric interactions. *arXiv preprint arXiv:2409.19650*, 2024. 1
- [28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettle,

- moyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. 3, 4
- [29] Jorge M. Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17:145–151, 2008. 6
- [30] Liangsheng Lu, Wei Zhai, Hongchen Luo, and Yang Cao. Phrase-based affordance detection via cyclic bilateral interaction. 2022. 3
- [31] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot affordance detection. In *IJCAI*, 2021. 3
- [32] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *CVPR*, 2022. 5
- [33] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Leverage interactive affinity for affordance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6809–6819, 2023. 3
- [34] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning visual affordance grounding from demonstration videos. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16857–16871, 2024. 3
- [35] Junyi Ma, Xieyanli Chen, Wentao Bao, Jingyi Xu, and Hesheng Wang. Madiff: Motion-aware mamba diffusion models for hand trajectory prediction on egocentric videos, 2024. 1
- [36] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt, 2023. 1
- [37] 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, 2016. 5
- [38] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
- [39] Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. In *Conference on Robot Learning (CoRL)*, 2021. 3
- [40] Bogdan Moldovan, Plinio Moreno, Martijn van Otterlo, José Santos-Victor, and Luc De Raedt. Learning relational affordance models for robots in multi-object manipulation tasks. In *2012 IEEE International Conference on Robotics and Automation*, pages 4373–4378, 2012. 1
- [41] Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. 2023. 1, 2, 3
- [42] Xingyu Peng, Yan Bai, Chen Gao, Lirong Yang, Fei Xia, Beipeng Mu, Xiaofei Wang, and Si Liu. Global-local collaborative inference with llm for lidar-based open-vocabulary detection, 2024. 1, 3
- [43] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 3, 4, 5, 6
- [44] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024. 3
- [45] Md.Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*, 2016. 6
- [46] Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z. Morley Mao. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*, 2022. 3
- [47] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991. 6
- [48] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibeи Yang. Cot-det: Affordance knowledge prompting for task driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3068–3078, 2023. 3
- [49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 1
- [50] Tuan Van Vo, Minh Nhat Vu, Baoru Huang, Toan Nguyen, Ngan Le, Thieu Vo, and Anh Nguyen. Open-vocabulary affordance detection using knowledge distillation and text-point correlation. 2024. 1, 2, 3
- [51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. 3
- [52] 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*, 30:79–82, 2005. 6
- [53] Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Det-toolchain: A new prompting paradigm to unleash detection ability of mllm, 2024. 3
- [54] Xinli Xu, Shaocong Dong, Tingfa Xu, Lihe Ding, Jie Wang, Peng Jiang, Liqiang Song, and Jianan Li. Fusionrcnn: Lidar-camera fusion for two-stage 3d object detection. *arXiv preprint arXiv:2209.10733*, 2022. 6
- [55] Jie Yang, Wang Zeng, Sheng Jin, Lumin Xu, Wentao Liu, Chen Qian, and Ruimao Zhang. Kptllm: Unveiling the power of large language model for keypoint comprehension, 2024. 3
- [56] 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 (ICCV)*, pages 10905–10915, 2023. 1, 2, 3, 6, 7
- [57] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16284–16295, 2024. 5

- [58] Yuhang Yang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *arXiv preprint arXiv:2405.13659*, 2024. 1
- [59] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. 3
- [60] Yao Yao, Zuchao Li, and Hai Zhao. Beyond chain-of-thought, effective graph-of-thought reasoning in language models, 2024. 3
- [61] Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection in the wild. *arXiv preprint arXiv:2108.03658*, 2021. 3
- [62] Wei Zhai, Yang Cao, Jing Zhang, and Zheng-Jun Zha. Exploring figure-ground assignment mechanism in perceptual organization. *Advances in Neural Information Processing Systems*, 35:17030–17042, 2022. 1
- [63] Wei Zhai, Pingyu Wu, Kai Zhu, Yang Cao, Feng Wu, and Zheng-Jun Zha. Background activation suppression for weakly supervised object localization and semantic segmentation. *International Journal of Computer Vision*, pages 1–26, 2023. 3
- [64] Wei Zhai, Yang Cao, Jing Zhang, Haiyong Xie, Dacheng Tao, and Zheng-Jun Zha. On exploring multiplicity of primitives and attributes for texture recognition in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):403–420, 2024. 1
- [65] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models, 2022. 3
- [66] Zichen Zhang, Hongchen Luo, Wei Zhai, Yang Cao, and Yu Kang. Bidirectional progressive transformer for interaction intention anticipation. *arXiv preprint arXiv:2405.05552*, 2024. 1
- [67] Zichen Zhang, Hongchen Luo, Wei Zhai, Yang Cao, and Yu Kang. Pear: Phrase-based hand-object interaction anticipation. *arXiv preprint arXiv:2407.21510*, 2024. 1
- [68] 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. 3