SeqAfford: Sequential 3D Affordance Reasoning via Multimodal Large Language Model

Chunlin Yu[1,†]   Hanqing Wang[1,†]   Ye Shi[1]   Haoyang Luo[1]   Sibei Yang[1]   Jingyi Yu[1]   Jingya Wang[1, *]

[1]ShanghaiTech University, Shanghai, China

{yuchl,wanghq2024,shiye,luohy2024, yangsb,yujingyi,wangjingya}@shanghaitech.edu.cn
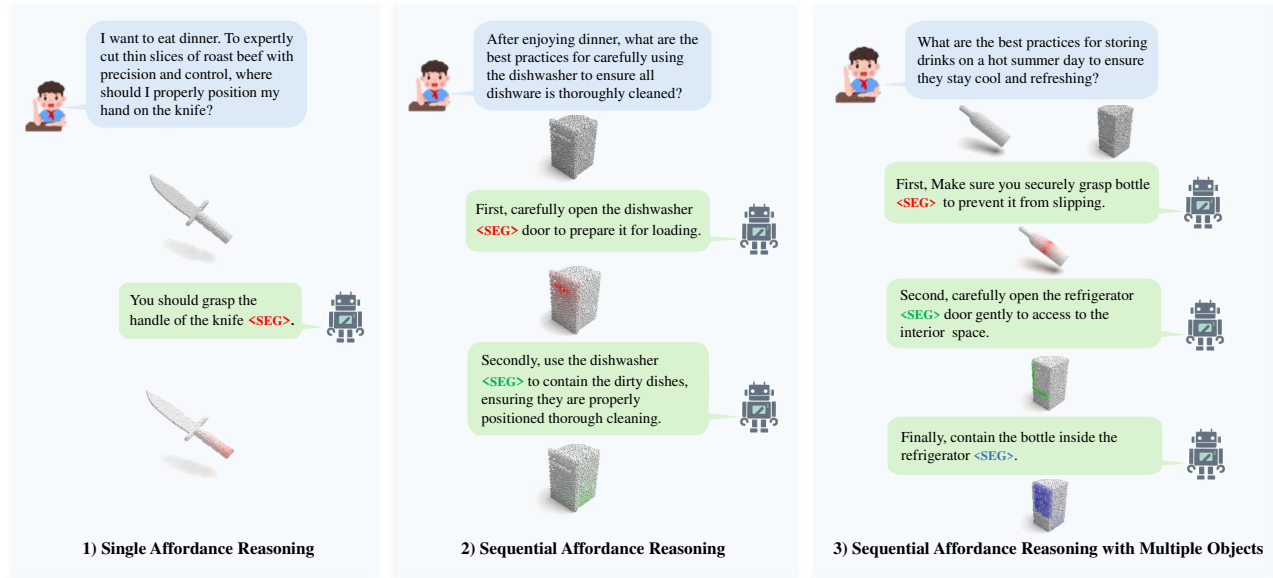
Figure 1. **Sequential 3D affordance reasoning task with different types of interactions.** We introduce SeqAfford, a Multi-Modal Language Model (MLLM) capable of serialized affordance inference implied in human instructions: 1) Single Affordance Reasoning; 2) Sequential Affordance Reasoning; 3) Sequential Affordance Reasoning with Multiple Objects

Abstract

3D affordance segmentation aims to link human instructions to touchable regions of 3D objects for embodied manipulations. Existing efforts typically adhere to single-object, single-affordance paradigms, where each affordance type or explicit instruction strictly corresponds to a specific affordance region and are unable to handle long-horizon tasks. Such a paradigm cannot actively reason about complex user intentions that often imply sequential affordances. In this paper, we introduce the Sequential 3D Affordance Reasoning task, which extends the traditional paradigm by reasoning from cumbersome user intentions and then decomposing them into a series of segmentation maps. Toward this, we construct the first instruction-based affordance segmentation benchmark that includes reasoning over both single and sequential affordances, comprising 180K instruction-point cloud pairs. Based on the benchmark, we propose our model, SeqAfford, to unlock the 3D multi-modal large language model with additional affordance segmentation abilities, which ensures reasoning with world knowledge and fine-grained affordance grounding in a cohesive framework. We further introduce a multi-granular language-point integration module to endow 3D dense prediction. Extensive experimental evaluations show that our model excels over well-established methods and exhibits open-world generalization with sequential reasoning abilities. Project page: https://seq-afford.github.io/.

[†]Equal contributions.
[*]Corresponding author.

1. Introduction

Affordance is a crucial lens through which humans and embodied agents interact with various objects of the world . When provided with human instructions, affordance aims to highlight the actionable possibilities of these objects, linking visual perception with manipulation. While 2D affordance [14, 35, 47] offers visual cues that suggest potential actions to embodied systems, 3D affordance [3, 6, 8, 24, 51] provides a more direct and intuitive guidance for executing tasks in the realistic 3D world, and thus solidify the foundation for downstream robot manipulation tasks [11, 30, 44, 45].

Previous work on 3D affordances has largely focused on the single-object, single-affordance paradigm, where affordance maps are grounded either in affordance categories [7, 29] or 2D demonstration images [8, 50]. Recently, language models have been employed to pair 3D objects with natural language questions [16], each designed to elicit a specific affordance. For example, the question "How can you go through the door?" can be interpreted by language models like BERT [12] or RoBERTa [20] as referring to the "openable" affordance component (e.g., the door handle). However, regardless of whether the grounding source is an affordance phrase, HOI image, or a natural question, each grounding source generally corresponds to a fixed affordance type.

Current systems can't actively reason complex user intentions, break them down into actionable primitives, or formulate chains of affordances from each primitive. Real-world physical interactions require modeling complex human instructions involving multiple objects and sequential affordances. For example, to 'reheat food in a bowl using the microwave', agents must first grasp the bowl, open the microwave, then contain the bowl inside. Thus, multi-object sequential reasoning is crucial for next-generation affordance systems.

Recently, Large-Language Models (LLMs) [2, 31, 55] have demonstrated exceptional sequential reasoning abilities, ingrained with internalized common-sense knowledge encoded from vast text data corpora. On top of that, the emergence of 3D Multimodal Large-Language Models (MLLMs) [34, 48] have further expanded their possibilities in understanding various object shapes in the 3D world. However, even the latest 3D MLLMs are not panaceas for reasoning about visual affordances from 3D objects, as they primarily focus on object-centered text generation tasks. This, therefore, highlights a pressing question: *Can we devise a 3D multi-modal large language model to sequentially reason and segment multi-object affordances based on long-horizon human instructions?*

In this paper, we introduce a new task called Sequential 3D Affordance Reasoning, designed to narrow the gap with real-world demands. Towards this, we construct a large-scale sequential affordance reasoning benchmark, containing 180K instruction-point cloud pairs. To ensure the diversity of the instruction data, four distinct methods are used for generating instructions. These methods include utilizing immersive 2D HOI images and detailed, colorful mesh renderings to prompt GPT-4o to generate diverse instructions, drawing on its inherent world knowledge, as illustrated in Fig. 2. Supported by this benchmark, we introduce our model SeqAfford, which unlocks the current 3D MLLMs with sequential affordance segmentation abilities. To further bolster the 3D dense prediction and reasoning task, we introduce a multi-granular language-point integration module, where dense point features conditioned on segmentation tokens of the large language model are integrated with sparse point features for subsequent dense prediction tasks. This module not only effectively injects the reasoning results of large language models into the dense point features, but enhances the affordance segmentation task with multi-granular levels of representation.

To summarize, our contributions are as follows:

- We introduce the Sequential 3D Affordance Reasoning task, which involves the sequential reasoning and segmentation of affordances based on complex human instructions. This paradigm is crucial for the development of next-generation affordance systems.
- We develop a large-scale sequential affordance reasoning benchmark with 180K instruction-point cloud pairs, serving as a comprehensive resource to advance research in affordance reasoning.
- We leverage the internalized world knowledge of pre-aligned 3D MLLMs to achieve fine-grained affordance-level multi-target sequential reasoning and explanations in a cohesive framework.

2. Related Work

3D Affordance Segmentation. With the rapid advancement of embodied AI, research on 3D affordance has increasingly garnered significant attention from both academia and industry. 3D AffordanceNet [7], built on point cloud data from PartNet [26], was the first to construct a fine-grained 3D affordance dataset, establishing a benchmark for 3D affordance research. Building on this, some researchers [38, 50] proposed leveraging universal knowledge extracted from 2D human-object interaction (HOI) images to assist in 3D affordance segmentation. However, these methods rely solely on visual information to infer affordances, without considering that embodied agents in the real world communicate with humans through language. Consequently, such methods are limited in their applicability for direct deployment in embodied agents. Recently, Li et al. [16] introduced a language-based affordance segmentation task, promoting the integration of natural language and affordance understanding.
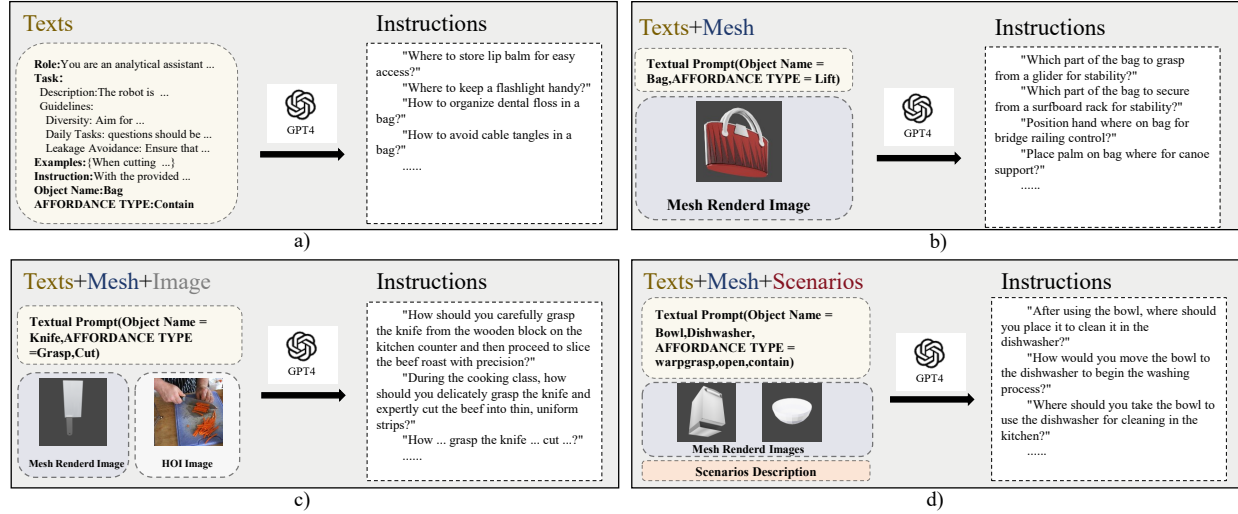
Figure 2. **Preparing the instructions.** To better utilize the world knowledge of GPT4, We prompt GPT-4o to generate diverse instructions based on 4 types of system prompts containing different modalities as input. Instructions are generated based on input prompts with modalities from a) purely textual affordance type, object name; b) the mesh-rendered image of the object; c) the mesh-rendered image and HOI images that reveal affordances of the object; d) the mesh-rendered image and textual description of the scenario.

Following this, Chu and Zhang [4] utilized LLMs to locate objects in 2D images and subsequently retrieve corresponding objects from a 3D dataset. Although these approaches leverage the reasoning capabilities of LLMs, they lack the ability to perform joint visual and language alignment, failing to bridge the gap between 2D and 3D modalities. Furthermore, their research assigns each text statement to a single specific affordance, overlooking the complexity of scenarios that often require the coordination of multiple affordances. In response to these limitations, we propose the integration of 3D multimodal large language models (MLLMs) into affordance segmentation. This novel approach enables the model to simultaneously comprehend contextual semantics and point cloud data, thereby facilitating affordance reasoning across a wide range of complex scenarios. By incorporating both natural language and 3D data, our model offers a more comprehensive and versatile understanding of affordances, rendering it better suited for real-world applications in embodied AI.

Multimodal Large Language Models. Large Language Models (LLMs) have achieved remarkable success in processing natural language, and researchers have been working to extend these models' reasoning capabilities into the visual domain. Early efforts [2, 18, 28, 55] have made significant strides by enabling LLMs to process both visual and textual information concurrently, thereby introducing an initial level of human-like multimodal reasoning. However, for many practical applications, such as visual segmentation, these models lack the necessary fine-grained perception required for detailed visual tasks. To address this issue, research

efforts [21, 41, 52, 54] enable the localization of specific regions within images by encoding spatial coordinates as tokens, improving the models' ability to reason about precise areas within the visual data.

Building on these 2D MLLM advancements, research has increasingly expanded into the 3D domain. Following the paradigm of LLaVA, PointLLM [48] replaces the vision encoder with a 3D point encoder, enabling the processing of 3D data within the latent space of LLMs and facilitating 3D object understanding. Similarly, ShapeLLM [34] is built upon the enhanced 3D encoder RECON++, which strengthens geometric understanding through multi-view image distillation. Other models, such as 3D-LLM [9], leverage 2D foundational models, like CLIP ViT [36], to process multi-view rendered images of 3D point clouds, thereby integrating the 3D world into LLMs. However, despite these advancements, the majority of existing MLLMs are primarily focused on scene-level and object-level understanding, lacking the ability to recognize and segment fine-grained affordances of 3D objects in diverse semantic contexts. Addressing this limitation, our study aims to endow MLLMs with affordance-aware perception, enabling them to interpret and act upon 3D objects more effectively in context-sensitive scenarios.

3. Dataset

Affordance segmentation involves understanding the operability of objects in various contexts, and the varying complexity of intentions adds significant challenges to the process. Simple instructions typically pertain to the direct usage

Method	#Sequential	#World Knowledge	#Multi-object	#Instruction-Point Pairs	#Point Cloud
3D AffordanceNet [7]	×	×	×	×	23k
O2O-Afford [27]	×	×	×	×	1.7k
Partafford [46]	×	×	×	×	25k
IAGNet [50]	×	×	×	×	7k
LASO [16]	×	×	×	19k	8.4k
Ours	✓	✓	✓	180k	18k

Table 1. **Comparison of Exisiting 3D Affordance Dateset with Ours.** #Point Cloud and #Instruction-Point Cloud Pairs denote the number of point clouds and instruction-point cloud pairs, respectively. × indicates that the dataset does not possess this attribute.

of an object, such as grasping a cup or opening a door. In contrast, complex instructions may involve multi-step actions or require contextual understanding, such as using a cup for a specific occasion or purpose. To address the challenge of affordance segmentation based on both simple and complex instructions, we constructed a dataset of instruction-point cloud pairs based on 3D AffordanceNet [7], encompassing both simple and complex types of intentions, which includes 162,386 instruction-point cloud pairs in the single affordance segmentation setting and 20,847 pairs in the sequential affordance segmentation setting, comprising a total of 18,371 point cloud instances across 23 object categories.

3.1. Dataset Collection

Point Cloud. Our point cloud data and affordance annotations are entirely sourced from 3D AffordanceNet [7]. In the simple instruction setting, we generated five instructions for each affordance of every point cloud instance. For the sequential affordance segmentation setting, we carefully selected point cloud categories that support this configuration and generated corresponding instructions for each affordance sequence combination.

Instruction. To create instructions, we developed four methods for generating instructions using GPT-4 [1]. Unlike LASO [16] which generates the same texts for all point cloud instances of a specific affordance type for each category, we generate text for each point cloud by utilizing the point cloud instance from 3D AffordanceNet [7] to trace back to the Mesh-rendered images in the PartNet [26] dataset. Additionally, we collect HOI images corresponding to the affordance types from IAGNet [50] to alleviate GPT's hallucination issues, enabling better understanding. Fig. 2 illustrates our ways of generating instructions in detail.

3.2. Statistics and Analysis

Our dataset comprises 162,386 instruction-point cloud pairs in the single affordance segmentation setting and 20,847 pairs in the sequential affordance segmentation setting, making up a total of 18,371 point cloud instances across 23 object categories.

Based on the complexity of the instructions, we have

divided them into two settings. The first setting is based on instructions that can only infer a single specific affordance, which we define as the "Single Affordance Segmentation Task". The second setting is based on instructions that can infer a combination of multiple affordances in sequence, which we define as the "Sequential Affordance Segmentation Task".

Inspired by the evaluation settings presented in LASO [16] and IAGNet [50], we propose two types of distinct dataset configurations including *Seen* and *Unseen*. *Seen*: This default setting maintains similar distributions of object classes and affordance types across both training and testing phases. *Unseen*: This configuration is specifically designed to evaluate the model's ability to generalize affordance knowledge. In this setup, certain affordance-object pairings are deliberately omitted from the training set but introduced during testing. For instance, while the model may learn to grasp objects like bags and mugs during training, it is expected to generalize the "grasp" affordance to earphones, a pairing absent from the training data.

4. Method

4.1. Architecture Overview

The overall architecture of SeqAfford is presented in Fig. 3. Generally, SeqAfford mainly consists of three components: 1) a 3D vision encoder benefited from large-scale 3D representation learning, which provides solid foundations for dense prediction tasks; 2) a 3D Multi-modal Large Language Model (MLLM) $\mathcal{F}$ that exhibits affordance reasoning ability with the aid of internalized world knowledge; 3) a Multi-Granular Language-Point Integration module that considers the effective integration the point features and the segmentation tokens of MLLM, synergizing both reasoning and segmentation tasks from a multi-granular feature perspective.

4.2. Network Architecture

3D MLLM Backbone. Recently, a series of 3D multi-modal language models have been proposed to deepen the understanding of open-world 3D objects, among which
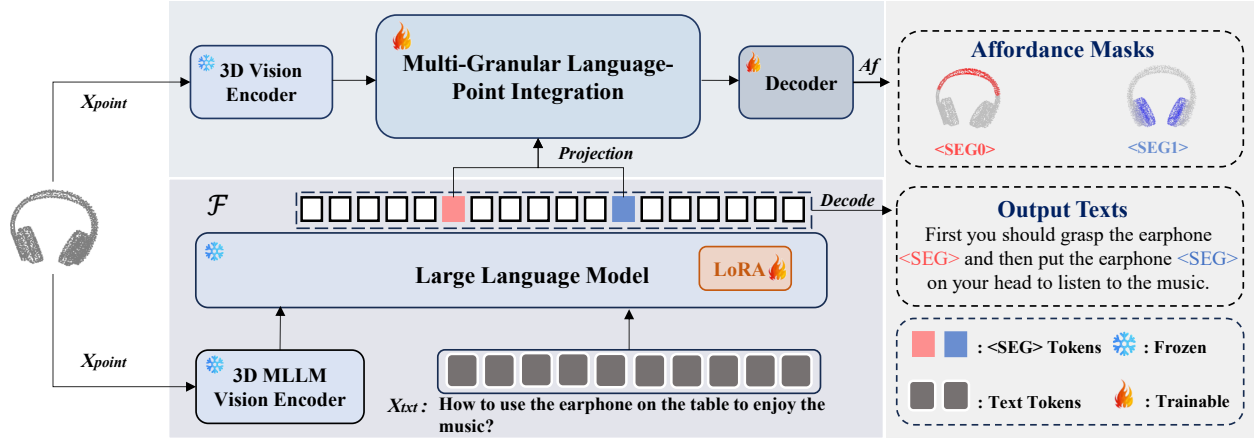
Figure 3. Main Pipeline. Given the point clouds of the target objects and a piece of complex human instruction, SeqAfford first reasons from this instruction and decomposes it into several hidden <SEG> tokens extracted from the last-layer embeddings, each representing an intermediate affordance segmentation result. Then, for each <SEG>, the point features extracted by the 3D vision encoder dynamically interact with the <SEG> token before being sent to the decoder for mask generation. The interaction is achieved through multi-granular language-point integration, synergizing both reasoning and affordance segmentation. We use LoRA for efficient fine-tuning.

ShapeLLM is recently pretrained for understanding various embodied interactions. In light of this, we adopt ShapeLLM [34] as our backbone, denoted as $\mathcal{F}$, where the point cloud encoder ReCon++ is pre-trained via multi-view distillation based on ReCon [33], and the LLM is drawn from LLaMa [40]. Previous work on 3D affordance tasks typically employed 3D backbones [29, 50] or utilized separate point-language encoders [16], which may fall short in reasoning and open-world generalization abilities. Here, we leap ahead by directly utilizing a unified 3D MLLM instead of relying solely on pure LLMs or other visual structures as our backbones, for two main reasons: 1) it opens new possibilities for open-world 3D objects understanding, bolstering the generalization of unseen objects or affordances; 2) it internalizes affordance perception ability, compressing it into natural language form, thus preparing for the subsequent affordance reasoning task.

Squential Affordance Reasoning. Despite the efficacy of 3D MLLMs in aligning 3D representations with natural language, they are primarily designed for object-oriented text generation tasks and lack the capability for 3D dense prediction tasks, particularly in fine-grained affordance segmentation. To encapsulate the segmentation ability into 3D MLLMs, a specific segmentation token <SEG> can be appended to the vocabulary set of the MLLM, inspired by [13].

Formally, when provided with point cloud $\mathbf{X}_{\text{point}}$ and a text instruction $\mathbf{X}_{\text{txt}}$ that demonstrates the user intentions on these potential objects, the 3D MLLM absorbs this multimodal information and generates a text response $\tilde{\mathbf{y}}_{\text{txt}}$. It can be formulated as

$$\tilde{\mathbf{y}}_{\text{txt}} = \mathcal{F}(\mathbf{X}_{\text{point}}, \mathbf{X}_{\text{txt}}), \tag{1}$$

where the output $\tilde{\mathbf{y}}_{txt}$ would include several <SEG> tokens, where a single <SEG> indicates a segmentation result within the sequence. We then extract the last-layer embeddings $\{\mathbf{h}_{\text{seg}}^{(i)}\}_{i=0}^{S-1}$ corresponding to the <SEG> tokens, where S is the number of predicted affordance sequences. Afterwards, an MLP projection layer to obtain $\{\mathbf{H}_{seg}^{(i)}\}_{i=0}^{S-1}$ as follows

$$\mathbf{H}_{\text{seg}}^{(i)} = \text{Proj}(\mathbf{h}_{\text{seg}}^{(i)}). \tag{2}$$

Multi-Granular Language-Point Integration. After obtaining several segmentation tokens that indicate a sequence of regions for reasoning where the given objects can be afforded, the remaining work entails integrating the abstracted reasoning results into 3D point clouds for dense affordance predictions. Therefore, the multi-granular language-point integration module mainly consists of two stages: 1) the multi-granular feature propagation process, which iteratively up-samples the point cloud features into dense features with multiple granularities considered; 2) the point-language integration stage, which distills the informative language features (i.e. the segmentation tokens) into the dense visual features (i.e. the 3D point features), and fuses the integrated dense features with global sparse features for final affordance segmentation. This serves as a crucial step towards reasoning and segmenting affordances in a cohesive framework.

To enable multi-granular feature propagation, as shown in Fig. 4, we hierarchically up-sample (UpS. in Fig. 4) the intermediate features from the 3D encoder, and propagate features through farthest point sampling (FPS in Fig. 4) process to sequentially generate $\mathbf{f}_1$, $\mathbf{f}_2$, and the ultimate dense feature $\mathbf{f}_{\text{dense}}$. During the up-sampling process of intermediate features, various feature up-sample techniques
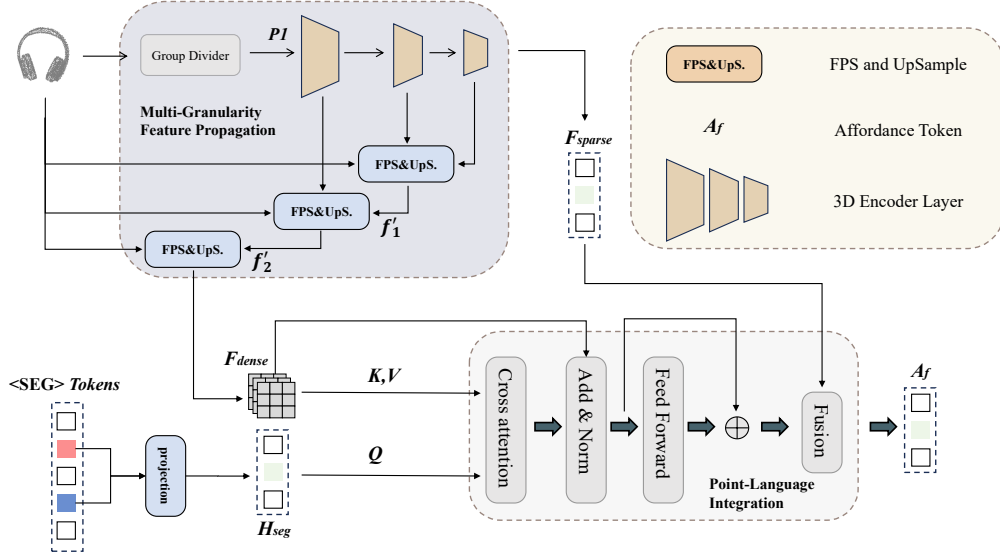
Figure 4. **Multi-Granular Language-Point Integration Module.** We propose an interaction module between `<SEG>` tokens from LLM and point features from the 3D vision encoder, to synergize both reasoning and segmentation in a cohesive framework. This module consists of the multi-granular feature propagation process, and the point-language integration stage.

are adopted, inspired by PointNet++ [32] and DGCNN [42]. More details about the multi-granular feature propagation process are revealed in the supplementary details.

During the point-language integration stage, the model takes the dense point cloud features $\mathbf{f}_{\text{dense}}$, the sparse point cloud features $\mathbf{f}_{\text{sparse}}$ and the instruction-rich $\mathbf{H}_{\text{seg}}^{(i)}$ as input. The $\mathbf{H}_{\text{seg}}^{(i)}$ and dense features $\mathbf{f}_{\text{dense}}$ are used as $\mathbf{Q}$ and $\mathbf{K}$, $\mathbf{V}$ respectively to perform cross-attention, and FFN, and the results obtained are fused with the sparse features $\mathbf{f}_{\text{sparse}}$ to get $\mathbf{A}_f^{(i)}$. Finally, the decoder takes $\mathbf{A}_f^{(i)}$ as the input to get affordance mask $\tilde{\mathbf{y}}_{\text{mask}}$

$$\mathbf{A}_f^{(i)} = \mathcal{G}(\mathbf{f}_{\text{dense}}, \mathbf{f}_{\text{sparse}}, \mathbf{H}_{\text{seg}}^{(i)}), \quad \tilde{y}_{\text{mask}} = \mathcal{D}(\mathbf{A}_f^{(i)}), \quad (3)$$

where $\mathcal{G}$ denotes point-language integration, and $\mathcal{D}$ denotes the decoder.

4.3. Training objectives

Our objective is to train an end-to-end MLLM capable of generating diverse texts while simultaneously predicting point-wise affordance masks. To this end, we employ auto-regressive cross-entropy loss $\mathcal{L}_c$ for text generation, Dice loss $\mathcal{L}_d$ and Binary Cross-Entropy loss $\mathcal{L}_b$ for guiding the segmentation mask prediction.

$$\mathcal{L} = \lambda_c \mathcal{L}_c(\mathbf{y}_{\text{txt}}, \tilde{\mathbf{y}}_{\text{txt}}) + \lambda_b \mathcal{L}_b(\mathbf{y}_{\text{mask}}, \tilde{\mathbf{y}}_{\text{mask}}) + \\ \lambda_d \mathcal{L}_d(\mathbf{y}_{\text{mask}}, \tilde{\mathbf{y}}_{\text{mask}}), \quad (4)$$

where the weights $\lambda_c, \lambda_b, \lambda_d$ are utilized to balance the different loss items.

5. Experiment

We conduct extensive experiments to evaluate the effectiveness of our proposed dataset, task, and method, including both Single and Sequential Affordance segmentation tasks. In Sec. 5.1, we assess the capability of our model to ground the Single Affordance with simple instruction. Sec. 5.2 studies a more challenging task where the model is requested to predict the sequential affordances. Various ablation experiments on our model are performed in Sec. 5.3.

Implementation Details. We employ ShapeLLM [34] as our 3D MLLM module in this paper, with the ShapeLLM-7B checkpoint as the default setting and we freeze the 3D encoder during training. We adopt Uni3D as the 3D vision encoder to enhance the 3D dense prediction tasks. Unless otherwise stated, the projection layer is implemented as a multi-layer perceptron. We employ LoRA [10] for efficient fine-tuning and set the rank of LoRA to 8 by default. Additionally, we utilize AdamW [23] optimizer with the learning rate and weight decay set to 0.0002 and 0, respectively. We adopt a cosine learning rate scheduler, with the warm-up iteration ratio set to 0.03. All attentions in ShapeLLM [34] are replaced by flash-attention [5] during training. The training is done on one A100 GPU for 10 epochs for the main experiments and during training, we use all mentioned datasets in Sec. 3 for joint training by leveraging task-specific prompts. For evaluation on a specific dataset, we finetune the trained model on the corresponding dataset.

Evaluation Metrics and Baseline. To provide a comprehensive and effective evaluation, we follow previous works

	Method	mIoU↑	AUC↑	SIM↑	MAE↓
Seen	ReferTrans [15]	11.4	77.2	0.449	0.135
	ReLA [17]	12.1	76.3	0.480	0.130
	3D-SPS [25]	10.1	75.2	0.413	0.141
	IAGNet [50]	14.2	81.7	0.510	0.117
	PointRefer [16]	16.3	84.3	0.568	0.108
	Ours	**19.5**	**86.9**	**0.594**	**0.098**
Unseen	ReferTrans [15]	9.1	67.4	0.427	0.151
	ReLA [17]	9.3	68.2	0.423	0.147
	3D-SPS [25]	7.1	66.9	0.397	0.162
	IAGNet [50]	11.7	73.6	0.438	0.143
	PointRefer [16]	12.4	76.1	0.502	0.132
	Ours	**13.8**	**82.4**	**0.518**	**0.128**
Sequential	ReferTrans* [15]	10.8	74.5	0.425	0.142
	ReLA* [17]	11.4	74.8	0.463	0.136
	3D-SPS* [25]	9.9	73.1	0.407	0.148
	IAGNet* [50]	13.5	78.2	0.496	0.131
	PointRefer* [16]	14.3	80.7	0.521	0.124
	Ours	**14.6**	**84.2**	**0.573**	**0.118**

Table 2. **Main Results.** The overall results of all comparative methods, the best results are in bold. Seen and Unseen are two partitions of the Single Affordance segmentation dataset. AUC and mIoU are shown in percentage. * means that baseline methods use **ground-truth** sequential order as all existing methods cannot predict sequential affordances.

and finally chose four evaluation metrics: Area Under the Curve (AUC) [22], Mean Intersection Over Union (mIOU) [37], SIMilarity (SIM) [39] and Mean Absolute Error (MAE) [43].To the best of our knowledge, LASO [16] is the most similar to our work. For a thorough comparison of our method, we conduct comparisons based on its setup in the Single Affordance segmentation task, comparisons with other baselines were also implemented following the approach mentioned in it. While in the sequential affordance segmentation task, we offer sequential information to these models enabling them to perform "sequential" reasoning.

5.1. Results on Language-Guided Single Affordance Segmentation Task

As detailed in Table. 2, our model demonstrates superior performance across all evaluation metrics compared to the baseline methods. Unlike conventional segmentation tasks, language-guided Single Affordance segmentation demands not just identification but the integration of perception and cognition, necessitating the model's reasoning capabilities and access to world knowledge. Existing approaches struggle with implicit queries due to their lack of integration of perception and cognition, which further underscores the task's inherent challenges. In contrast, our model lever-

ages MLLMs to bridge this gap, demonstrating superior performance by comprehending and interpreting the queries accurately.

5.2. Results on Language-Guided Sequential Affordance Reasoning Task

The sequential affordance reasoning task implies the ability to infer multiple affordances from a single text, which requires a more profound integration of cognitive and perceptual capabilities compared to the Single Affordance segmentation task. In our model, to make the Multimodal Large Language Model (MLLM) more comprehensible, we use the format <SEG> to represent the sequence of affordances and decode to obtain the affordance mask based on them. The previous models did not possess this capability. To compare them with our method, we used GPT to decompose the original instructions into new sequence instructions, then fed the decomposed instructions separately as inputs to these models, enabling them to perform "sequential" reasoning. The main results are shown in Table. 2, and we speculate that the reason our model performs better is that, compared to LASO [16] which merely uses language models to encode the input text, we have introduced a multimodal 3D large language model. This model has a much stronger capability for integrating 3D data and text than a purely linguistic model, and it possesses a richer knowledge of the world, enabling it to handle multimodal sequential tasks more effectively. The sequential reasoning results are visualized in Fig. 5 (b), where our model can understand how a task is connected with sequential actionable affordances involving multi-objects. This ability is not only attributed to the challenging benchmark collected from diverse sources but also the powerful world knowledge internalized in 3D MLLMs.

5.3. Ablation Study

We conduct various ablation studies to assess the impact of different model implementations on our model SeqAfford's performance, including the multi-granular language-point integration module and the different choices of 3D vision encoder backbones.

Multi-Granular Language-Point Integration Module. Introducing the Multi-Granular Language-Point Integration (MGLP) module results in a substantial improvement over the baseline, as shown in Table. 3. This underscores our method's capability to minimize information loss and this indicates that our method enables deep integration of high-dimensional semantic features representing instructions with dense features from point clouds and semantically rich but sparse features from point clouds, thereby making affordance reasoning more effective.

Choice of 3D Vision Encoder Backbone. We conducted ablation experiments to study the influence of the 3D Vision Encoder backbone, and we investigated the performance of

USER: When slicing bread with this knife, at what position should I place my hand for optimal control? Please output the segmentation result.

ASSISTANT: You should place your hand at the handle of the knife <SEG>.

USER: When using the knife for cutting through dense root vegetables, which part of the knife should I focus on? Please output the segmentation result.

ASSISTANT: When cutting, the blade of the knife <SEG> mostly affords this operation, you should focus mainly on this.

USER: If I need to stab a piece of fruit to test its ripeness, which area of the knife is the most important? Please output the segmentation result.

ASSISTANT: The most important area is the tip of the knife <SEG>.

(a) Different Affordance Types via Different Instructions on the Same Object

USER: When I step outside and notice strong sunlight, how would I handle this item to protect myself ?

ASSISTANT: First you should grasp the hat <SEG>.

ASSISTANT: Then put the hat <SEG> on your head.

USER: What should I do when I want to use the scissors to cut window grilles?

ASSISTANT: First you should hold the scissors <SEG> in a proper way.

ASSISTANT: Then use the scissors <SEG> to cut the window grilles carefully.

USER: How should I properly store my earphone to keep it safe and in good condition when I do not use it ?

ASSISTANT: First you should grasp the earphone<SEG> from your head.

ASSISTANT: Then put it on the table <SEG> to prevent it from being damaged.

(b) Sequential Affordance Reasoning on Both Single Object and Multiple Objects
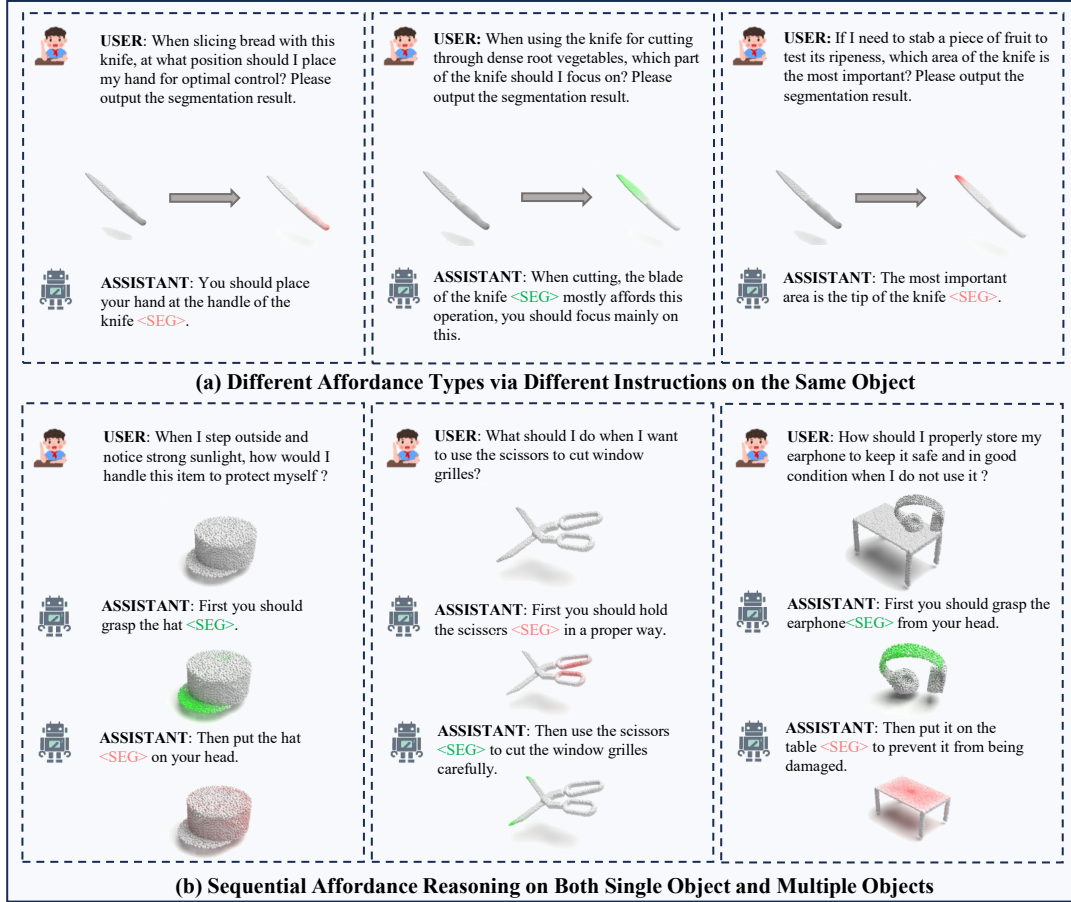
Figure 5. Qualitative results of our model. SeqAfford understands human instruction and accurately segments the target affordance.

3D SeqAfford with some alternative backbones. As shown in Table. 4, Uni3D [53] performs better in this task due to its strong representation ability, Thus, we set it as our default 3D vision encoder backbone.

Variants	Task	mIoU↑	AUC↑	SIM↑	MAE↓
w/o MGLP	Single	12.1	83.4	0.552	0.117
Ours	Single	**19.5**	**86.9**	**0.594**	**0.098**
w/o MGLP	Sequential	11.7	80.3	0.518	0.129
Ours	Sequential	**14.6**	**84.2**	**0.573**	**0.118**

Table 3. Ablation study on multi-granular language-point module.

3D Vision Backbone	mIoU↑	AUC↑	SIM↑	MAE↓
ULIP [49]	17.9	84.8	0.574	0.109
OpenShape [19]	18.4	85.3	0.582	0.103
Recon++ [34]	19.1	86.4	0.588	0.099
Uni3D [53]	**19.5**	**86.9**	**0.594**	**0.098**

Table 4. Ablation study on the choices of 3D vision backbone.

6. Conclusion

In this paper, we introduce a novel task called the Sequential 3D Affordance Reasoning Task, designed to address complex user intentions that may involve sequential or long-horizon subtasks, transcending the conventional single-object, single-affordance paradigm. Our contributions include the construc-
tion of a large-scale 3D affordance benchmark comprising 180K instruction-point pairs collected from diverse sources, and the development of an advanced 3D multimodal LLM that leverages world knowledge to interpret complex user intentions, producing sequential, actionable affordance maps with reasonable explanations. SeqAfford advances 3D affordance reasoning by integrating LLM capabilities, enhancing the model's ability to handle complex, sequential tasks in real-world contexts. Our future work will focus on scene-level fine-grained reasoning to enable more sophisticated and context-aware affordance segmentation for embodied intelligent agents.

7. Acknowledgement

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2, 3

[3] Hengshuo Chu, Xiang Deng, Xiaoyang Chen, Yinchuan Li, Jianye Hao, and Liqiang Nie. 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. *arXiv preprint arXiv:2502.20041*, 2025. 2

[4] Meng Chu and Xuan Zhang. Iris: Interactive responsive intelligent segmentation for 3d affordance analysis. *arXiv preprint arXiv:2409.10078*, 2024. 3

[5] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 6

[6] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scene-fun3d: fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14531–14542, 2024. 2

[7] 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. 2, 4

[8] Xianqiang Gao, Pingrui Zhang, Delin Qu, Dong Wang, Zhigang Wang, Yan Ding, Bin Zhao, and Xuelong Li. Learning 2d invariant affordance knowledge for 3d affordance grounding. *arXiv preprint arXiv:2408.13024*, 2024. 2

[9] 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. 3

[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. *arXiv preprint arXiv:2106.09685*, 2021. 6

[11] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2024. 2

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2. Minneapolis, Minnesota, 2019. 2

[13] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 5

[14] 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*, pages 3086–3096, 2024. 2

[15] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. 7

[16] 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, 4, 5, 7

[17] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 7

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[19] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36, 2024. 8

[20] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019. 2

[21] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 3

[22] 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. 7

[23] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Dongyue Lu, Lingdong Kong, Tianxin Huang, and Gim Hee Lee. Geal: Generalizable 3d affordance learning with cross-modal consistency. *arXiv preprint arXiv:2412.09511*, 2024. 2

[25] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage

3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022. 7

[26] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 2, 4

[27] 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*, pages 1666–1677. PMLR, 2022. 4

[28] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704*, 2024. 3

[29] 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

[30] Chuanruo Ning, Ruihai Wu, Haoran Lu, Kaichun Mo, and Hao Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *Advances in Neural Information Processing Systems*, 36:4585–4596, 2023. 2

[31] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2

[32] 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. 6

[33] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pages 28223–28243. PMLR, 2023. 5

[34] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. *arXiv preprint arXiv:2402.17766*, 2024. 2, 3, 5, 6, 8

[35] 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. 2

[36] 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. 3

[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. 7

[38] Yawen Shao, Wei Zhai, Yuhang Yang, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. *arXiv preprint arXiv:2411.19626*, 2024. 2

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

[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. 5

[41] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[42] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 6

[43] 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. 7

[44] 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*, 36:60966–60983, 2023. 2

[45] Shijie Wu, Yihang Zhu, Yunao Huang, Kaizhen Zhu, Jiayuan Gu, Jingyi Yu, Ye Shi, and Jingya Wang. Afforddp: Generalizable diffusion policy with transferable affordance. *arXiv preprint arXiv:2412.03142*, 2024. 2

[46] 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. 4

[47] Peiran Xu and MU Yadong. Weakly-supervised affordance grounding guided by part-level semantic priors. In *The Thirteenth International Conference on Learning Representations*. 2

[48] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. PointLLM: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023. 2, 3

[49] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 8

[50] 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. 2, 4, 5, 7

[51] 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. 2

[52] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 3

[53] Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, and Yu Qiao. Uni3d: A unified baseline for multi-dataset 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9253–9262, 2023. 8

[54] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 3

[55] 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. 2, 3