SEQAFFORD: SEQUENTIAL 3D AFFORDANCE REASONING VIA MULTIMODAL LARGE LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

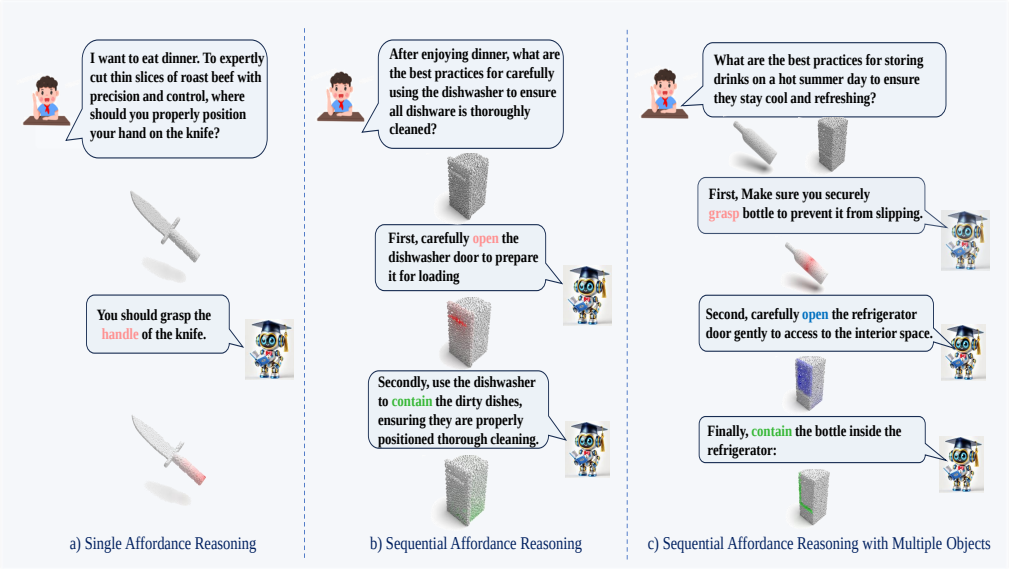


Figure 1: **Sequential 3D affordance reasoning task with different types of interactions.** We introduce SeqAfford, a Multi-Modal Language Model (MLLM) capable of serialized affordance inference implied in human instructions: a) single affordance reasoning; b) sequential affordance Reasoning; c) sequential 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 with multiple objects involved. 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 sequence of segmentations. 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.

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 offers visual cues that suggest potential actions to embodied systems, 3D affordance 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.

Pioneering works on 3D affordance have primarily focused on the single-object, single-affordance paradigm, where one afforance map can be grounded from either an affordance category Deng et al. (2021); Nguyen et al. (2023) or a 2D demonstration image Yang et al. (2023). More recently, language models have been employed to pair 3D objects with natural language questions Li et al. (2024), each designed to probe a distinct affordance part. Unfortunately, current systems are incapable of actively reasoning based on complex user intentions, breaking them down into actionable primitives, and formulating a chain of affordances derived from each primitive. To further enable real-world physical interactions, the inherent complexity of human instructions can no longer be neglected, which often involves handling multiple objects and navigating through sequential affordances. For example, when agents are required to "*use the microwave to reheat the food inside the bowl*", they must first *grasp* the bottle, then *open* the microwave before using it to *contain* the bottle, as shown in Figure 1. From this perspective, the multi-object sequential reasoning ability is indispensable for shaping the next-generation affordance systems.

Recently, Large-Language Models (LLMs) Alayrac et al. (2022); Zhu et al. (2023); Ouyang et al. (2022) 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) Qi et al. (2024); Xu et al. (2023) have further expanded their possibilities in understanding various object shapes in 3D world. However, even current 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 diversity of instruction data, the instructions are generated in four different ways. Supported by this benchmark, we introduce our model SeqAfford that 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 are the first to propose a novel approach for 3D affordance segmentation using multi-modal large language models, leveraging the common-sense knowledge acquired during pre-training to enhance affordance reasoning. Our model achieves state-of-the-art performance across various settings and demonstrates robust generalization capabilities to open-world scenarios.

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 (Deng et al., 2021), built on point cloud data from PartNet (Mo et al., 2019), was the first to construct a fine-grained 3D affordance dataset, establishing a benchmark for 3D affordance research. Building on this, Yang et al. (2023) 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. (2024) introduced a language-based affordance segmentation task, promoting the integration of natural language and affordance understanding. Following this, Chu & Zhang (2024) 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 such as Flamingo (Alayrac et al., 2022), MiniGPT-4 (Zhu et al., 2023), and LLaVA (Liu et al., 2024a), 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 such as GPT4RoI (Zhang et al., 2023b), VisionLLM (Wang et al., 2024), InternGPT (Liu et al., 2023b), and Ferret (You et al., 2023) 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 (Xu et al., 2023) 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 (Qi et al., 2024) is built upon the enhanced 3D encoder RECON++, which strengthens geometric understanding through multi-view image distillation. Other models, such as 3D-LLM (Hong et al., 2023), leverage 2D foundational models, like CLIP ViT (Radford et al., 2021), 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 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 (Deng et al., 2021), encompassing both simple and complex

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.

Method	#Sequential	#Multi	#Part	#Point Cloud	#Instruction-Point Pairs
3D AffordanceNet (Deng et al., 2021)	×	×	✓	23K	×
O2O-Afford (Mo et al., 2022)	×	×	✓	1.7k	×
Partafford (Xu et al., 2022)	×	×	✓	25k	×
IAGNet (Yang et al., 2023)	×	×	✓	7k	×
LASO (Li et al., 2024)	×	×	✓	8.4k	19k
Ours	✓	✓	✓	23K	180k

types of intentions, which includes 162,386 instruction-point cloud pairs in the single affordance segmentation setting and 628,847 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 (Deng et al., 2021). 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 (Achiam et al., 2023). Unlike LASO (Li et al., 2024) 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 (Deng et al., 2021) to trace back to the Mesh-rendered images in the PartNet (Mo et al., 2019) dataset. Additionally, we collect HOI images corresponding to the affordance types from IAGNet (Yang et al., 2023) to alleviate GPT's hallucination issues, enabling better understanding. Figure 2. illustrates our ways of generating instructions in detail.

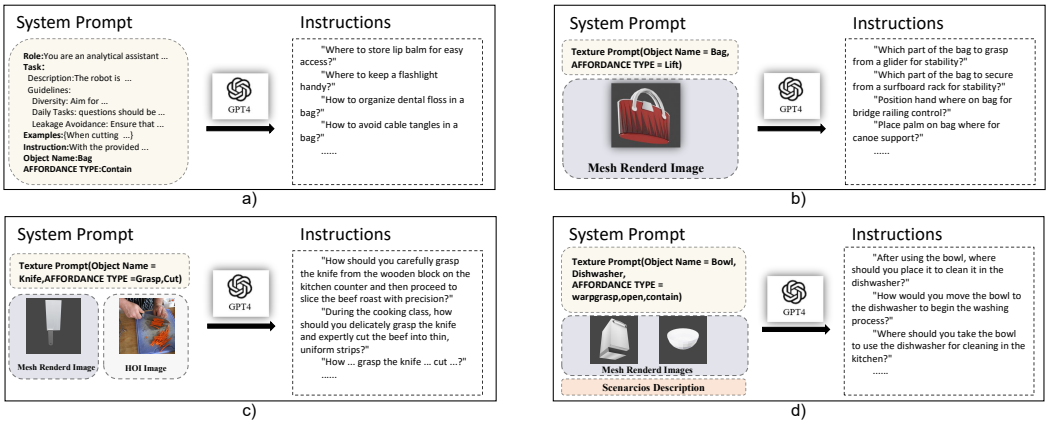


Figure 2: **Preparing the instructions.** We prompt GPT-4o to generate diverse instructions based on 4 types of system prompts containing different modalities as input. Instructions are generated based on input prompts with modalities from a) purely textual affordance type, object name; b) the mesh-rendered 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.

3.2 Statistics and Annlysis

Our dataset comprises 162,386 instruction-point cloud pairs in the single affordance segmentation setting and 628,847 pairs in the sequential affordance segmentation setting, making up a total of

Table 2: **Dataset Statistics**. This table showcases the distribution of our instruction-tuning benchmark for Sequential Affordance Reasoning task, which introduces a wide range of variances and realistic simulations, including seen/unseen settings, single/sequential scenarios.

Task	Setting	Train		Test	
		Shapes	Instruction-Point Cloud Pairs	Shapes	Instruction-Point Cloud Pairs
Single	Seen	16084	157,820	2287	4566
Single	Unseen	14307	124,140	2287	4566
Sequential	Seen	8156	531,393	2,946	97454

18,371 point cloud instances across 23 object categories. We have visualized word clouds for the object categories, affordance types, and the introduction in our dataset, demonstrating the richness of our dataset.

(a) Instructions (b) Affordance Types (c) Object Categories

Figure 3: Word clouds of (a) instructions, (b) affordance types, and (c) 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". We present the statistical information of our dataset in Table 2 and three word-clouds in Figure 3 to illustrate the diversity of our dataset.

Inspired by the methodologies presented in LASO (Li et al., 2024) and IAGNet (Yang et al., 2023), 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 capacity to generalize affordance knowledge. In this setting, certain affordance-object pairings are deliberately excluded from the training set but introduced during testing. For example, while the model may learn to grasp objects such as bags and mugs during training, it is required to generalize the 'grasp' affordance to earphones—a combination not encountered during training.

4 Method

4.1 Architecture Overview

The overall architecture of SeqAfford is presented in Figure 4. 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 ShapeLLM is recently pretrained

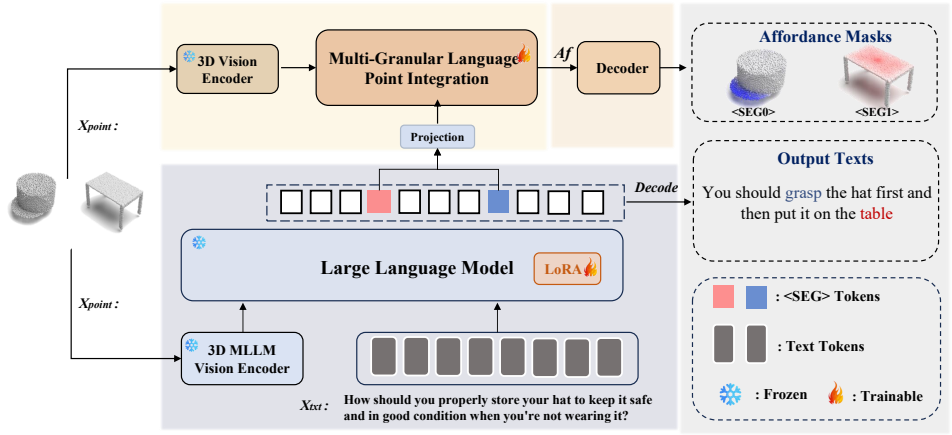


Figure 4: **Main Pipeline.** Given the point clouds of the target objects and a complex human instruction, SeqAfford first reasons from this instruction and decomposes it into several hidden <SEG> tokens extracted from the last-layer embeddings, each representing an intermediate affordance segmentation 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.

for understanding various embodied interactions. In light of this, we adopt ShapeLLM (Qi et al., 2024) as our backbone, denoted as $\mathcal{F}$, where the point cloud encoder ReCon++ is pre-trained via multi-view distillation based on ReCon Qi et al. (2023), and the LLM is drawn from LLaMa Touvron et al. (2023). Previous work on 3D affordance tasks typically employed 3D backbones Yang et al. (2023); Nguyen et al. (2023) or utilized separate point-language encoders Li et al. (2024), which may fall short in reasoning and open-world generalization abilities. Here, we take a 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 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 Multi-object 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 Lai et al. (2024).

Formally, when provided with M point clouds $\mathbf{X}_{\text{point}} \in \mathbb{R}^{M\times N\times 3}$ and a text instruction $\mathbf{X}_{\text{txt}}$ that demonstrates the user intentions on these potential objects, the 3D MLLM absorbs these multi-modal 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

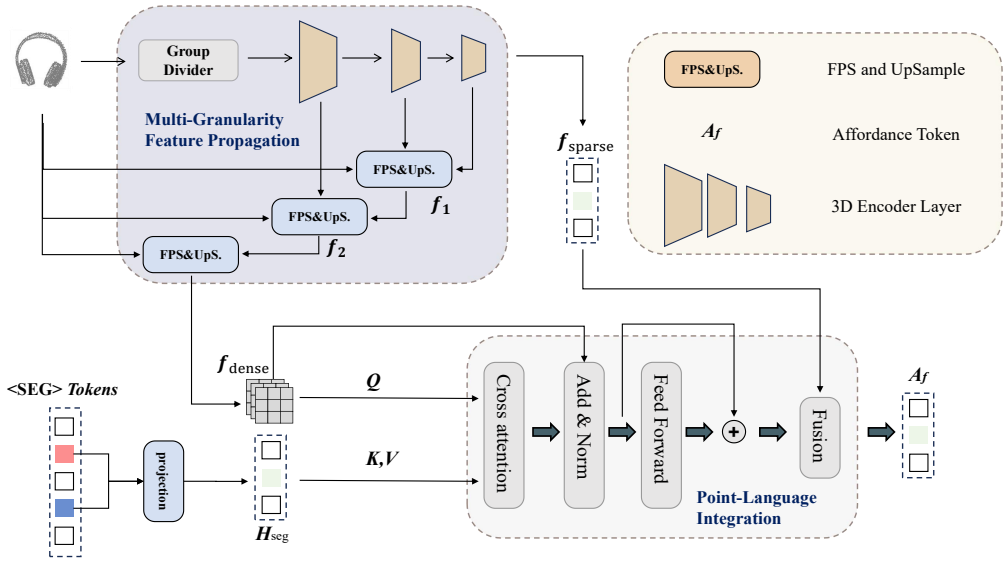


Figure 5: **Multi-Granular Language-Point Integration Module.** We propose an interaction module between <SEG> tokens from LLM and point features from 3D vision encoder, to synergize both reasoning and segmentation in a cohesive framework. This module consists of the multi-granular feature propagation process, and the point-language integration stage.

stages: 1) multi-granular feature propagation process, which iteratively up-samples the point cloud features into dense features with multiple granularities considered; 2) 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. 5, we hierarchically up-sample (UpS. in Fig. 5) the intermediate features from the 3D encoder, and propagate features through farthest point sampling (FPS in Fig. 5) process to sequentially generate $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, and the ultimate dense feature $\boldsymbol{f}_{\text{dense}}$. During the up-sampling process of intermediate features, various feature up-sample techniques are adopted, inspired by PointNet++ (Qi et al., 2017) and DGCNN Wang et al. (2019). 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 $\boldsymbol{f}_{\text{dense}}$, the sparse point cloud features $\boldsymbol{f}_{\text{sparse}}$ and the instruction-rich $\mathbf{H}_{\text{seg}}^{(\mathbf{i})}$ as input. The dense features $\boldsymbol{f}_{\text{dense}}$ and $\mathbf{H}_{\text{seg}}^{(i)}$ are used as Q and K, V respectively to perform cross-attention, and FFN, and the results obtained are fused with the sparse features $\boldsymbol{f}_{\text{sparse}}$ to get $\mathbf{A_f}$. Finally, the decoder takes A_f as the input to get affordance mask $\tilde{\mathbf{y}}_{\text{mask}}$:

$$\mathbf{A_f} = \mathcal{G}(\boldsymbol{f}_{\text{dense}}, \boldsymbol{f}_{\text{sparse}}, \mathbf{H}_{\text{head}}), \quad \tilde{\mathbf{y}}_{\text{mask}} = \mathcal{D}(\mathbf{A_f}), \tag{3}$$

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 $\mathcal{L}_{\text{b}}$ for guiding the segmentation mask prediction.

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

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 (Qi et al., 2024) as our 3D vision-language 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 (Hu et al., 2021) for efficient fine-tuning and set the rank of LoRA o 8 by default. Additionally, we utilize AdamW (Loshchilov, 2017) optimizer with the learning rate and weight decay set to 0.00001 and 0, respectively. We adopt a cosine learning rate scheduler, with the warm-up iteration ratio set to 0.03. All attentions in ShapeLLM are replaced by flash-attention (Dao et al., 2022) 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 and finally chose four evaluation metrics: Area Under the Curve (AUC) (Lobo et al., 2008), Mean Intersection Over Union (mIOU) (Rahman & Wang, 2016), SIMilarity (SIM) (Swain & Ballard, 1991) and Mean Absolute Error (MAE) (Willmott & Matsuura, 2005).To the best of our knowledge, LASO (Li et al., 2024) 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 3, 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 leverages MLLMs to bridge this gap, demonstrating superior performance by comprehending and interpreting the queries accurately.

5.2 RESULTS ON LANGUAGE-GUIDED SEQUENTIAL AFFORDANCE SEGMENTATION TASK

The sequential affordance segmentation 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 3, we believe that the reason our model performs better is that, compared to LASO (Li et al., 2024) 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.

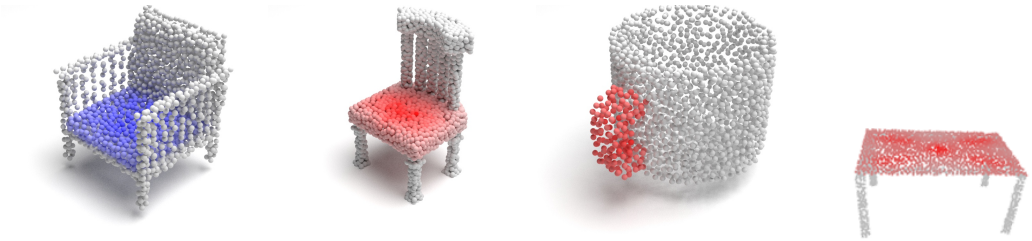
We present the 3D visualization rendering results of point cloud segmentation in Figure 6. As shown, when the instruction is "To sit on a chair with a lightweight design, which section should provide portability?";" To sit comfortably for a long period, which part of the chair should support your

Table 3: **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 lack the capability to predict sequential affordances.

	Method	*mIoU*↑	*AUC*↑	*SIM*↑	*MAE*↓
Seen	ReferTrans (Li & Sigal, 2021)	11.4	77.2	0.449	0.135
	ReLA (Liu et al., 2023a)	12.1	76.3	0.480	0.130
	3D-SPS (Luo et al., 2022)	10.1	75.2	0.413	0.141
	IAGNet (Yang et al., 2023)	14.2	81.7	0.510	0.117
	PointRefer (Li et al., 2024)	16.3	84.3	0.568	0.108
	Ours	**19.5**	**86.9**	**0.594**	**0.098**
Unseen	ReferTrans (Li & Sigal, 2021)	9.1	67.4	0.427	0.151
	ReLA (Liu et al., 2023a)	9.3	68.2	0.423	0.147
	3D-SPS (Luo et al., 2022)	7.1	66.9	0.397	0.162
	IAGNet (Yang et al., 2023)	11.7	73.6	0.438	0.143
	PointRefer (Li et al., 2024)	12.4	76.1	0.502	0.132
	Ours	**13.8**	**82.4**	**0.518**	**0.128**
Squential	ReferTrans* (Li & Sigal, 2021)	10.8	74.5	0.425	0.142
	ReLA* (Liu et al., 2023a)	11.4	74.8	0.463	0.136
	3D-SPS* (Luo et al., 2022)	9.9	73.1	0.407	0.148
	IAGNet* (Yang et al., 2023)	13.5	78.2	0.496	0.131
	PointRefer* (Li et al., 2024)	14.3	80.7	0.521	0.124
	Ours	**16.6**	**83.2**	**0.573**	**0.118**

thighs?";"which part of the mug should you hold when you want you use it to drink"" and "If you want to put off your hat,where should you put it?"Our model can segment the corresponding areas of each object.

To conclusion,SeqAfford demonstrates excelling performance in terms of both Single Affordance Reasoning and Sequential Affordance Reasoning tasks. Previous models either fail to comprehend complex human instructions, or still cannot decompose a cumbersome user intention into several actionable primitives. By contrast, our approach can not only reason from the various forms of user intentions, but also is capable of producing accurate and high-quality affordance segmentation results. More illustrations about the sequential affordance reasoning task are given in the supplementary material due to space limitations, more details are provided in Appendix E.



(a) Chair Sit (b) Chair Sit (c) Mug Grasp (d) Table Support

Figure 6: Visualization Results

5.3 Ablation Study

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

Multi-Granularity Language-Point Integration Module. Introducing the Multi-Granularity Language-Point Integration (MGLP) module results in a substantial improvements over the baseline, as shown in Table 4. 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.

Table 4: Ablation study on the semantic-point fusion module.

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**

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 3D SeqAfford with some alternative backbone. As shown in table 5, Uni3D (Zhang et al., 2023a) performs better in this task due to its strong representation ability, Thus, we set it as our default 3D vision encoder backbone.

Table 5: Ablation study on the choices of 3D vision backbone.

3D Vision Backbone	*mIoU*↑	*AUC*↑	*SIM*↑	*MAE*↓
OpenShape (Liu et al., 2024b)	18.4	85.3	0.582	0.103
Recon++ (Qi et al., 2024)	19.1	86.4	0.588	0.099
Uni3D (Zhang et al., 2023a)	**19.5**	**86.9**	**0.594**	**0.098**

6 CONCLUSION, LIMITATION, AND FUTURE WORK

In this paper, we have advanced the field of 3D affordance segmentation by introducing the Sequential 3D Affordance Reasoning task. Unlike traditional paradigms that focus on single-object, single-affordance scenarios, our task addresses the complexity of sequential affordances involving multiple objects and nuanced user instructions. To support this new task, we developed an extensive benchmark consisting of 180K instruction-point cloud pairs, laying the foundation for future research in this domain. Our proposed model, SeqAfford, integrates multimodal large language models with additional affordance segmentation capabilities, utilizing world knowledge and fine-grained affordance grounding in a cohesive framework. Moreover, the introduction of a multi-granular language-point integration module has significantly enhanced 3D dense prediction capabilities. Extensive experimental evaluations demonstrate that SeqAfford outperforms existing methods, showcasing robust sequential reasoning abilities and remarkable generalization to open-world scenarios.

While our work represents a significant advancement in 3D affordance segmentation, there are still some limitations. One primary constraint is the reliance on high-quality 3D point cloud data, which may not always be readily available or easy to obtain in real-world settings. Additionally, our research primarily focuses on object-centric reasoning and lacks scene-level fine-grained reasoning, which is crucial for embodied intelligent agents.

For future work, we plan to address these limitations through several approaches. First, we aim to develop methods for augmenting or synthetically generating diverse 3D point cloud data to improve the robustness of our models in varied environments. Additionally, we intend to extend our research to encompass scene-level fine-grained reasoning, enabling more sophisticated and context-aware affordance segmentation that aligns with the complexities faced by embodied intelligent agents in real-world scenarios.

REFERENCES

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.

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.

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

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.

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*, pp. 1778–1787, 2021.

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.

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.

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*, pp. 9579–9589, 2024.

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.

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*, pp. 14251–14260, 2024.

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*, pp. 23592–23601, 2023a.

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

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, 2024b.

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*, 2023b.

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.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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*, pp. 16454–16463, 2022.

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*, pp. 909–918, 2019.

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*, pp. 1666–1677. PMLR, 2022.

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)*, pp. 5692–5698. IEEE, 2023.

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.

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.

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*, pp. 28223–28243. PMLR, 2023.

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.

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*, pp. 8748–8763. PMLR, 2021.

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

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

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.

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.

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.

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.

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.

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.

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*, pp. 10905–10915, 2023.

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.

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*, pp. 9253–9262, 2023a.

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*, 2023b.

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.

A TEXTURE PROMPT FORMAT

This section outlines the specific design of the Texture Prompt component included in the four generative paradigms demonstrated earlier. By designing the Texture Prompt, it helps to enhance GPT-4's understanding and generalization ability for the relevant tasks, providing suitable introductions for subsequent tasks. We have utilized the principles of prompt engineering to design modules such as Role, Task, Example, and Instruction for the Texture prompt. Additionally, we have formatted the input in JSON format to improve GPT-4's comprehension of the text, standardizing the input of Object Name and Affordance Type.

A.1 ROLE

In the role module, we have preset the Texture Prompt template as follows: "Role: You are an analytical assistant specializing in robotic affordance grounding. Your expertise is in creating questions that facilitate the training of robotic affordance grounding, enabling robots to reason about task execution, such as determining the appropriate part of an object to grasp."

A.2 TASK

In the task module, we have preset the Texture Prompt template as follows: "Task Description: You will be provided with the name of an object. The robot is expected to use this tool to perform a variety of everyday tasks. Along with the tool name, you will receive an affordance type that the object can afford and a list of questions that have already been generated for this tool. Guidelines: Diversity: Aim for a wide range of tasks, ensuring that there is no overlap with previous ones. Daily Tasks: questions should be common and representative of the ones encountered in daily life. Leakage Avoidance: Ensure that the generated questions tasks can only afford the given affordance."

A.3 EXAMPLES

Examples illustrate the types of questions that can be generated based on the object name and affordance type. These examples guide the creator of the prompt in formulating new and diverse questions. In the task module, we have preset the Texture Prompt template as follows: "Examples: When cutting a rope with these scissors, which part of it should your palm touch? If you want to lift the bag, at which point is your finger most likely to carry it? Point out the areas on the microwave ideal for opening. How would you grasp the hat to best maintain its condition? If you want to boil water, at which points on the tap would you open the water valve? And so on..."

A.4 INSTRUCTION

In the task module, we have preset the Texture Prompt template as follows: "Instruction: With the provided OBJECT NAME: '+object+' and AFFORDANCE TYPE: '+affordance types+', generate fifteen new affordance grounding question tasks. Use the HISTORY of generated tasks as a reference to ensure compliance with the diversity guideline. Output should be in the JSON format."

B VISUALIZATION OF THE RESULTS

In this section, we present more visualization details of the 3D visualization rendering results of point cloud segmentation that could not be displayed in the main text due to space limitations. It is easy to see that our SeqAfford has achieved good results in the segmentation of point clouds using the affordance attribute.

In the image set presented above, our SeqAfford system demonstrates the point cloud segmentation and rendering for a Knife object, based on the corresponding Introductions and Affordances. In Figure 7a, we initially obtain the relevant question: "When cutting through a ripe pineapple, how should you grip the knife for control?". We then employ SeqAfford to calculate the affordance for this specific query regarding the Knife. Utilizing the derived affordance, we proceed with the point cloud segmentation. The final segmentation outcome is rendered using the Python library Mitsuba, and the resulting visual can be seen in Figure 7a. It accurately showcases the point cloud segmentation.

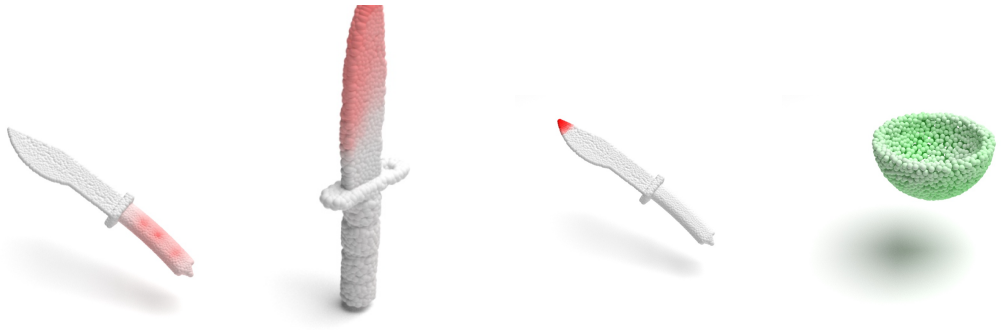
The process begins with the extraction of the question from a JSON data structure that includes a list of queries related to the grasp affordance of a knife. This particular question is chosen for its relevance to a common kitchen task. SeqAfford then analyzes the query and identifies the key affordance points on the knife that are instrumental for controlling cuts through a ripe pineapple. These points are critical for grip and control, ensuring the safety and precision required for the task.

Following the affordance calculation, our system segments the point cloud data, focusing on the areas of the knife that are pertinent to the task. This segmentation is crucial as it allows for a detailed understanding of the physical interaction with the object, which is vital for robotic manipulation and simulation.

Finally, to visualize the segmentation, we employ the Mitsuba rendering library, a powerful tool that enables the creation of realistic images from the segmented point cloud data. The rendered image in Figure 7a is a testament to the effectiveness of our approach, as it clearly delineates the areas of the knife that are optimal for gripping when cutting through a ripe pineapple.

This comprehensive process not only enhances our understanding of how objects like knives are interacted with but also provides a visual confirmation of the accuracy of our segmentation techniques.

For Figure 7b, a different affordance type, "cut," was applied. Our SeqAfford model once again delivered accurate point cloud predictions, successfully identifying the affordance associated with cutting, which is the upper side of the knife. This showcases the model's capability to adapt to various affordance types and predict their outcomes effectively. For Figure 7c and 7d, SeqAfford has effectively calculated affordance predictions based on the "cut" type, identifying the edge of the knife as the predicted affordance. The rendered results align with the predictions, confirming the accuracy of SeqAfford's calculations.



(a) knife grasp (b) knife cut (c) knife stab (d) bowl contain

Figure 7: Visualization results. This figure visualizes the affordance results given

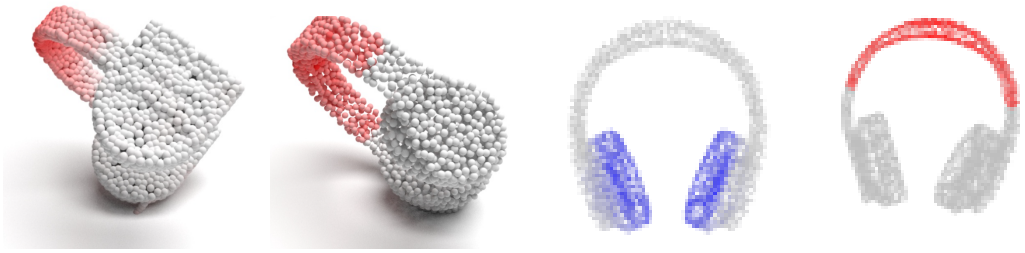
In this subsection, we primarily utilize SeqAfford for affordance prediction on the point cloud of headphones. Figures 8a and 8b both pertain to the 'grasp' affordance type of the headphones. By employing SeqAfford for prediction, it is evident that in response to the question, "When grabbing headphones to take them out, which part would you hold?", our predictions using SeqAfford for the 'grasp' related queries have yielded excellent results. Both predictions successfully identified the upper part of the headphones as the area to grip. This is clearly visible in the rendered visualizations of Figures 8a and 8b, showcasing the effectiveness of our predictions.

Our approach involves leveraging the SeqAfford system to analyze the point cloud data of the headphones and predict the most ergonomic and intuitive areas for grasping. The question posed is a common scenario that users would encounter, and our system is designed to mimic human intuition to determine the optimal points for interaction.

The success of these predictions is a testament to the accuracy and reliability of the SeqAfford system. It demonstrates the system's ability to not only understand the physical attributes of an object but also to predict how a user would naturally interact with it. The rendered images provide a visual confirmation of the predictions, allowing us to validate the results in a manner that is both intuitive and easy to understand.

Furthermore, the consistency of the predictions across two different queries related to the 'grasp' affordance type highlights the robustness of the SeqAfford system. It showcases the system's capability to generalize its understanding of affordances across a variety of potential user interactions. This level of sophistication is crucial for applications in fields such as user interface design, robotics, and virtual reality, where the ability to predict and understand user interactions is paramount.

In conclusion, the visualization results of Figure 7 serve as a compelling demonstration of the SeqAfford system's prowess in affordance prediction.



(a) Earphone grasp (b) Earphone grasp (c) Earphone listen (d) Earphone grasp

Figure 8: Visualization results

C Detail Design of Feature Propagation.

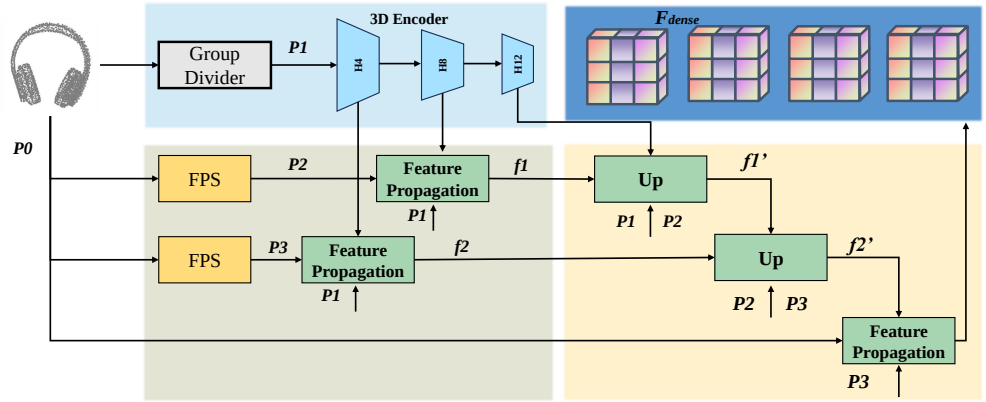


Figure 9: Detail Design of Feature Propagation.

In this section, we provide a detailed design of Feature Propagation. Our feature propagation is then applied to obtain features f_1 and f_2:

$$f_1 = \text{FP}(P_1, P_2, H_8), \quad f_2 = \text{FP}(P_1, P_3, H_4) \tag{5}$$

Next, we follow DGCNN (Wang et al., 2019) using its upsample technique to obtain f_1 and f_2:

$$f_1' = \text{Up}(P_2, f_1, P_2, H_8), \quad f_2' = \text{Up}(P_3, f_2, P_2, f_1) \tag{6}$$

Finally, we apply feature propagation once again to obtain the final dense point cloud features F_{dense}:

$$F_{\text{dense}} = \text{FP}(P_0, P_3, f_2'). \tag{7}$$