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

Supplementary Materials

Supplementary organization:

A. Code and Dataset

We will release our code and dataset once the paper is accepted.

B. Textual Prompt Format

This section outlines the specific design of the Textual Prompt component included in the four generative paradigms demonstrated earlier. By designing the Textual 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 Textual 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.We show the details in Fig. A1.

B.1. Role

In the role module, we have preset the textual 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."

B.2. Task

In the task module, we have preset the textual 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."

Table A1. **Dataset Statistics**. This table showcases the distribution of our instruction-tuning benchmark for the 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	13393	2,946	7454

B.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 textual 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...".

B.4. Instruction

In the task module, we have preset the textual 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.5. Template

Here is the template that encapsulates all the elements of the textual prompt:

B.6. Genereted Instructions Example

We illustrated detailed examples of our generated instructions in Fig. A2, A3, A4, A5

C. Format of Questions and and Answers

We have several formats of input, we show the Question List template , Explanatory Question List template and Answer List template in Fig. A6

D. Dataset Details

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. We have visualized word clouds for the object categories, affordance types, and the introduction in our dataset, demonstrating the richness of our dataset. Fig. A7 shows the distribution of the generated question length and Fig. A8 shows the words cloud of the generated instructions.

E. Detailed Methodology

E.1. Point Upsample

Fig. A9 shows the detailed method of our upsample ways. Here, we adopt Uni3D[**?**] as our 3D encoder, which is capable of encoding point clouds to obtain semantically rich point cloud features Fp_{sparse}. To perform dense prediction tasks, we need to obtain finer-grained point cloud features from the sparse point cloud features Fp_{sparse}. Following PointNet++[**?**] and DGCNN[**?**], we utilized their feature propagation techniques to build our feature propagation process. Specifically, we extract features from the intermediate layers of the 3D encoder and use Farthest Point Sampling (FPS) to sample different numbers of

[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.

[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.
[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?

Instruction:With the provided OBJECT NAME and AFFORDANCE TYPE, generate fifteen new affordance grounding question tasks. Use the HISTORY of generated tasks as a reference to ensure compliance with the diversity guideline. The output should be in the JSON format.

Figure A1. textual Prompt Format

points from the point cloud:

$$P2 = FPS(X_{point}), \tag{1}$$

$$P3 = FPS(X_{point}). \tag{2}$$

Feature propagation is then applied to obtain features $f1$ and $f2$:

$$f1 = FP(P1, P2, H8), \tag{3}$$

$$f2 = FP(P1, P3, H4). \tag{4}$$

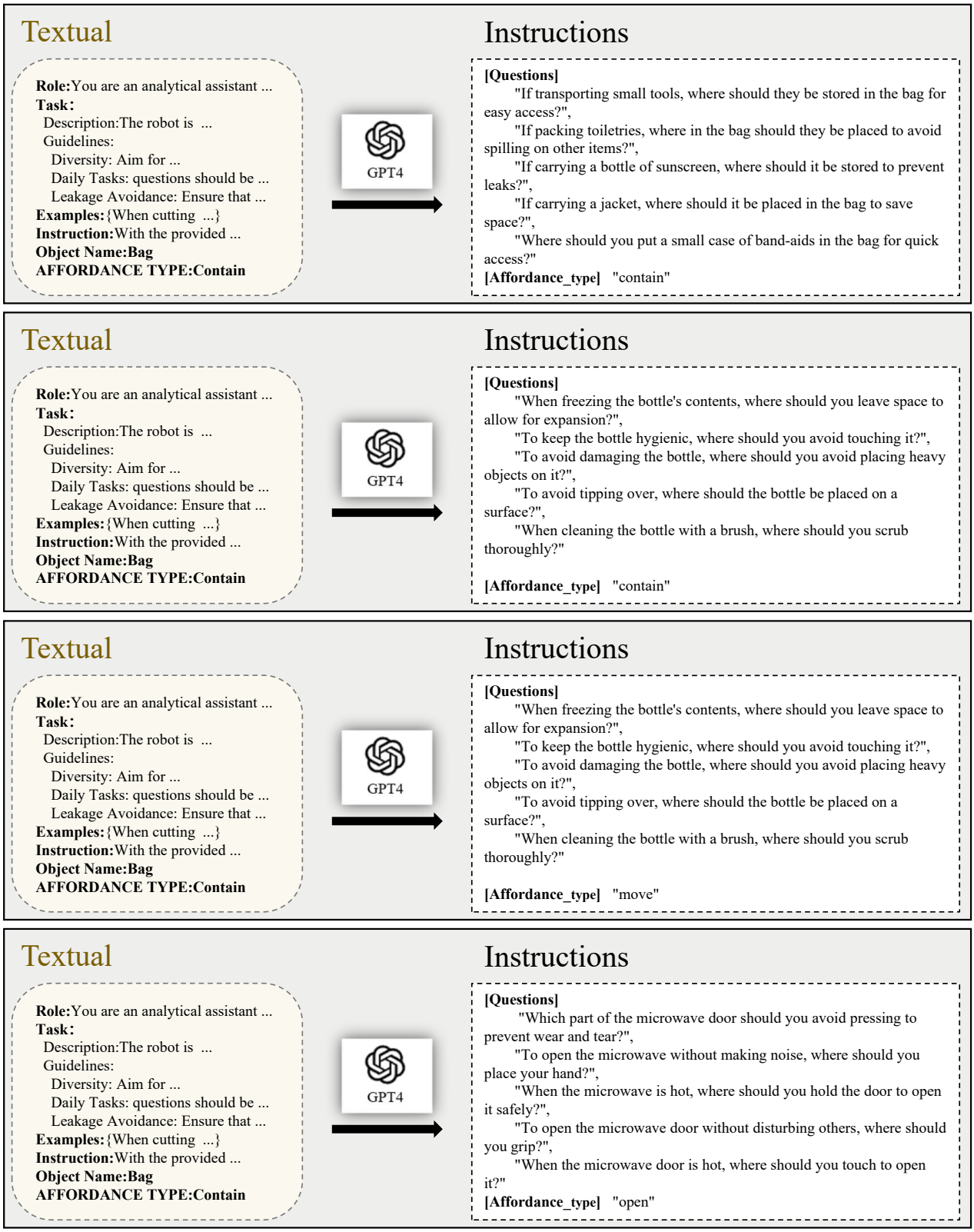


Figure A2. Genereted Instructions Example:Textual Prompt

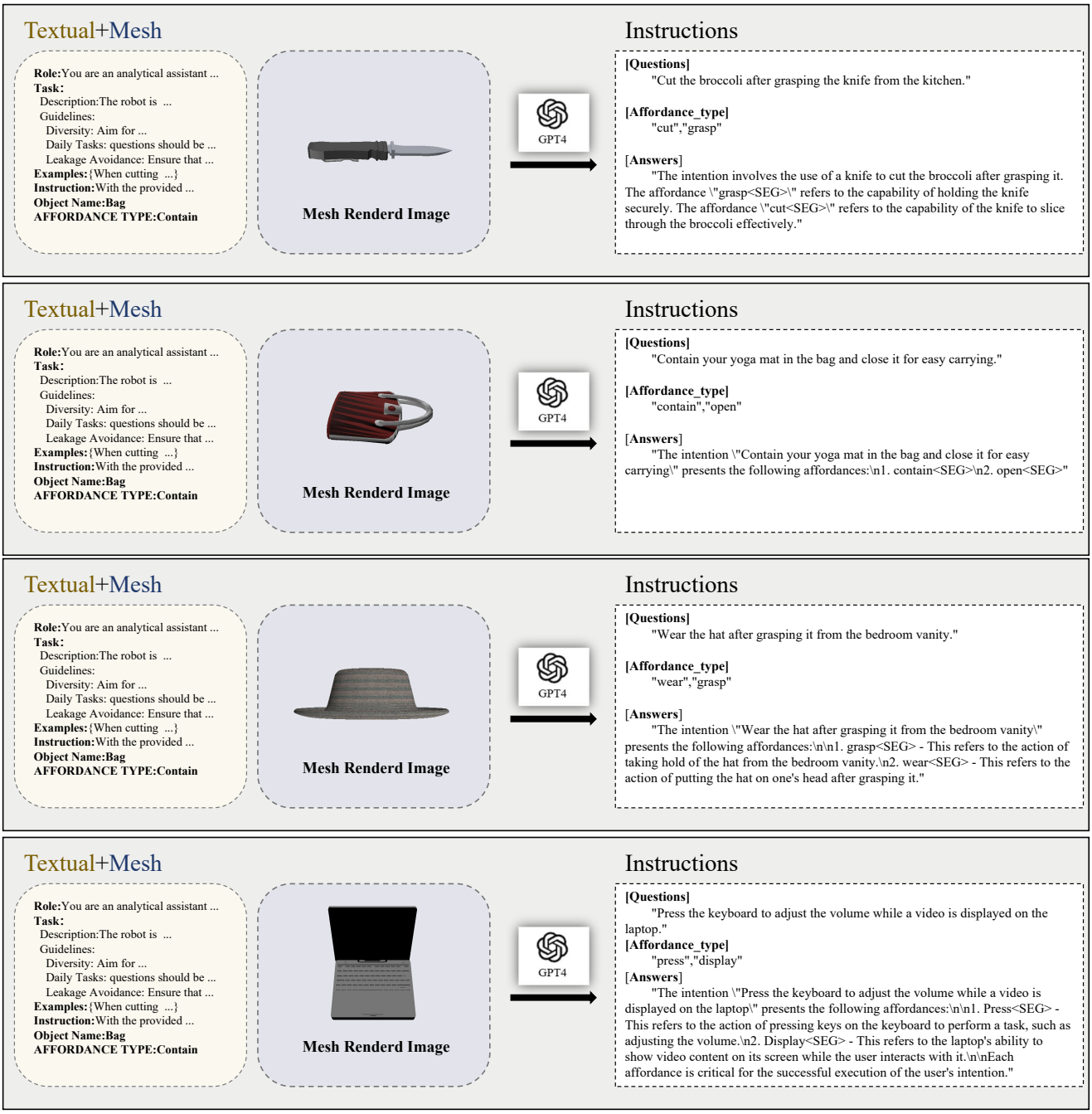


Figure A3. Genereted Instructions Example:Textual Prompt with Mesh Images

Next, we follow DGCNN[**?**] using its upsample technique to obtain $f1^{'}$ and $f2^{'}$:

$$f1^{'} = Up(P2, f1, P2.H8), \tag{5}$$

$$f2^{'} = Up(P3, f2, P2, f1^{'}) \tag{6}$$

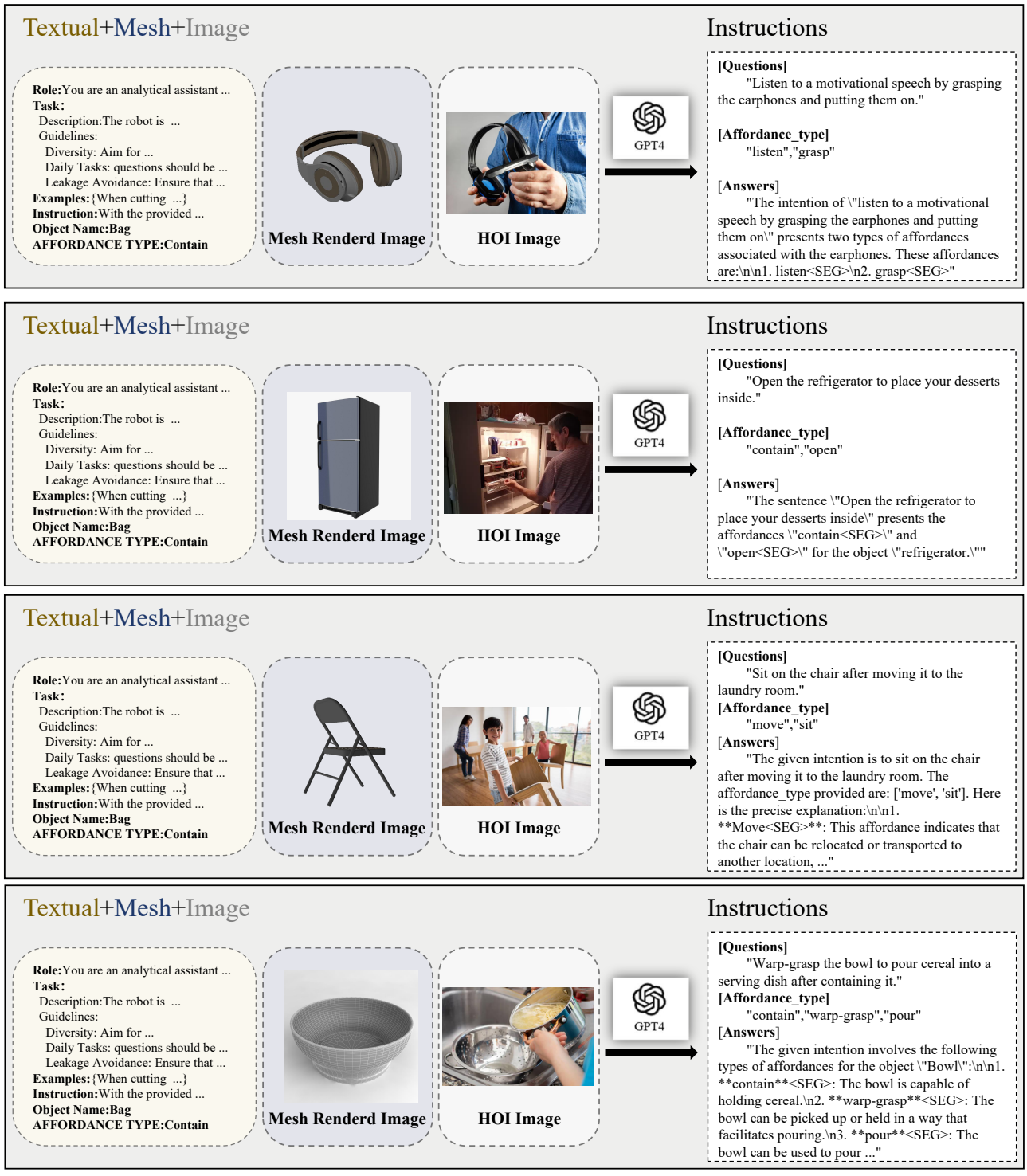


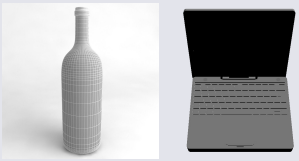
Figure A4. Genereted Instructions Example:Textual Prompt with Mesh and HOI Images

E.2. Mutil-Object

Here, we describe the detailed implementation of the Mutil-Object affordance reasoning. During training, we add $< SEG >$ token after the object name, and our MLLM will learn the paragrim and when it needs to generate text output, it will also

Textual+Mesh+Scenarios

Role: You are an analytical assistant ...
Task:
Description:The robot is ...
Guidelines:
Diversity: Aim for ...
Daily Tasks: questions should be ...
Leakage Avoidance: Ensure that ...
Examples:{When cutting ...}
Instruction: With the provided ...
Object Name:Bag
AFFORDANCE TYPE:Contain



Mesh Renderd Images

A person is having a video conference on their laptop, with a bottle of juice next to it, taking occasional sips during the break in the conversation.

GPT4

Instructions

[Questions]
"When grasping the bottle next to the laptop, where should the person's grip focus to ensure a secure hold without disturbing the laptop?",
"How should a person's gripper be positioned to wrap around the bottle's body without knocking over the laptop?",
"Which part of the bottle offers the best grip for a person to lift it without touching the laptop's keyboard or screen?",
"How can a person adjust its grip on the bottle to accommodate different bottle sizes while keeping the laptop stable?"
[Affordance_type] "wrapgrasp"

Textual+Mesh+Scenarios

Role: You are an analytical assistant ...
Task:
Description:The robot is ...
Guidelines:
Diversity: Aim for ...
Daily Tasks: questions should be ...
Leakage Avoidance: Ensure that ...
Examples:{When cutting ...}
Instruction: With the provided ...
Object Name:Bag
AFFORDANCE TYPE:Contain

Mesh Renderd Images

A chef is in the kitchen, preparing a meal. They have a knife in one hand and are using it to chop vegetables on a cutting board placed on the table.

GPT4

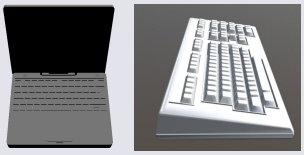
Instructions

[Questions]
"When a person is cutting vegetables, which part of the knife should it use to make precise slices?",
"Which part of the knife is ideal for a person to use when chopping herbs finely?",
"How can a person use the knife to cut a pineapple into manageable chunks?",
"How should a person position the knife to cut a carrot into thin, even sticks?",
"Which part of the knife should a robot use to core an apple without wasting too much of the fruit?"

[Affordance_type] "cut"

Textual+Mesh+Scenarios

Role: You are an analytical assistant ...
Task:
Description:The robot is ...
Guidelines:
Diversity: Aim for ...
Daily Tasks: questions should be ...
Leakage Avoidance: Ensure that ...
Examples:{When cutting ...}
Instruction: With the provided ...
Object Name:Bag
AFFORDANCE TYPE:Contain



Mesh Renderd Images

A software developer is at their desk, rapidly typing code on a backlit keyboard attached to a high-performance laptop.

GPT4

Instructions

[Questions]
"Where should you press to take a screenshot on your computer?",
"When you want to duplicate a selected file on a Mac, which keys should you press?",
"When you want to print a document, where should you press?",
"To open the settings app on a Windows computer, where should you press?",
"To open a new document in many software programs, where should you press?"

[Affordance_type] "press"

Textual+Mesh+Scenarios

Role: You are an analytical assistant ...
Task:
Description:The robot is ...
Guidelines:
Diversity: Aim for ...
Daily Tasks: questions should be ...
Leakage Avoidance: Ensure that ...
Examples:{When cutting ...}
Instruction: With the provided ...
Object Name:Bag
AFFORDANCE TYPE:Contain

Mesh Renderd Images

A person is loading dirty dishes into a dishwasher, carefully placing a bowl into the upper rack.

GPT4

Instructions

[Questions]
"If you need to wash utensils with silicone handles, where should you place them in the dishwasher?",
"If you need to wash small bowls, where should you place them in the dishwasher?",
"Where should you put mixing bowls to ensure they are properly contained during the wash?",
"Where should you place bowls to prevent them from moving during the cycle?",
"If you need to wash food storage containers with compartments, where should you place them in the dishwasher?"

[Affordance_type] "contain"

Figure A5. Genereted Instructions Example:Textual Prompt with Mesh Images and Scenarios Descriptions

[Answer List Template]

ANSWER_LIST = ["{Affordance_type} are {seg}, separately.", "{Affordance_type} are {seg}.", "Sure, {Affordance_type} are {seg}, separately.", "Sure, {Affordance_type} are {seg}.", "the segmentation result of {Affordance_type} are {seg}.", "the segmentation result of {Affordance_type} are {seg}, separately.", "Sure, the segmentation result of {Affordance_type} are {seg}.", "Sure, the segmentation result of {Affordance_type} are {seg}, separately.",]

[Mutil Answer List Template]

MR_MULTI_ANSWER_LIST = [
"{class_name} are {seg}, separately.",
"{class_name} are {seg}.",
"Sure, {class_name} are {seg}, separately.",
"Sure, {class_name} are {seg}.",
"the segmentation result of {class_name} are {seg}.",
"the segmentation result of {class_name} are {seg}, separately.",
"Sure, the segmentation result of {class_name} are {seg}.",
"Sure, the segmentation result of {class_name} are {seg}, separately.",]

[Explanatory Question List Template]

EXPLANATORY_QUESTION_LIST = [
DEFAULT_POINT_TOKEN + "\n" + "{sent}. Please respond with segmentation mask." Please output segmentation mask and explain why.
DEFAULT_POINT_TOKEN + "\n" + "{sent}. Please output segmentation mask", Please output segmentation mask and explain the reason.",]
DEFAULT_POINT_TOKEN + "\n" + "{sent} .Can you give me the segmentation mask?" Please output segmentation mask and give some explanation.]

[Question List Template]

LONG_QUESTION_LIST = [
DEFAULT_POINT_TOKEN + "\n" + "{sent}. Please respond with segmentation mask.",
DEFAULT_POINT_TOKEN + "\n" + "{sent}. Please output segmentation mask.",]
DEFAULT_POINT_TOKEN + "\n" + "{sent} .Can you give me the segmentation mask?",]

Figure A6. The details of our templates

follow this paragtim. We extract the nouns from the output of the language model, we then use the order of these nouns as the order of the model affordance reasoning. This is justified because the large language model is able to output the correct text sequence.

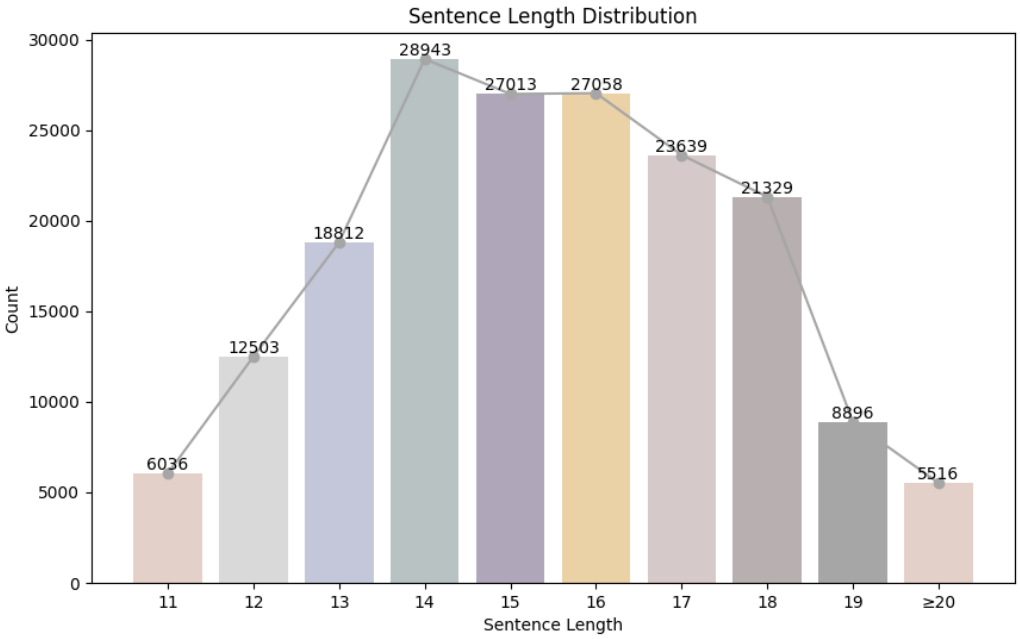
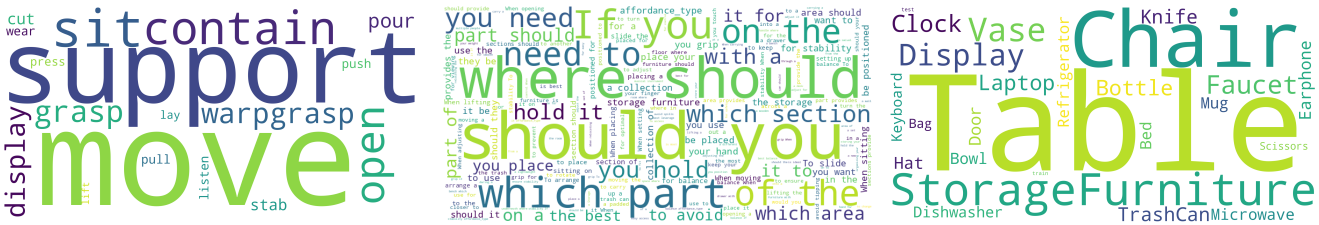


Figure A7. Question Length Distribution



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

Figure A8. Word clouds of (a) instructions, (b) affordance types, and (c) object categories.

F. More Visualization Results

Here, we show more qualitative results of our model. Fig. A10 shows the additional visualization results of Single Affordance Reasoging, and Fig. A11 shows the additional visualization results of

G. Comparison with the Previous SOTA Method

We show the qualitative comparison with LASO[**?**]: obviously, when the question is complex, LASO[**?**] failed to understand the intention and do not perform well in connecting the affordance with the intention.

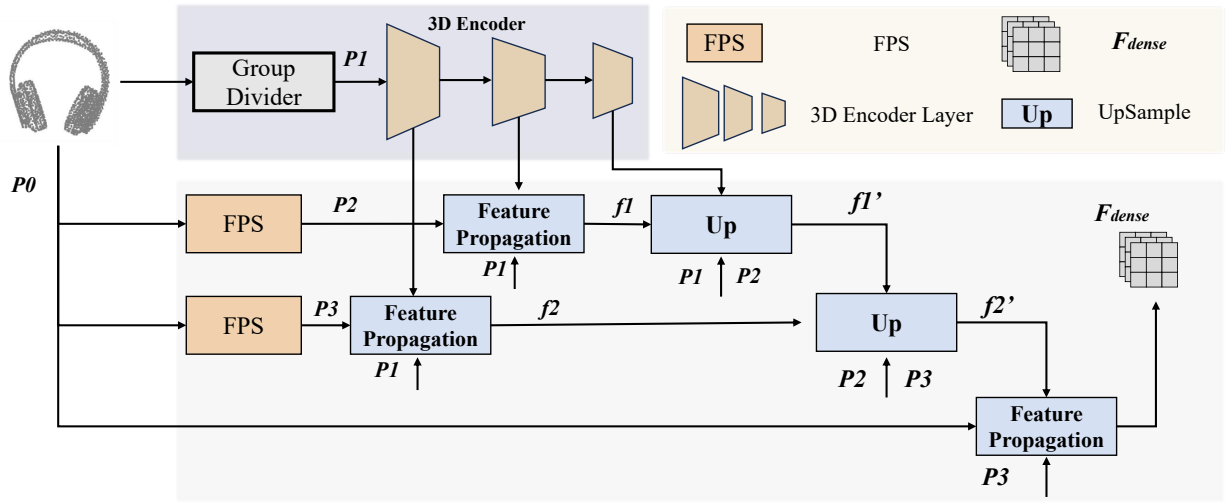


Figure A9. **Upsample.** We show the details of the Upsample method.Specifically, we use FPS and feature propagation to get the hierarchical point cloud features and upsample the sparse features into dense features

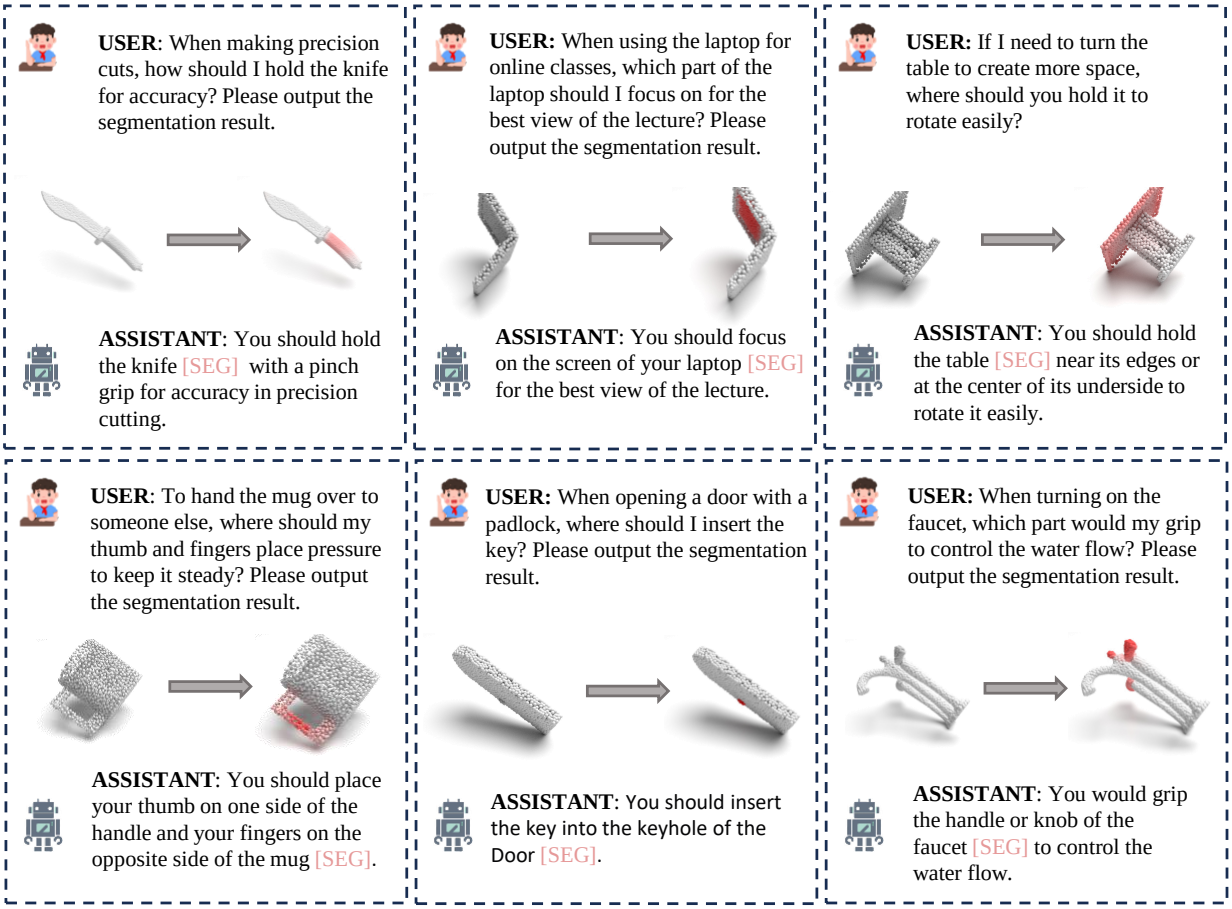
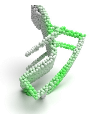


Figure A10. More Qualitative Visualization Results of Sequential Affordance Reasoning

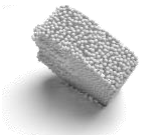
USER: Scenario: Rearrange the seating area by relocating the chair from the study room to the family room and then comfortably sit on it to enjoy a favorite book? In this scenario, what should I do?

ASSISTANT: First you should move the chair [SEG].

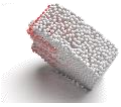


ASSISTANT: Then sit on the chair [SEG] comfortably.

USER: How can you ensure that your folded clothes are neatly placed inside a bag and the drawstring is tightened for a secure closure?

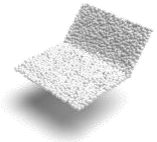


ASSISTANT: Firstly, open the bag [SEG] to access its interior for placing the folded clothes.

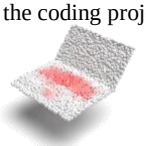


ASSISTANT: Secondly, makes use of the bag [SEG] to contain the clothes.

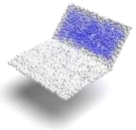
USER: How can you initiate a program's execution by pressing the keyboard and subsequently view its output on the laptop screen?



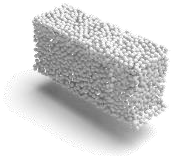
ASSISTANT: First, you press the keyboard of the laptop [SEG] to execute the coding project.



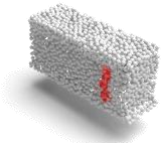
ASSISTANT: Second, the laptop [SEG] screen displays the output of the program.



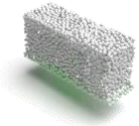
USER: How should I meticulously prepare to gently warm a bowl of soup in a microwave oven during a rainy day?



ASSISTANT: First you should open the microwave [SEG] gently.

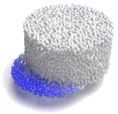


ASSISTANT: Then make sure that the microwave [SEG] has enough space to contain it.



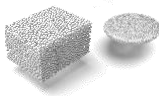
USER: How should you properly store your hat to keep it safe and in good condition when you're not wearing it?

ASSISTANT: First you should carefully grasp the hat [SEG] from the your head.

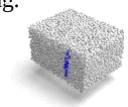


ASSISTANT: Then put your hat on the table [SEG] to support it.

USER: How should I properly reheat the food in the bowl?



ASSISTANT: First open the microwave [SEG] to prepare for loading.



ASSISTANT: Then you should carefully grasp the bowl [SEG] and make sure the food does not fall out.

ASSISTANT: Finally, use the microwave [SEG] to contain the bowl and start heating.

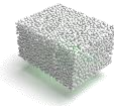


Figure A11. More Qualitative Visualization Results of Sequential Affordance Reasoning

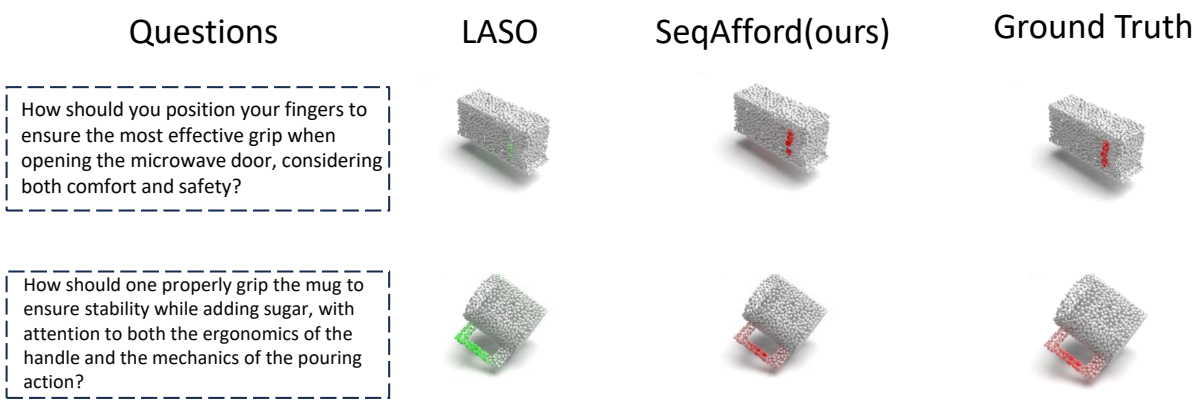


Figure A12. Comparison with the Previous SOTA Method