

# ShapeLLM: Universal 3D Object Understanding for Embodied Interaction

---

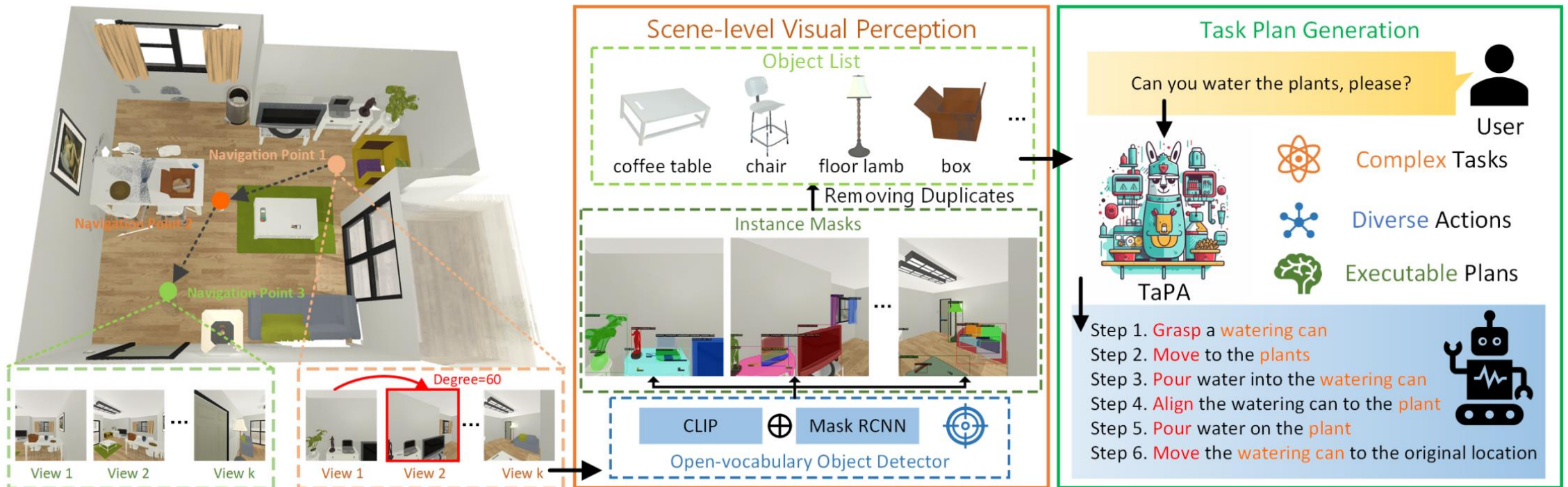
**Zekun Qi**

**28 September 2024**

# Background

①: LLMs have rich knowledge, especially **planning and decision-making** abilities.

Which can be used for Embodied Intelligence.



## Background

②: By supervised fine-tuning with instruction-following data, LLM's knowledge can be easily **aligned with the specific field**.

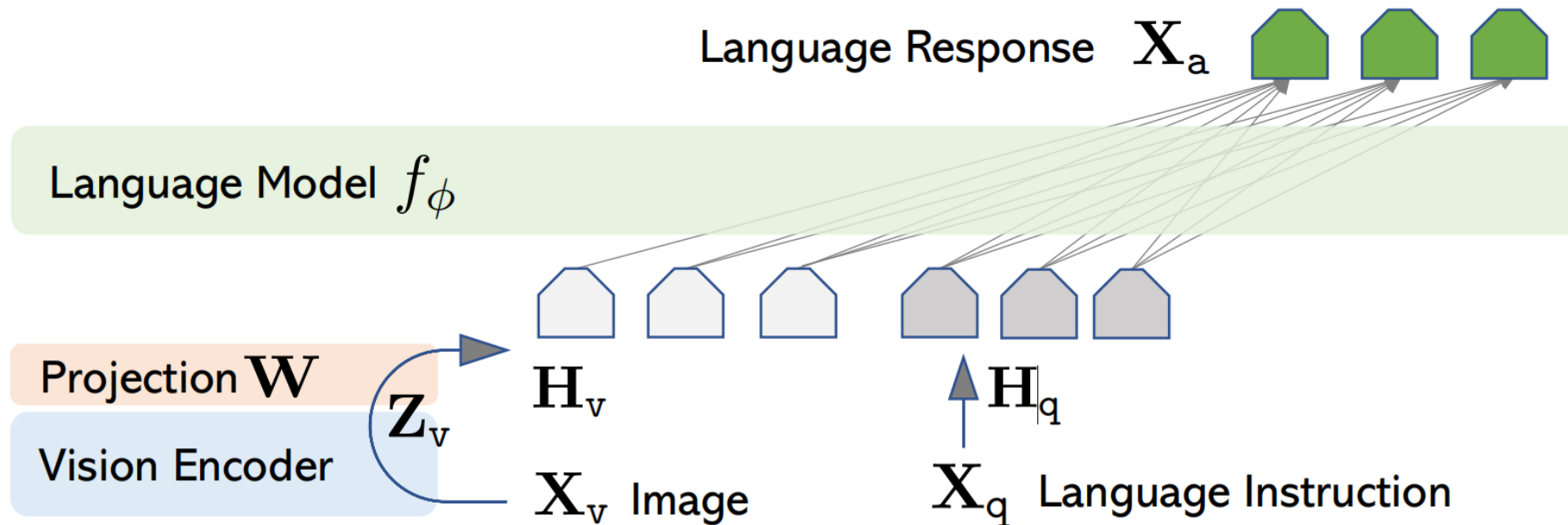
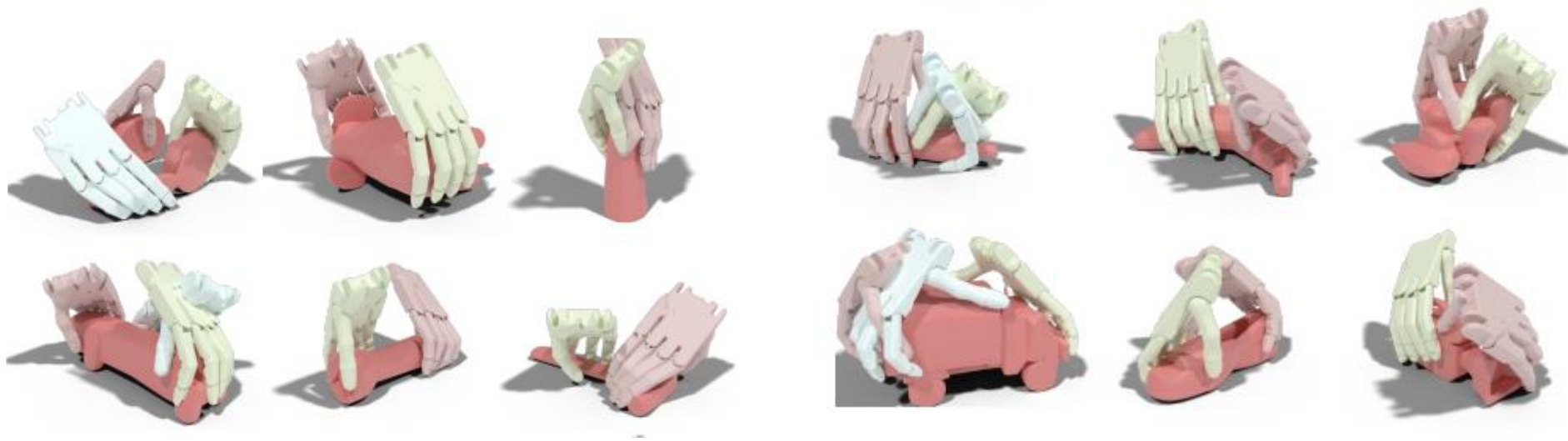


Figure 1: LLaVA network architecture.

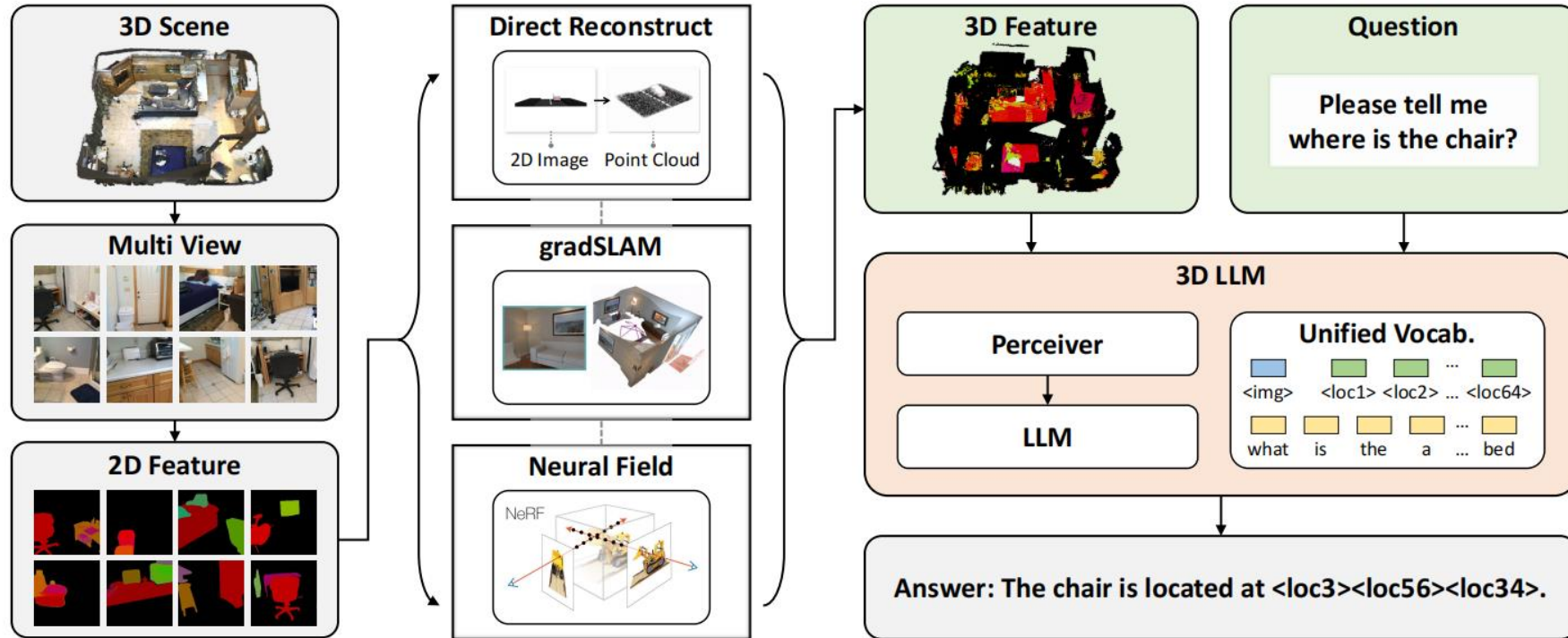
## Background

③: **3D Point Clouds as Inputs.** Compared to 2D images, 3D point clouds provide a **more accurate representation** of the physical environment, which is necessary for embodied tasks like 6-DoF grasping.



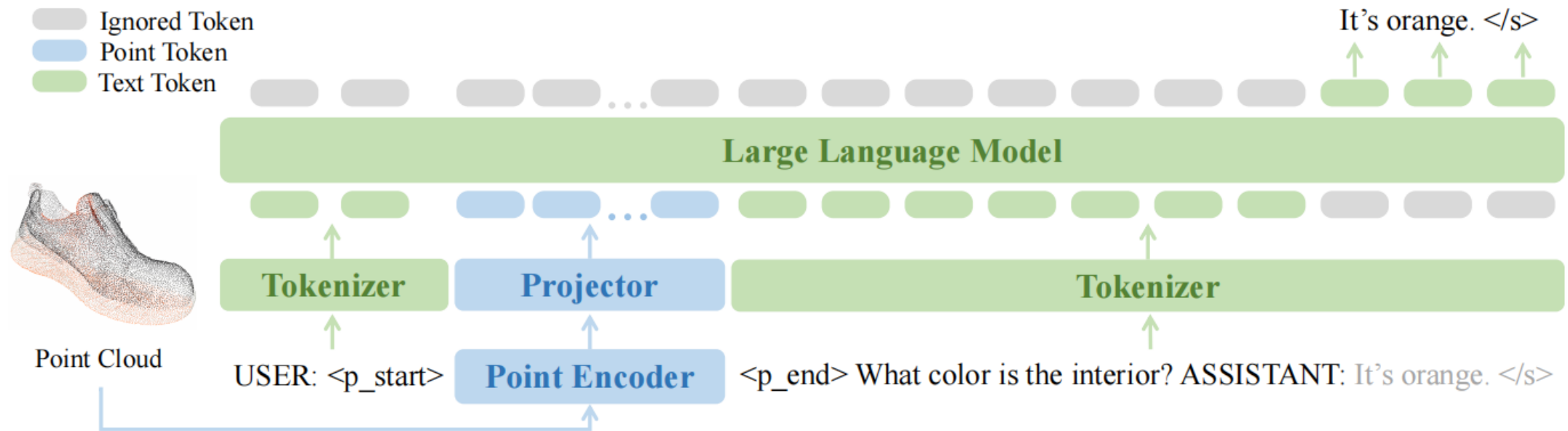
# Background

④: Embodied manipulation is typically **object-centric**, operating a **part of the object**, while current 3D LLMs mainly focus on the **scene level**, and full scene point clouds are needed



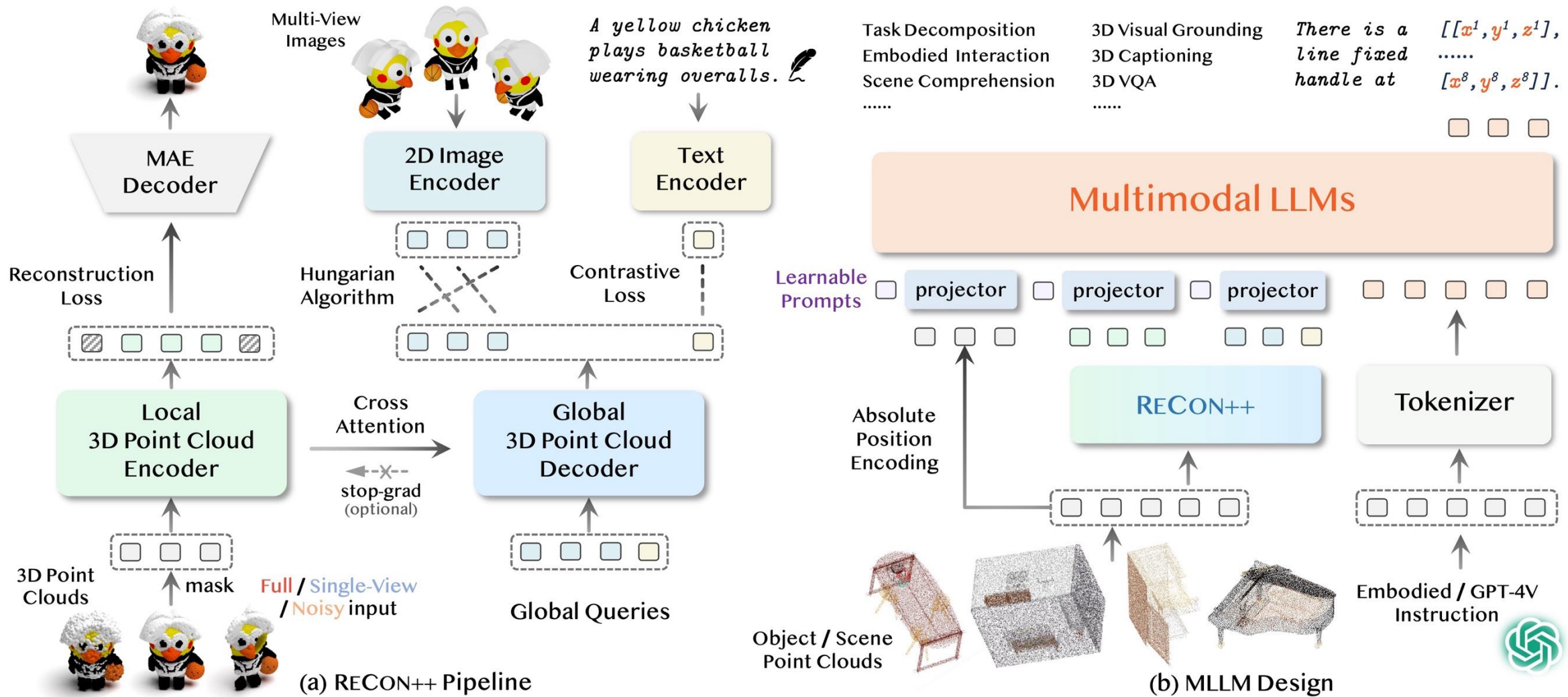
## Background

⑤: Most robots are only equipped with a **single-view RGBD camera**, which poses challenges in acquiring full point clouds.





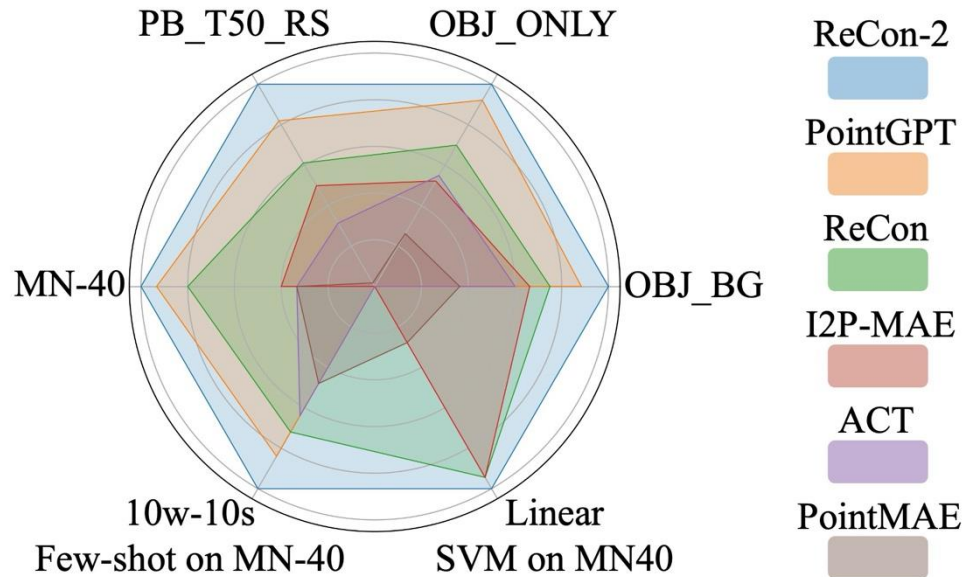
# Pipeline



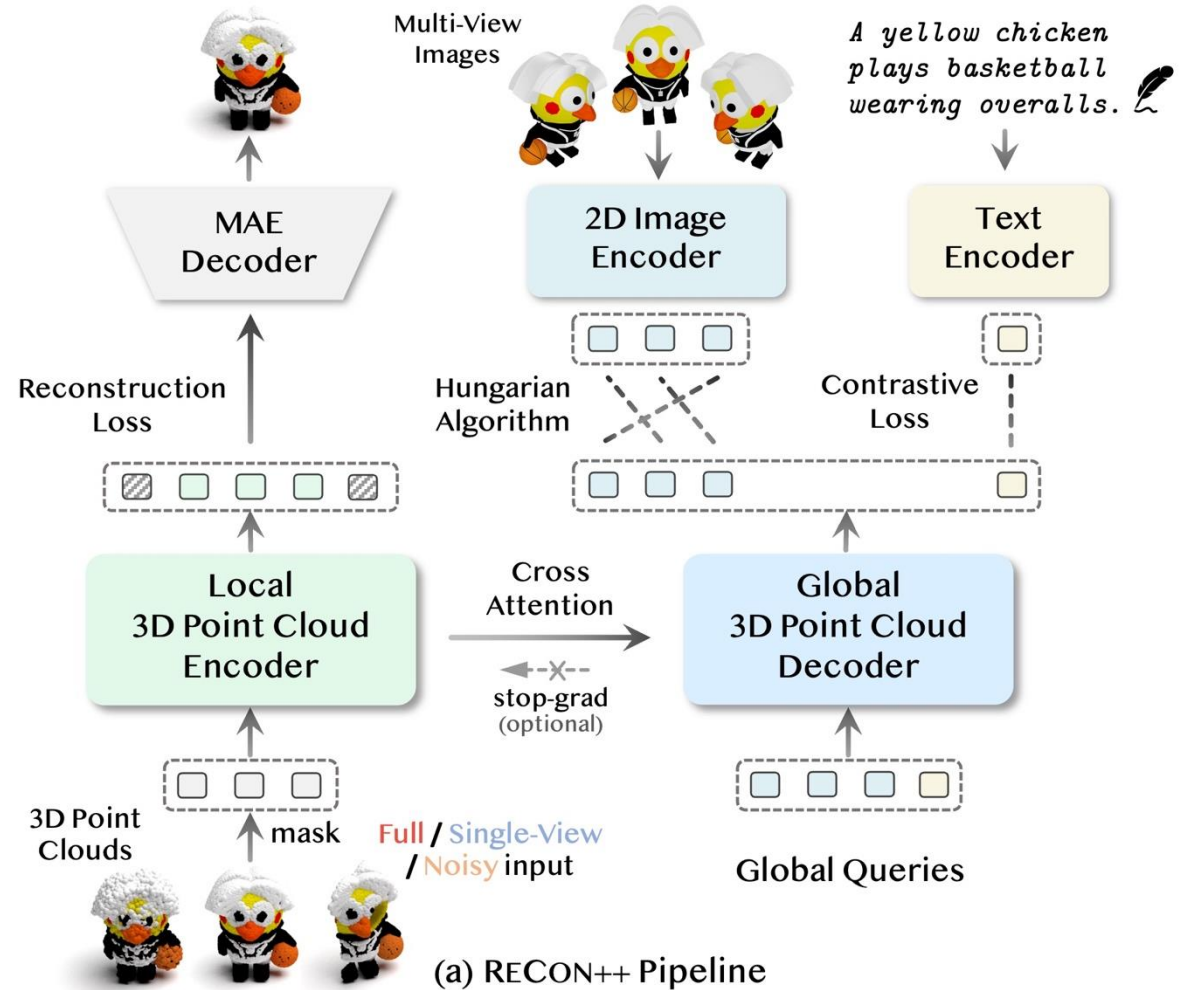
# Pipeline

## Point Encoder Encoder

1. **Scaling up** to Objaverse
2. Distillation on multi-view images
3. Training with **single-view & noisy points**



## 3D Representation Learning





## Transfer Learning

Table 1. **Fine-tuned 3D recognition** on ScanObjectNN and ModelNet40. BG, ON, and RS are short for OBJ\_BG, OBJ\_ONLY, and PB\_T50\_RS, respectively. Overall accuracy (%) with voting [98] is reported. †: results with a post-pretraining stage [18].

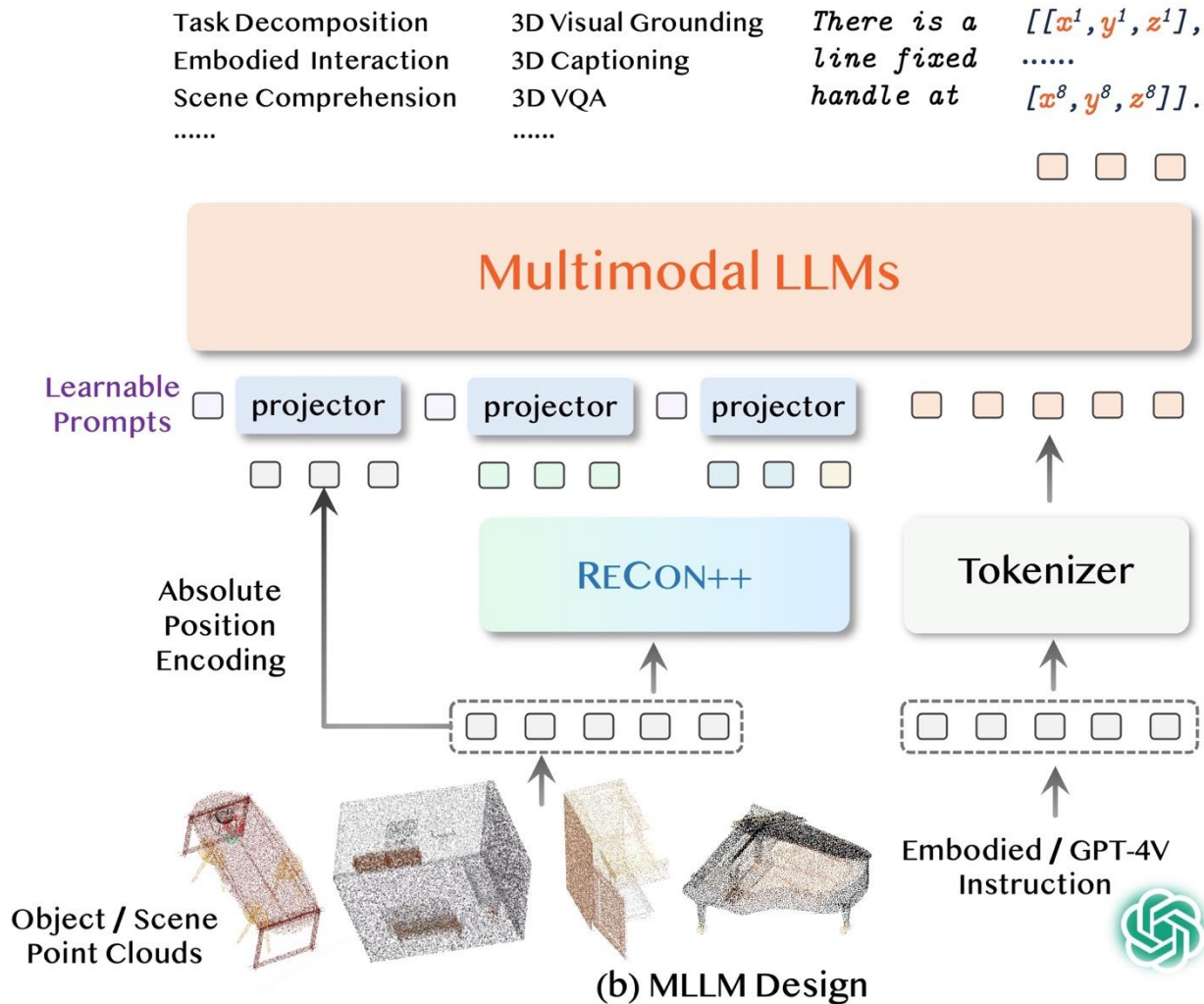
Method	ScanObjectNN			ModelNet40	
	BG	ON	RS	1k P	8k P
<i>Supervised Learning Only</i>					
PointNet [127]	73.3	79.2	68.0	89.2	90.8
PointNet++ [128]	82.3	84.3	77.9	90.7	91.9
DGCNN [158]	82.8	86.2	78.1	92.9	-
PointMLP [108]	-	-	85.4	94.5	-
PointNeXt [132]	-	-	87.7	94.0	-
Transformer [154]	83.04	84.06	79.11	91.4	91.8
<i>with Self-Supervised Representation Learning</i>					
Point-BERT [181]	87.43	88.12	83.07	93.2	93.8
Point-MAE [120]	90.02	88.29	85.18	93.8	94.0
Point-M2AE [186]	91.22	88.81	86.43	94.0	-
ACT [36]	93.29	91.91	88.21	93.7	94.0
TAP [159]	-	-	88.5	94.0	-
VPP [131]	93.11	91.91	89.28	94.1	94.3
I2P-MAE [189]	94.15	91.57	90.11	94.1	-
ULIP-2 [171]	-	-	91.5	-	-
RECON [130]	95.35	93.80	91.26	94.5	94.7
PointGPT-B <sup>†</sup> [18]	95.8	95.2	91.9	94.4	94.6
PointGPT-L <sup>†</sup> [18]	97.2	96.6	93.4	94.7	94.9
<b>RECON++-B<sup>†</sup></b>	<b>98.62</b>	<b>96.21</b>	<b>93.34</b>	<b>94.6</b>	<b>94.8</b>
<b>RECON++-L<sup>†</sup></b>	<b>98.80</b>	<b>97.59</b>	<b>95.25</b>	<b>94.8</b>	<b>95.0</b>

## Zero-shot 3D Shape Recognition

Table 2. **Zero-shot 3D recognition** on Objaverse-LVIS [29], ModelNet40 [165] and ScanObjectNN [152]. Ensembled: trained on Objaverse [29], ShapeNet [13], ABO [22] and 3D-FUTURE [42], following OpenShape [94]. †: Uni3D employs a larger EVA-CLIP-E [147] as the cross-modal teacher [36] that improves performance, while other methods employ OpenCLIP-bigG [74].

Method	Objaverse-LVIS			ModelNet40			ScanObjectNN		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
<i>2D Inference</i>									
PointCLIP [187]	1.9	4.1	5.8	19.3	28.6	34.8	10.5	20.8	30.6
PointCLIP2 [199]	4.7	9.5	12.9	63.6	77.9	85.0	42.2	63.3	74.5
<i>Trained on ShapeNet</i>									
RECON [130]	1.1	2.7	3.7	61.2	73.9	78.1	42.3	62.5	75.6
CLIP2Point [69]	2.7	5.8	7.9	49.5	71.3	81.2	25.5	44.6	59.4
ULIP [170]	6.2	13.6	17.9	60.4	79.0	84.4	51.5	71.1	80.2
OpenShape [94]	10.8	20.2	25.0	70.3	86.9	91.3	47.2	72.4	84.7
<i>Trained on Ensembled</i>									
ULIP-2 [171]	26.8	44.8	52.6	75.1	88.1	93.2	51.6	72.5	82.3
OpenShape [94]	46.8	69.1	77.0	84.4	96.5	98.0	52.2	79.7	88.7
Uni3D-B <sup>†</sup> [196]	51.7	74.1	80.8	86.3	<b>96.5</b>	<b>97.9</b>	63.8	<b>82.7</b>	90.2
Uni3D-L <sup>†</sup> [196]	53.1	75.0	81.5	86.3	<b>96.8</b>	<b>98.3</b>	58.2	81.8	89.4
<b>RECON++-B</b>	<b>53.2</b>	<b>75.3</b>	<b>81.5</b>	<b>86.5</b>	94.7	95.8	<b>63.6</b>	80.2	<b>90.6</b>
<b>RECON++-L</b>	<b>53.7</b>	<b>75.8</b>	<b>82.0</b>	<b>87.3</b>	95.4	96.1	<b>65.4</b>	<b>84.1</b>	<b>89.7</b>

# Pipeline



## Feature Bridge

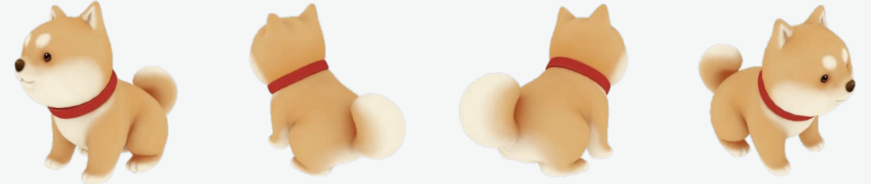
- Multi-level feature projection
  - Global contrastive feature
  - Local geometry reconstruction
  - Absolute position encoding
- Visual prompt tuning

$E_{APE}$	$E_{prompt}$	$E_{local}$	$E_{global}$	3D MM-Vet	GAPartNet
✓	✗	✗	✗	30.8	<b>12.3</b>
✓	✓	✗	✗	32.0	11.4
✗	✗	✓	✗	42.2	10.0
✗	✓	✗	✓	50.3	10.5
✓	✗	✓	✓	52.3	10.5
✓	✓	✗	✓	50.3	11.7
✗	✗	✗	✓	52.4	11.7
✗	✗	✓	✓	49.6	10.1
✗	✓	✓	✓	51.7	10.1
✓	✓	✓	✓	<b>53.1</b>	11.7

## Instruction-Following Data

1. General instruction data from Objaverse-LVIS based on GPT-4V
2. Embodied interaction data from GPartNet bounding box based on GPT-4

### [Image Upload]



### [System Prompt]

You are a 3D Computer visual assistant. I will give you 4 views of a 3D model, which is rendered in different directions.

Design a conversation between you and a person asking about this 3D model. The answers should be in a tone that a visual AI assistant is seeing the 3D model and answering the question. Ask diverse questions and give corresponding answers. You need to consider the spatial relation in different views.

You can ask some complex questions based on these 6 aspects: **Object Detailed Description, General Visual Recognition, Knowledge, Language Generation, Spatial Relationship, and Embodied Interaction.**

I will give you several examples, and please generate questions and answer pairs with the format of my examples. For example:

Q: <Example Q1> A: <Example A1> Q: <Example Q2> A: <Example A2>

Provide detailed answers when answering complex questions. You can give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary. Do not ask about uncertain details.

Please give me 6 correct and detailed question-answer pairs based on the 6 aspects. You should try to avoid asking "Q: Is there xxx? A: Yes, xxx."

# ShapeLLM

point cloud input



original mesh  
(only for reference)

User: What kind of animal is this like? And what color is it?

User



The figure resembles a bull or a buffalo. Its primary color is blue.

User: What kind of sports might he be doing?

User



He appears to be doing a sport that involves jumping or leaping, possibly basketball or volleyball.

Visual Question Answering

point cloud input



original mesh  
(only for reference)

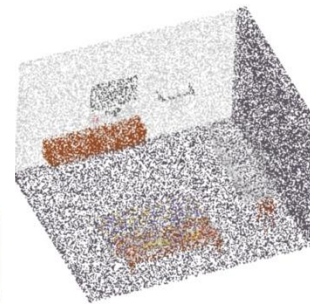
User: Generate a brief caption for this 3D model.

User



A playful 3D model of a cat, rendered in a soft pastel color palette.

3D Captioning



User: Is the sofa on the carpet?

User



Yes, the sofa is on the carpet.

User: How many doors are there beneath the sink cabinet?

User




There are two doors beneath the sink cabinet.

Scene Understanding




# ShapeLLM

point cloud input




original mesh (only for reference)



I can't find my file, where is it most likely to be in this model?

User

This storage furniture has a drawer, and files are likely to be in the drawer. You need to pull the drawer to view it, and the bounding box of the drawer is  $[[[-0.14, -0.44, 0.75], [-1.05, -0.44, 0.75]], [-1.05, 0.45, 0.75], [-0.14, 0.45, 0.75], [-0.14, -0.44, 0.42], [-1.05, -0.44, 0.42], [-1.05, 0.45, 0.42], [-0.14, 0.45, 0.42]]$ .



Embodied Visual Grounding

point cloud input



original mesh (only for reference)



What might have just happened?

User

The egg has just been cracked, revealing its contents.


As an embodied AI assistant, please give me steps to cook this object.

User


1. Crack an egg into a bowl.
2. Add salt, pepper, and any desired ingredients.
3. Scramble them together with a spatula until they are well combined.

Planning & Task Decomposition

single-view point cloud input



original point cloud (only for reference)



Please give me several steps to weigh 2 things and sort them by the weight.

User

1. Make sure the scale is set to zero.
2. Place the first item on the scale.
3. Record the weight of the first item.
4. Place the second item on the scale.
5. Record the weight of the second item.
6. Compare the weights of the 2 items.
7. Sort the items based on their weights.

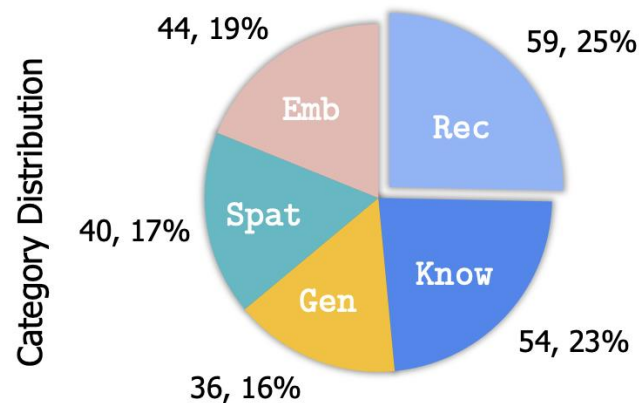
Single-View Point Cloud Understanding



## 3D MM-Vet: 3D Multimodal Comprehension Evaluation Benchmark

### Five Levels of 3D Understanding

- General Visual Recognition
- Knowledge
- Language Generation
- Spatial Recognition
- Embodied Interaction



[Point Clouds]



[Question<sup>1</sup> Type: Knowledge]

What is this woman's occupation?

[Answer<sup>1</sup>]

She is a medical worker.

[Question<sup>2</sup> Type: General Visual Recognition]

What clothes are the women wearing and what colors are they?

[Answer<sup>2</sup>]

She is wearing protective clothing and two medical gloves. the protective clothing is blue while the medical gloves are green.

[Question<sup>3</sup> Type: Embodied Interaction]

As an AI assistant, please give me some steps to put her in all the protective equipment.

[Answer<sup>3</sup>]

Step 1: Put her in the protective clothing and tie the knot.

Step 2: Put her hair up and cover it with the hat.

Step 3: Open the gloves and put them on her.

## 3D MM-Vet: 3D Multimodal Comprehension Evaluation Benchmark

### [Point Clouds]

#### [Question<sup>1</sup> Type: General Visual Recognition]

What subparts are there in the scene?

#### [Answer<sup>1</sup>]

There is a bag of cookies, a mug of milk and a China bowl.

#### [Question<sup>2</sup> Type: Embodied Interaction]

As an AI robot, please give me steps to mix the milk and cookies in a bowl.

#### [Answer<sup>2</sup>]

Step 1: Pour the milk into the bowl.

Step 2: Put the cookies into the bowl.

Step 3: Stir with a spoon.

#### [Question<sup>3</sup> Type: Knowledge]

Describe the physical properties of the milk.

#### [Answer<sup>3</sup>]

The milk is a kind of liquid with a white color, whose density and boiling point is higher than water while the freezing point is lower than water, has a mild, slightly sweet odor and taste.



## 3D MM-Vet-C: Robustness Evaluation Benchmark

**Single-View:** randomly select a camera viewpoint within the unit sphere and generate a **single viewpoint** within the FoV on polar coordinates. **Jitter:** Gaussian jittering with noise  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  and  $\sigma = 0.01$ . **Rotate:** random SO(3) rotation sampling over X-Y-Z Euler angle  $(\alpha, \beta, \gamma) \sim \mathcal{U}(-\theta, \theta)$  and  $\theta = \pi/6$ .

Method	3D MM-Vet-C Variants			
	Clean	Single-View	Jitter	Rotate
PointBind&LLM [54]	23.5	20.4	19.7	19.5
PointLLM-7B [167]	41.2	33.6	38.8	40.6
PointLLM-13B [167]	46.6	41.3	42.3	44.2
<b>SHAPELLM-7B</b>	<b>47.4</b>	<b>38.3</b>	<b>45.8</b>	<b>42.7</b>
<b>SHAPELLM-13B</b>	<b>53.1</b>	<b>43.6</b>	<b>47.8</b>	<b>49.3</b>

single-view point cloud



single-view point cloud

