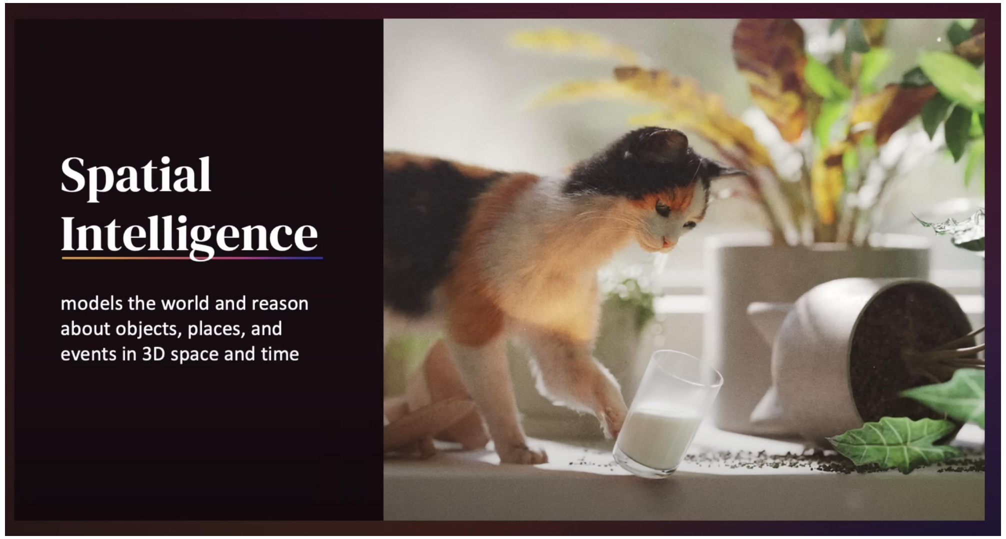
SoFar: Language-Grounded Orientation Bridges Spatial Reasoning and Object Manipulation

Background: From Seeing to Doing with Spatial Intelligence

Question *How to proceed from seeing to doing?*

Seeing is
for doing



Spatial
Intelligence

models the world and reason
about objects, places, and
events in 3D space and time

Quote:

"Sight turning into ***insight****, seeing becomes* ***understanding****, understanding leads to* ***actions****."*

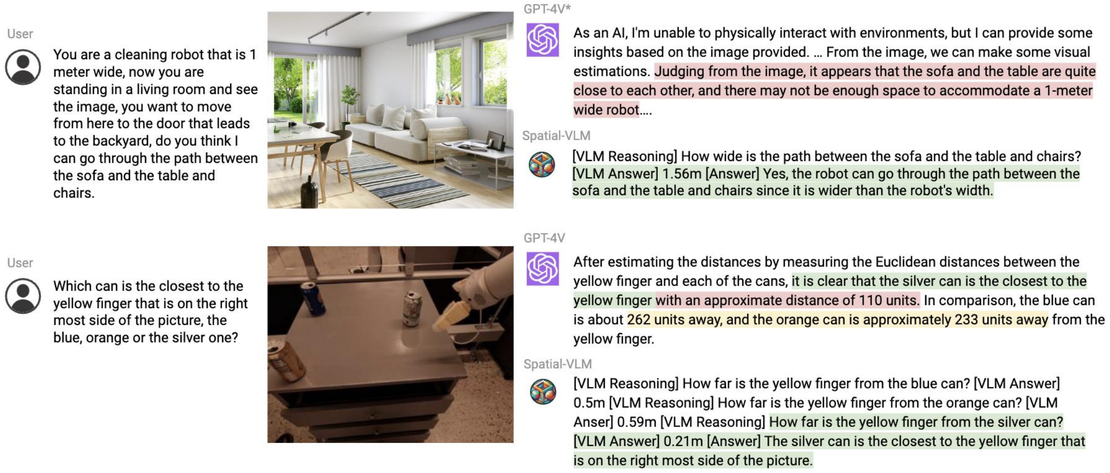
– Li Fei-Fei

Direction:

All of this leads to Spatial Intelligence.

Background: Previous Works

Question *How to proceed from seeing to doing?*



You are a cleaning robot that is 1 meter wide, now you are standing in a living room and see the image, you want to move from here to the door that leads to the backyard, do you think I can go through the path between the sofa and the table and chairs.

As an AI, I'm unable to physically interact with environments, but I can provide some insights based on the image provided. ... From the image, we can make some visual estimations. Judging from the image, it appears that the sofa and the table are quite close to each other, and there may not be enough space to accommodate a 1-meter wide robot....

[VLM Reasoning] How wide is the path between the sofa and the table and chairs? [VLM Answer] 1.56m [Answer] Yes, the robot can go through the path between the sofa and the table and chairs since it is wider than the robot's width.

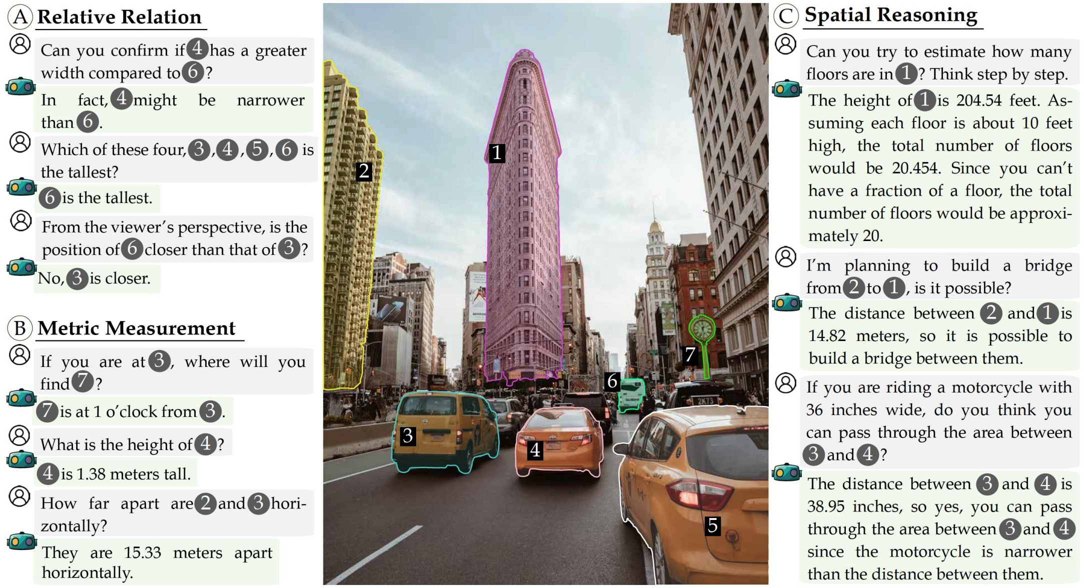
Which can is the closest to the yellow finger that is on the right most side of the picture, the blue, orange or the silver one?

After estimating the distances by measuring the Euclidean distances between the yellow finger and each of the cans, it is clear that the silver can is the closest to the yellow finger with an approximate distance of 110 units. In comparison, the blue can is about 262 units away, and the orange can is approximately 233 units away from the yellow finger.

[VLM Reasoning] How far is the yellow finger from the blue can? [VLM Answer] 0.5m [VLM Reasoning] How far is the yellow finger from the orange can? [VLM Anser] 0.59m [VLM Reasoning] How far is the yellow finger from the silver can? [VLM Answer] 0.21m [Answer] The silver can is the closest to the yellow finger that is on the right most side of the picture.

SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities

Boyuan Chen*,†,1, Zhuo Xu*,1, Sean Kirmani[1], Danny Driess[1], Pete Florence[1]
Brian Ichter[1], Dorsa Sadigh[1], Leonidas Guibas[2], Fei Xia[1]
[1]Google DeepMind, [2]Google Research
Correspond to: boyuanc@mit.edu, zhuoxu@google.com, xiafei@google.com
Website: https://spatial-vlm.github.io/



Ⓐ Relative Relation

Can you confirm if ④ has a greater width compared to ⑥?

In fact, ④ might be narrower than ⑥.

Which of these four, ③,④,⑤,⑥ is the tallest?

⑥ is the tallest.

From the viewer's perspective, is the position of ⑥ closer than that of ③?

No, ③ is closer.

Ⓑ Metric Measurement

If you are at ③, where will you find ⑦?

⑦ is at 1 o'clock from ③.

What is the height of ④?

④ is 1.38 meters tall.

How far apart are ② and ③ horizontally?

They are 15.33 meters apart horizontally.

Ⓒ Spatial Reasoning

Can you try to estimate how many floors are in ①? Think step by step.

The height of ① is 204.54 feet. Assuming each floor is about 10 feet high, the total number of floors would be 20.454. Since you can't have a fraction of a floor, the total number of floors would be approximately 20.

I'm planning to build a bridge from ② to ①, is it possible?

The distance between ② and ① is 14.82 meters, so it is possible to build a bridge between them.

If you are riding a motorcycle with 36 inches wide, do you think you can pass through the area between ③ and ④?

The distance between ③ and ④ is 38.95 inches, so yes, you can pass through the area between ③ and ④ since the motorcycle is narrower than the distance between them.

SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models

An-Chieh Cheng[1], Hongxu Yin[3], Yang Fu[1], Qiushan Guo[2], Ruihan Yang[1],
Jan Kautz[3], Xiaolong Wang[1,3], Sifei Liu[3]
[1]UC San Diego, [2]The University of Hong Kong, [3]NVIDIA

Challenges:

- Accurate **3D understanding**
- How to use it for **robotics**?
- Complex scenes
- **Language-grounded** manipulation

Previous works:

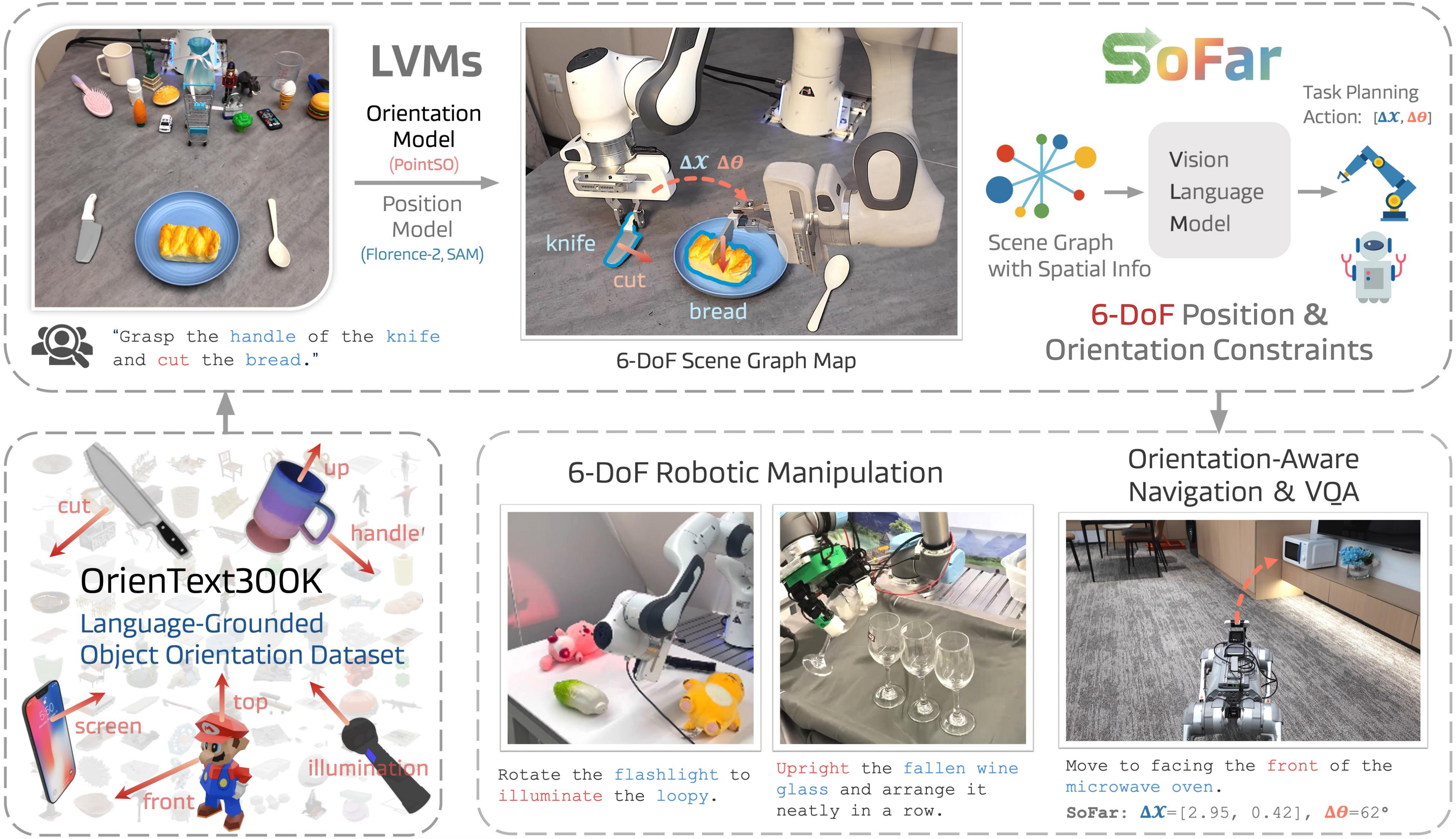
- Only **positional** understanding
- Understanding only (VQA only)

Our Insights:

- **6-DoF** spatial understanding
- Unifying **scene parsing and planning** with languages
- Open-world and

SoFar Overview

Key *Bridge Spatial Understanding & Object Manipulation*



LVMs

Orientation Model (PointSO)

Position Model (Florence-2, SAM)

"Grasp the handle of the knife and cut the bread."

6-DoF Scene Graph Map

Scene Graph with Spatial Info → Vision Language Model → Task Planning Action: $[\Delta X, \Delta\theta]$

6-DoF Position & Orientation Constraints

OrienText300K

Language-Grounded Object Orientation Dataset

6-DoF Robotic Manipulation

Rotate the flashlight to illuminate the loopy.

Upright the fallen wine glass and arrange it neatly in a row.

Orientation-Aware Navigation & VQA

Move to facing the front of the microwave oven.

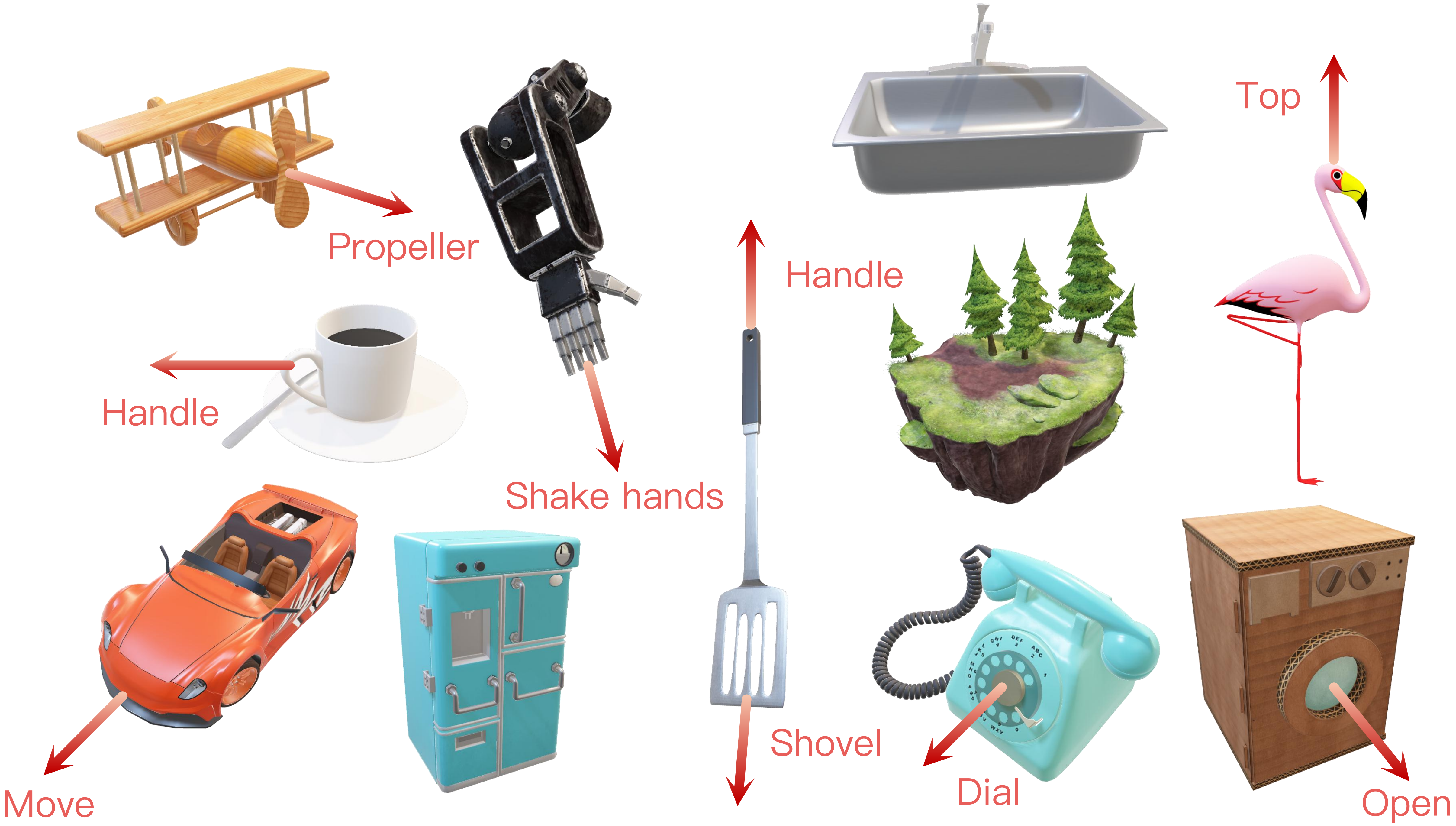
SoFar: ΔX=[2.95, 0.42], $\Delta\theta$=62°

Key Ideas:

- Semantic Orientation grounded by languages
- Modular design with large models

Highlights:

- Zero-shot
- Open-world
- Generalization
- Cross-embodiment


Propeller
Shake hands
Handle
Top
Handle
Move
Shovel
Dial
Open

cut
screen
illumination
to
front
up
handle

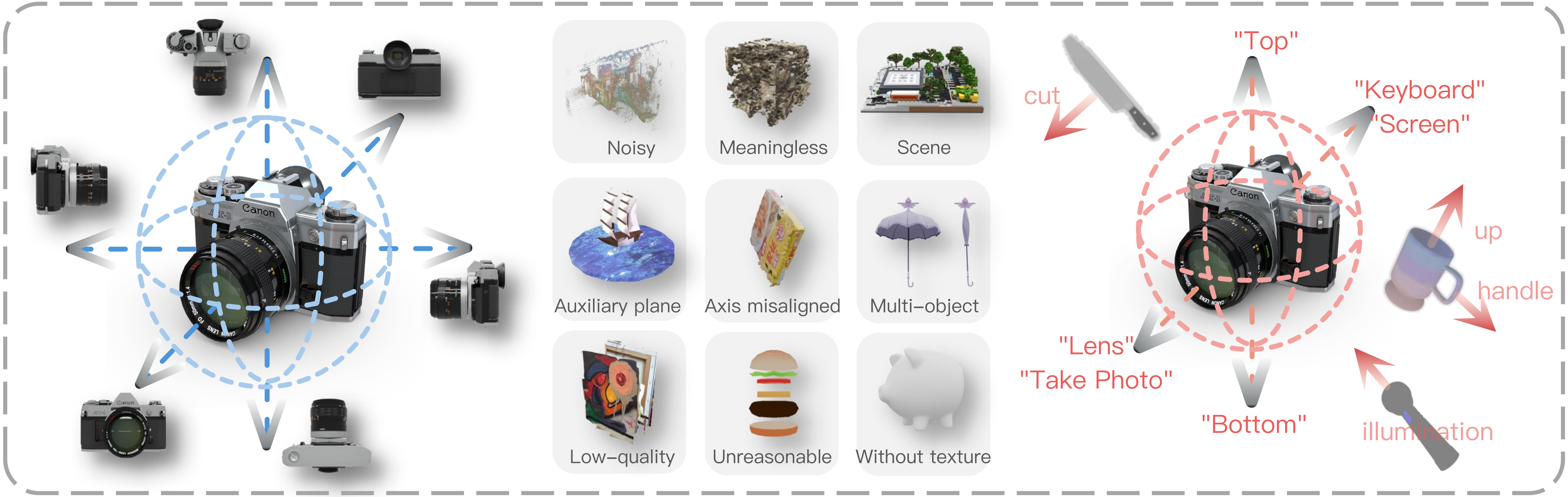
OrienText300K

Language-Grounded Object Orientation Dataset

350K+ High-quality 3D Assets & Orientations
8M Rendered Images

OrienText300K: Orientation-Text Paired Data at Scale

Data *Scalable and High-Quality Data Construction & Filtering (Objaverse)*



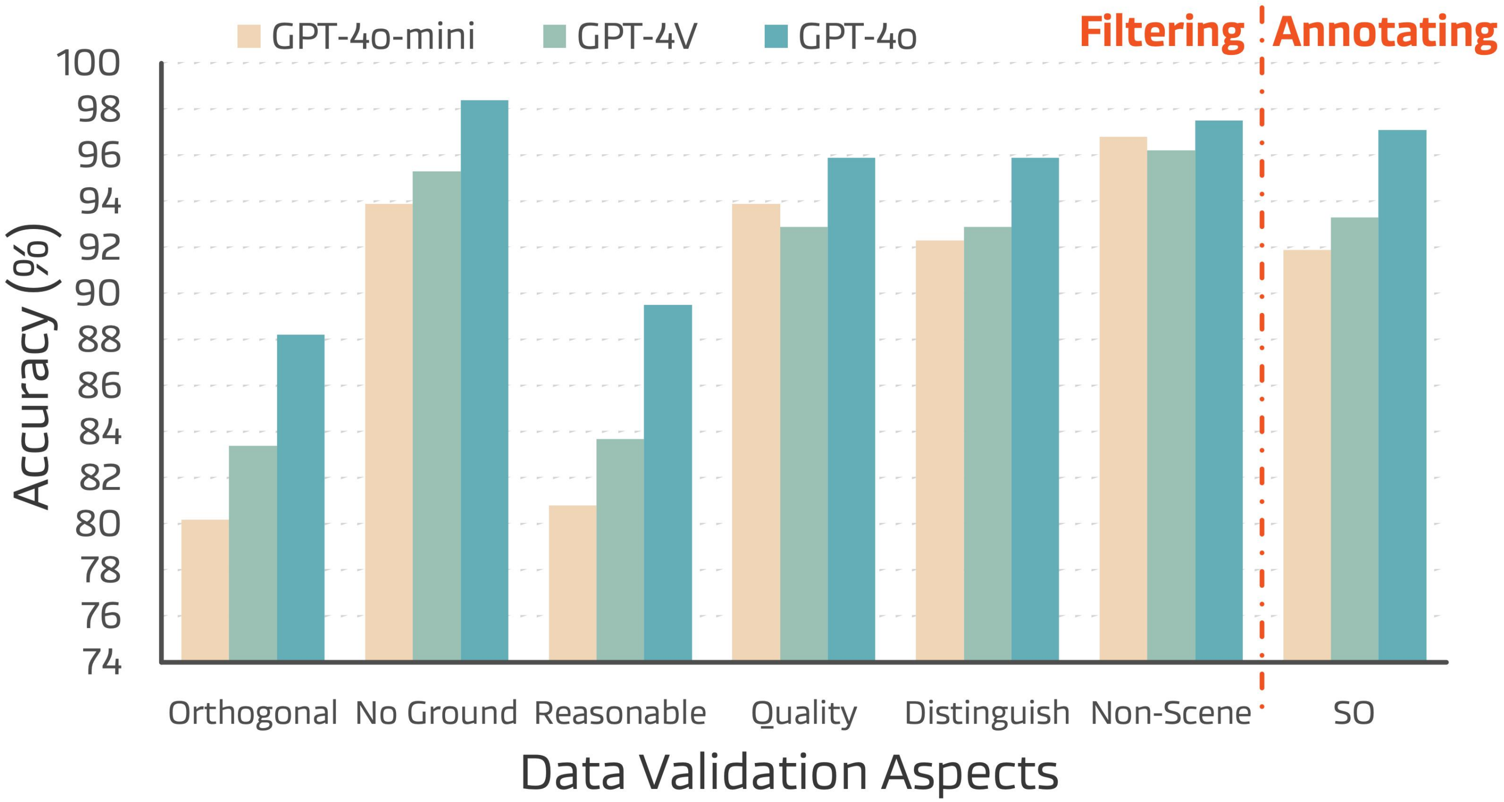
blender **Standard Views**

ChatGPT **Filtering Bad Data**

ChatGPT **Orientation–Text Pair**

OrienText300K: Orientation-Text Paired Data at Scale

Data *Data Construction & Filtering (Objaverse)*

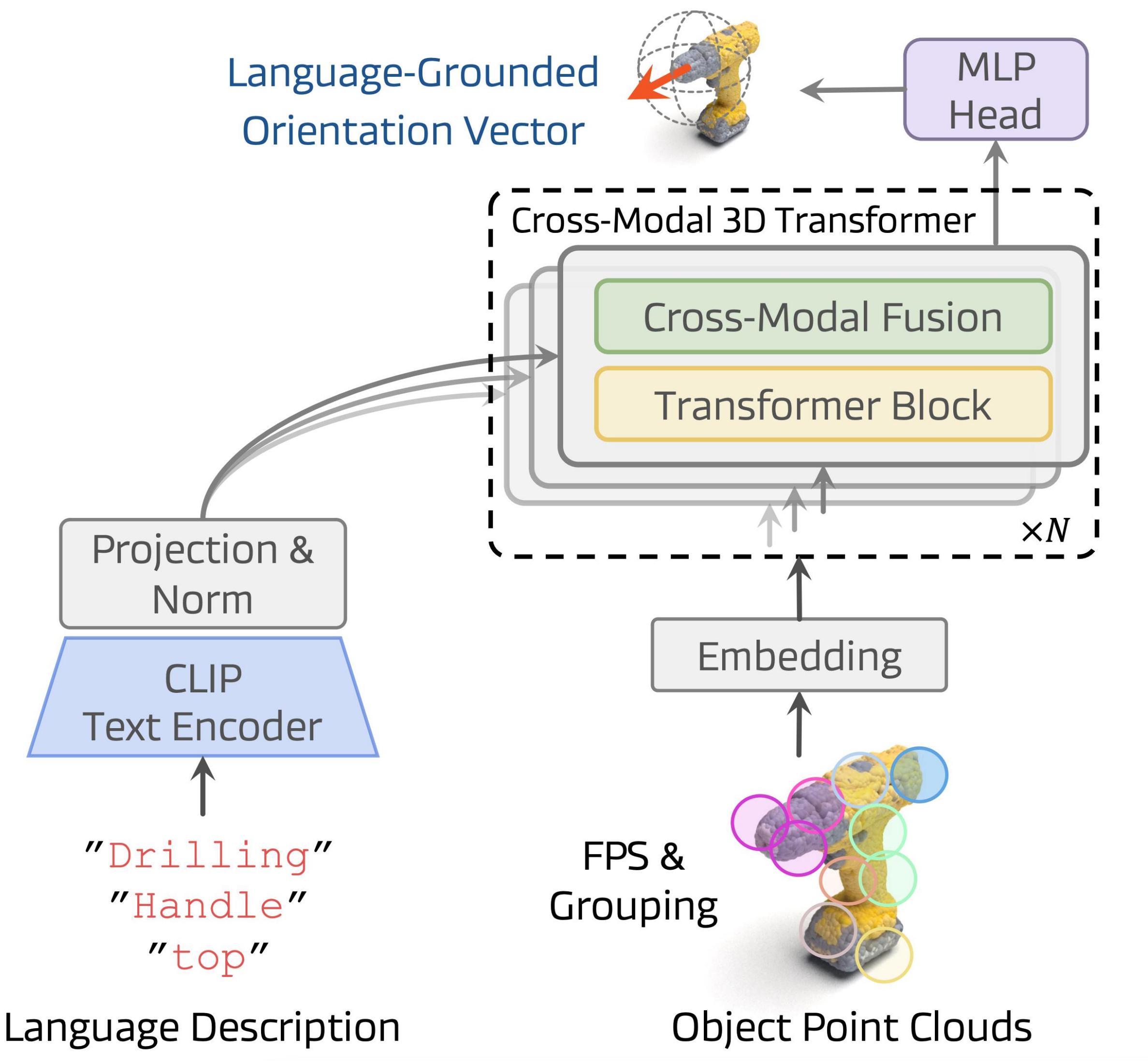


Validation:

- 210 **human-labeled** filtering labels and annotations
- GPT-4o achieves 88.3% and 97.1% mean acc

PointSO: A Cross-Modal 3D Transformer for Semantic Orientation Prediction

PointSO *Accurate, scalable, and robust SO prediction*



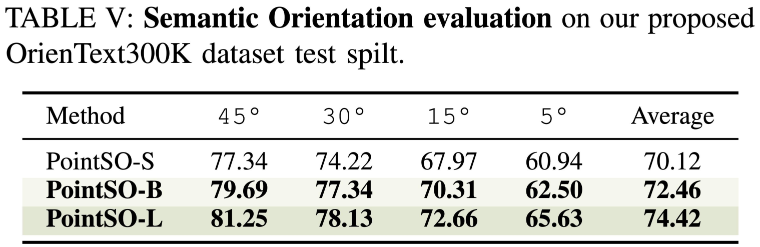


TABLE V: **Semantic Orientation evaluation** on our proposed OrienText300K dataset test spilt.

Method	45°	30°	15°	5°	Average
PointSO-S	77.34	74.22	67.97	60.94	70.12
PointSO-B	**79.69**	**77.34**	**70.31**	**62.50**	**72.46**
PointSO-L	**81.25**	**78.13**	**72.66**	**65.63**	**74.42**

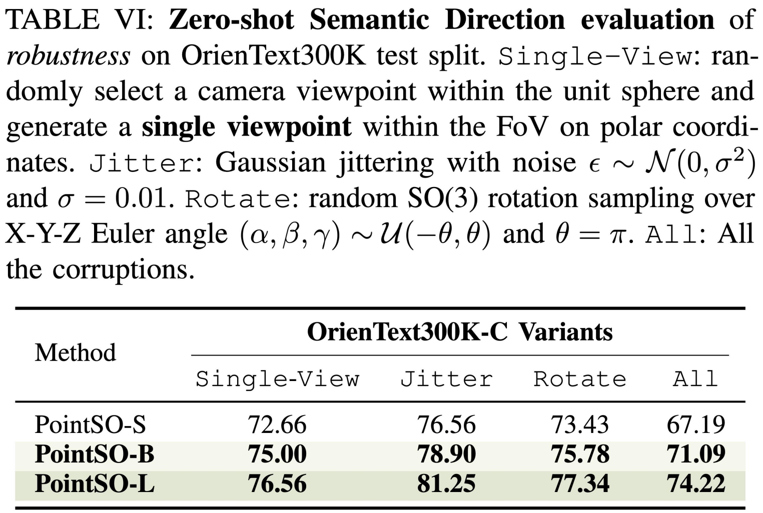


TABLE VI: **Zero-shot Semantic Direction evaluation** of *robustness* on OrienText300K test split. `Single-View`: randomly select a camera viewpoint within the unit sphere and generate a **single viewpoint** within the FoV on polar coordinates. `Jitter`: Gaussian jittering with noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ 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 $\theta = \pi$. `All`: All the corruptions.

Method	**OrienText300K-C Variants**			
	`Single-View`	`Jitter`	`Rotate`	`All`
PointSO-S	72.66	76.56	73.43	67.19
PointSO-B	**75.00**	**78.90**	**75.78**	**71.09**
PointSO-L	**76.56**	**81.25**	**77.34**	**74.22**

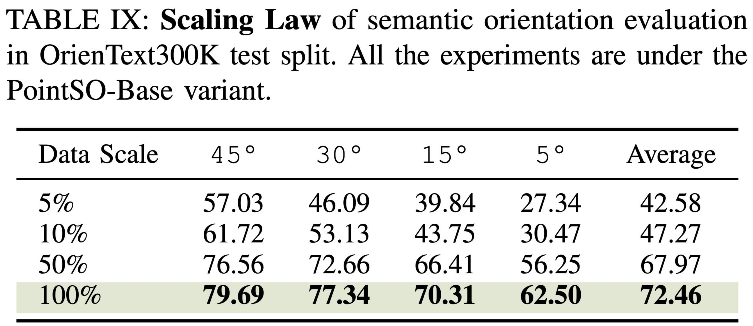


TABLE IX: **Scaling Law** of semantic orientation evaluation in OrienText300K test split. All the experiments are under the PointSO-Base variant.

Data Scale	45°	30°	15°	5°	Average
5%	57.03	46.09	39.84	27.34	42.58
10%	61.72	53.13	43.75	30.47	47.27
50%	76.56	72.66	66.41	56.25	67.97
100%	**79.69**	**77.34**	**70.31**	**62.50**	**72.46**

Model:

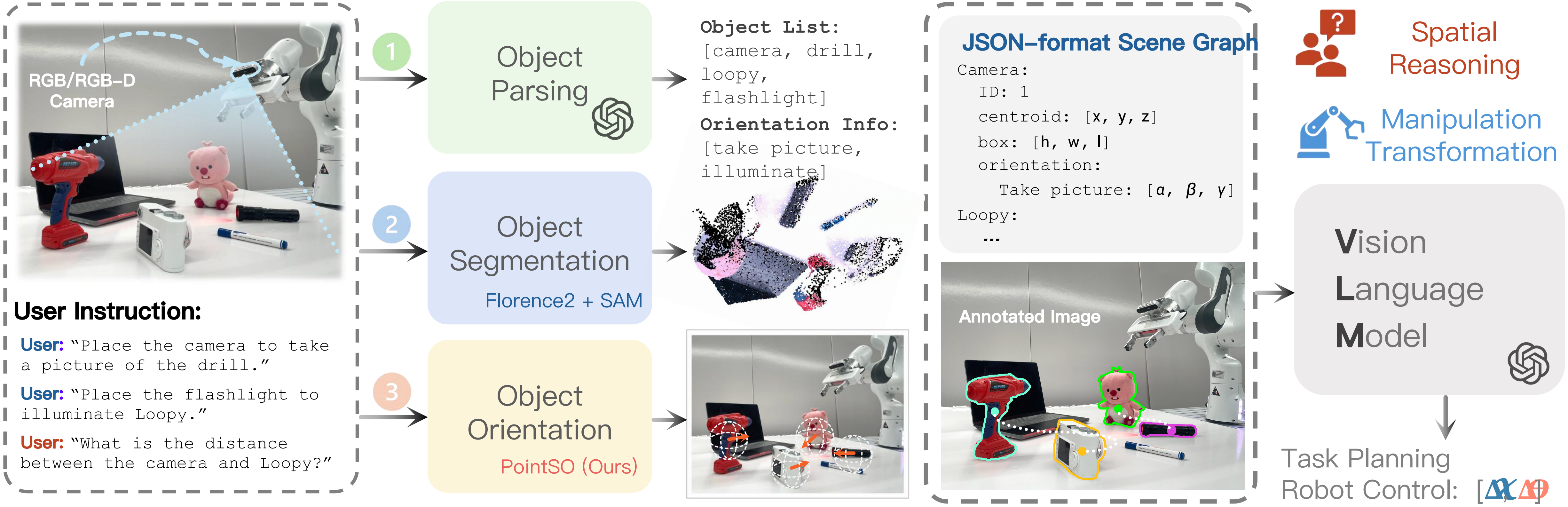
- Transformer-based Model
- Cross-Model Fusion
- 3D Augmentation

Features:

- High Accuracy
- Scalable
- Robust

SoFar Framework

Key *A hierarchical & modular framework with various Foundation Models*



User Instruction:

User: "Place the camera to take a picture of the drill."

User: "Place the flashlight to illuminate Loopy."

User: "What is the distance between the camera and Loopy?"

① Object Parsing

② Object Segmentation
Florence2 + SAM

③ Object Orientation
PointSO (Ours)

```
Object List:
[camera, drill,
loopy,
flashlight]
Orientation Info:
[take picture,
illuminate]
```

JSON-format Scene Graph

```
Camera:
  ID: 1
  centroid: [x, y, z]
  box: [h, w, l]
  orientation:
    Take picture: [α, β, γ]
Loopy:
  ...
```

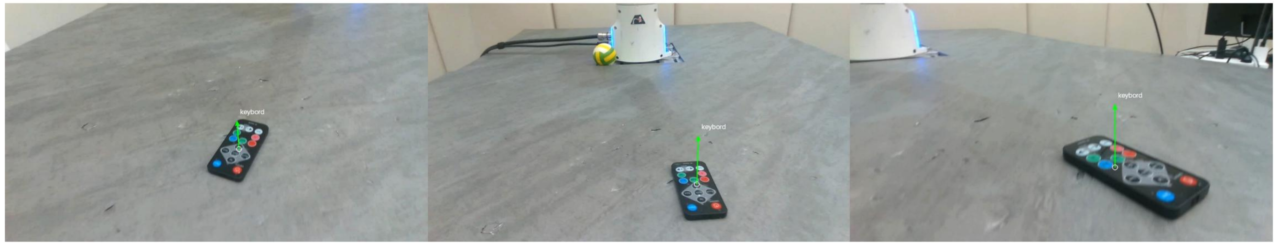
Spatial Reasoning

Manipulation Transformation

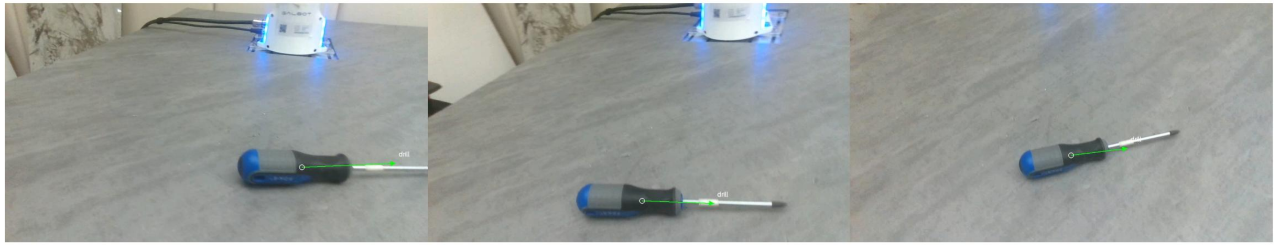
Vision
Language
Model

Task Planning
Robot Control: $[\Delta x, \Delta \theta]$

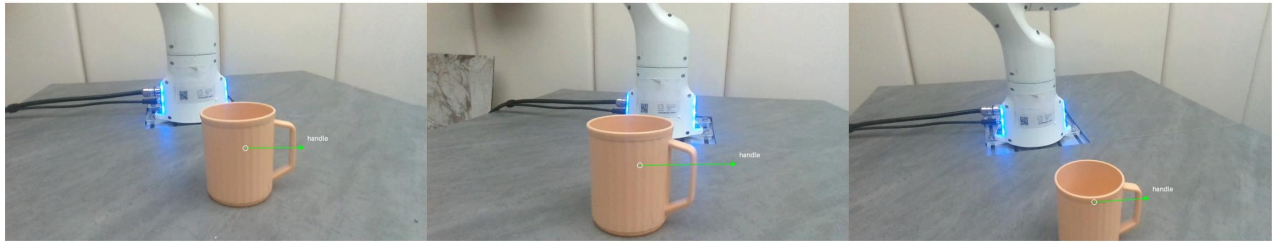
Real-World SO Prediction



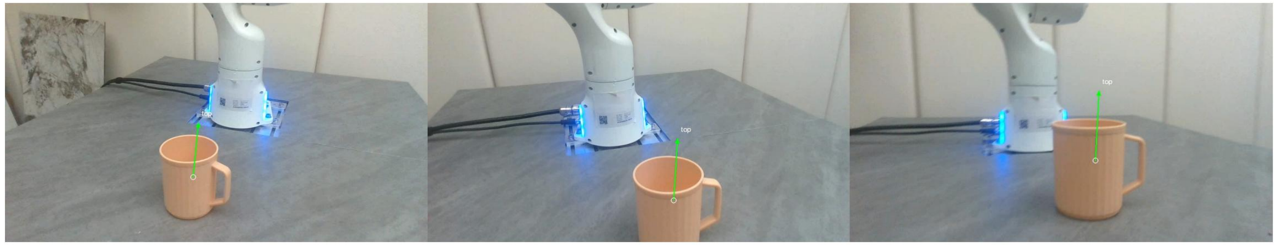
The “Screen” direction of a remote.



The “drill” direction of a screwdriver.

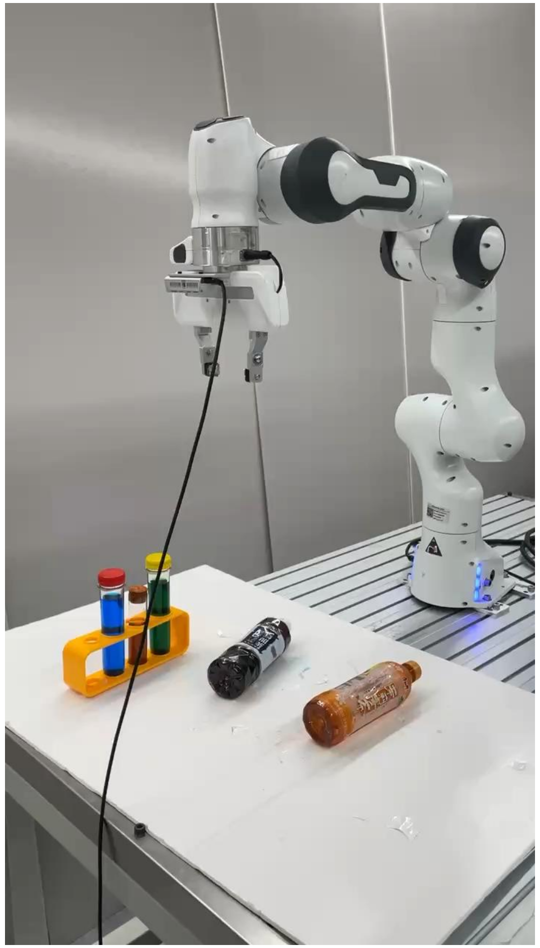


The “handle” direction of a mug.

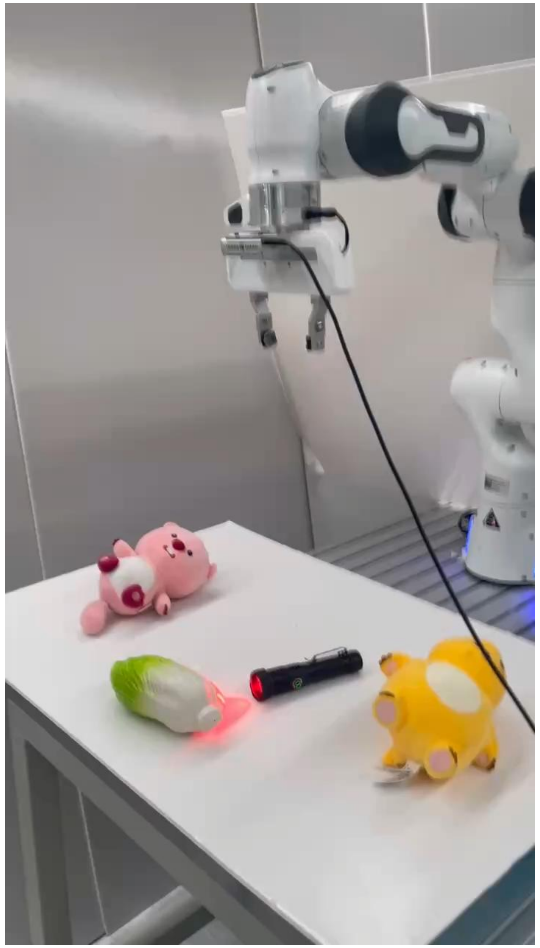


The “top” direction of a mug.

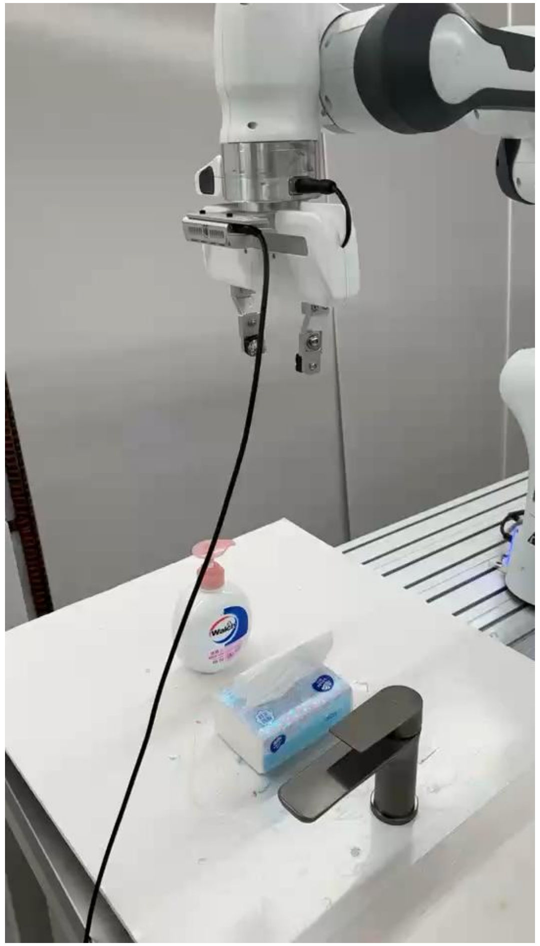
Language-Grounded Manipulation



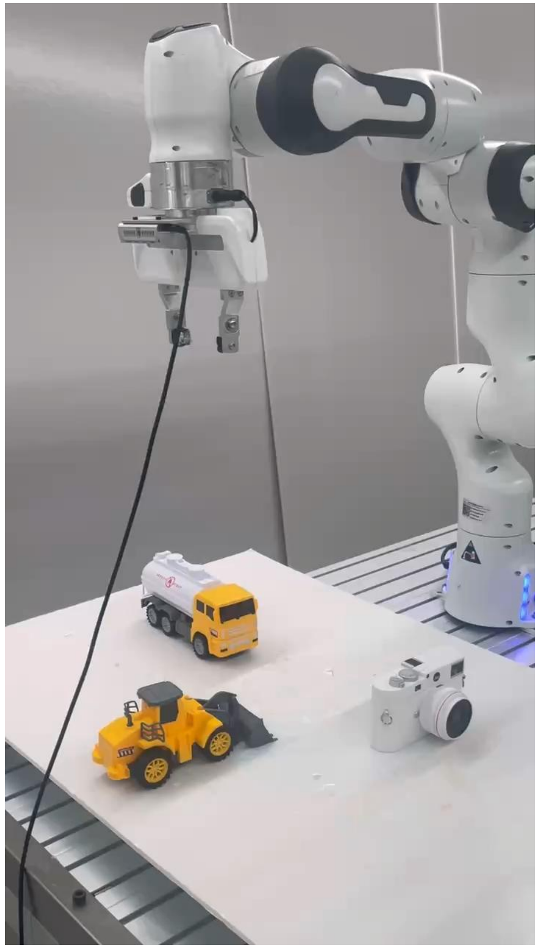
Pick up the nearest test tube and place it in the center of the



Rotate the flashlight to illuminate the loopy

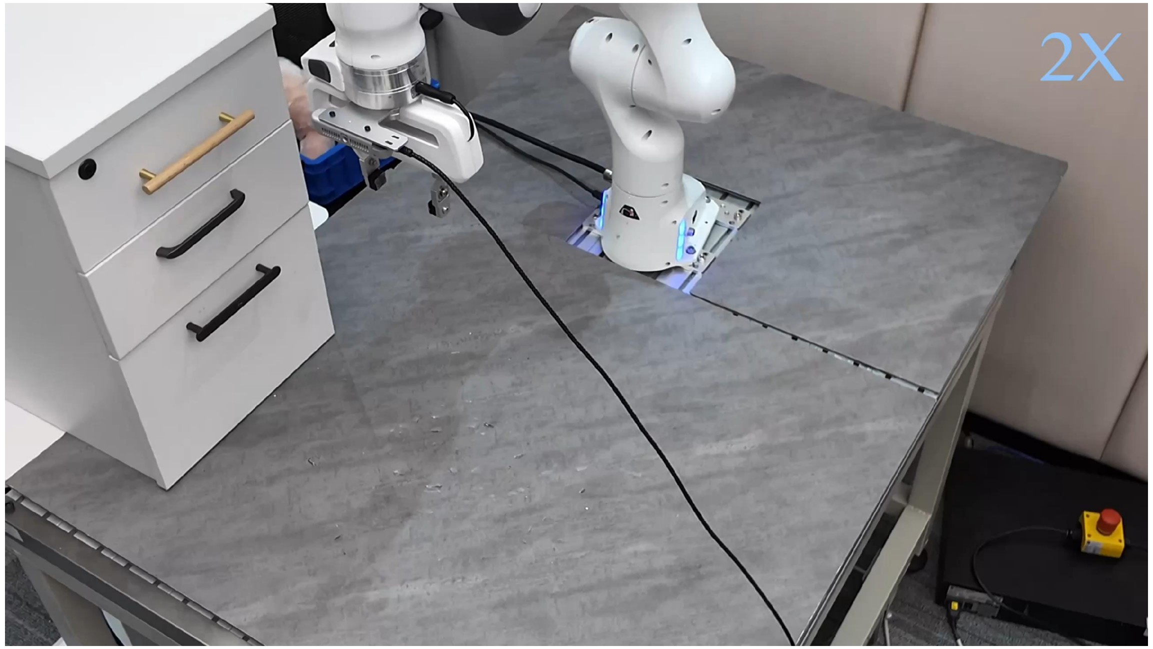


Pull out a tissue

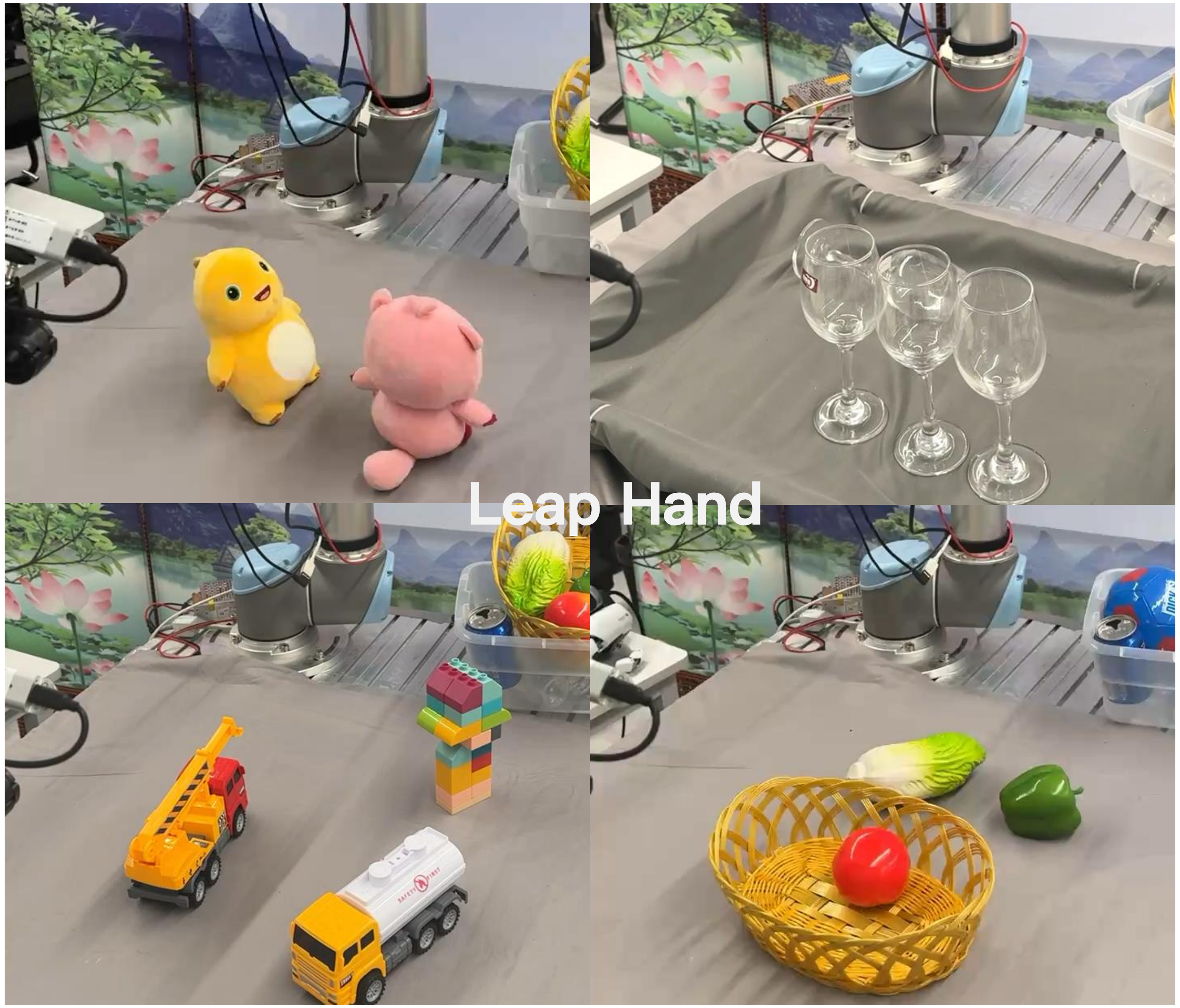


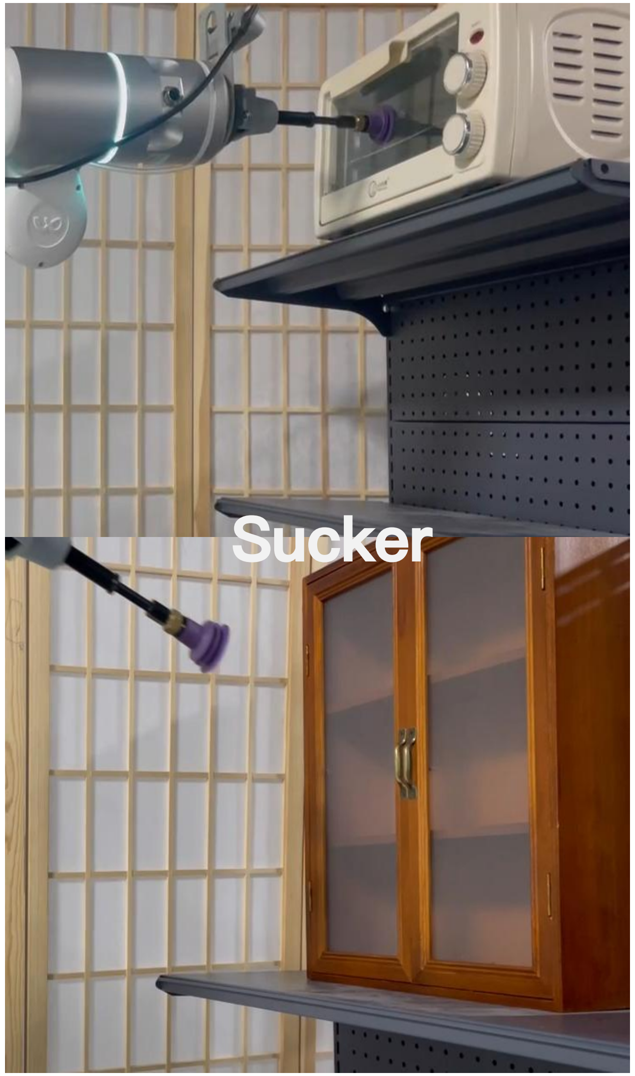
Aim the camera at the toy truck

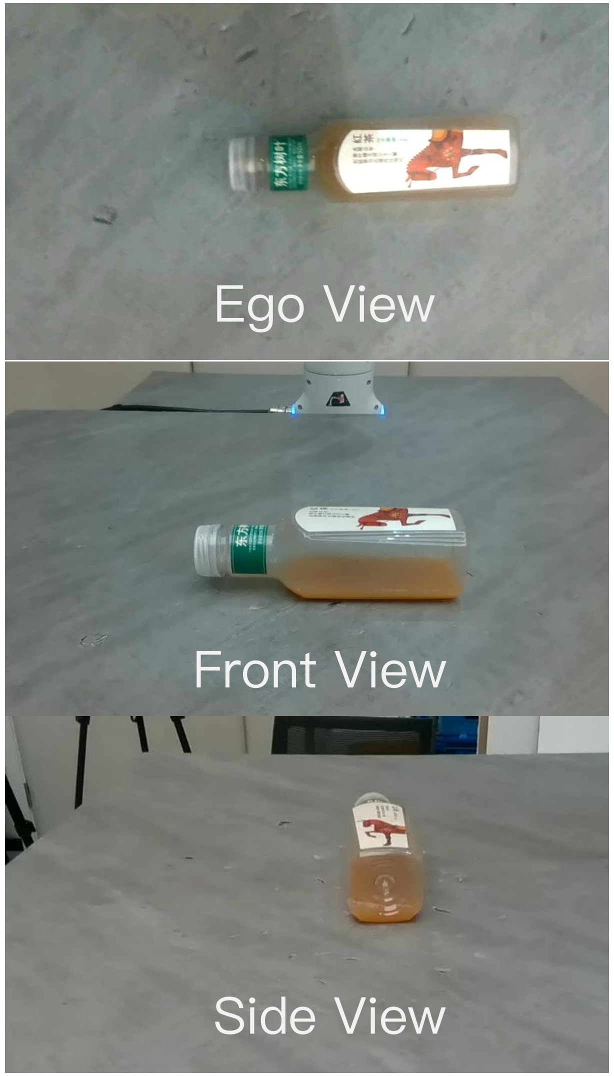
Long-Horizon Manipulation



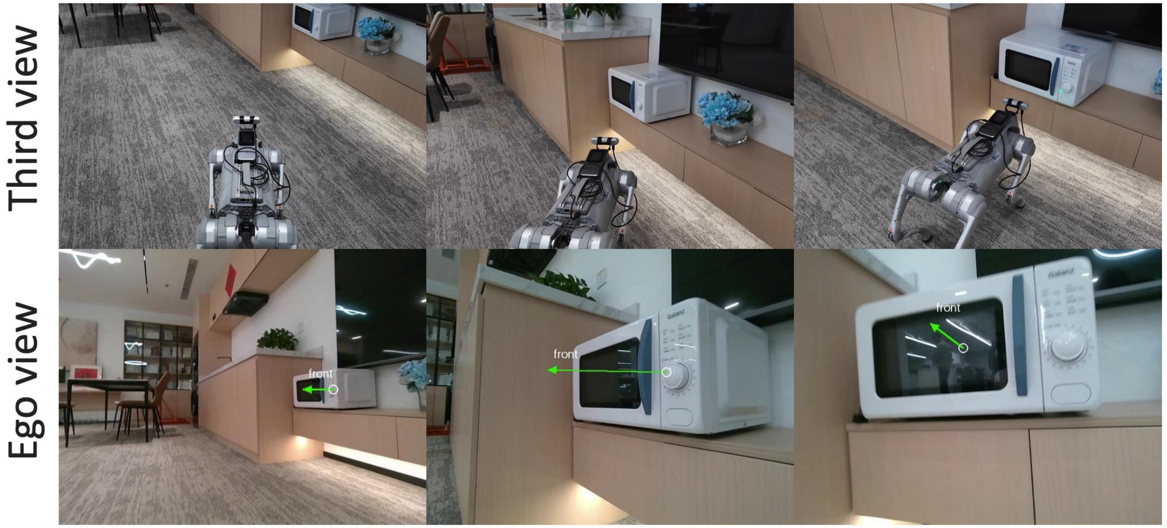
Cross Embodiment & Cross View Generalization



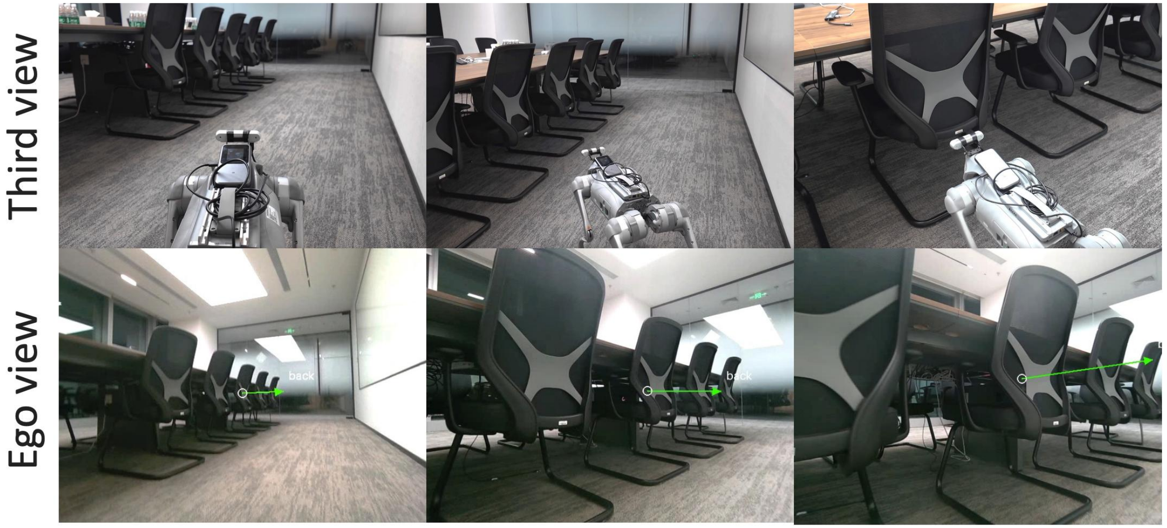




Language–Grounded Navigation

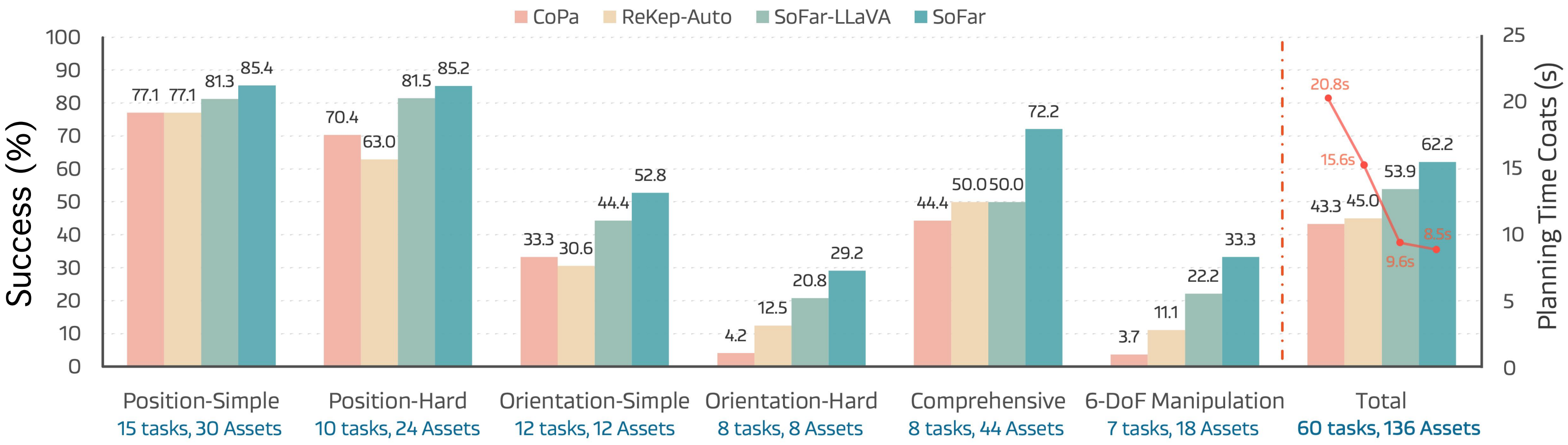


Move to facing the front of the microwave oven.



Move to facing the third chair’s back.

Results



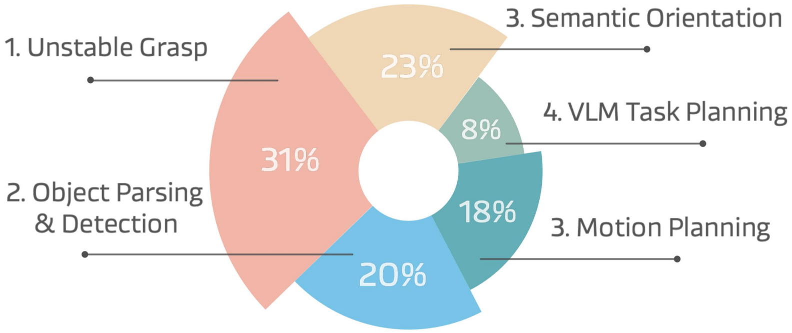


Fig. 16: **Failure case distribution analysis of our SoFar.**

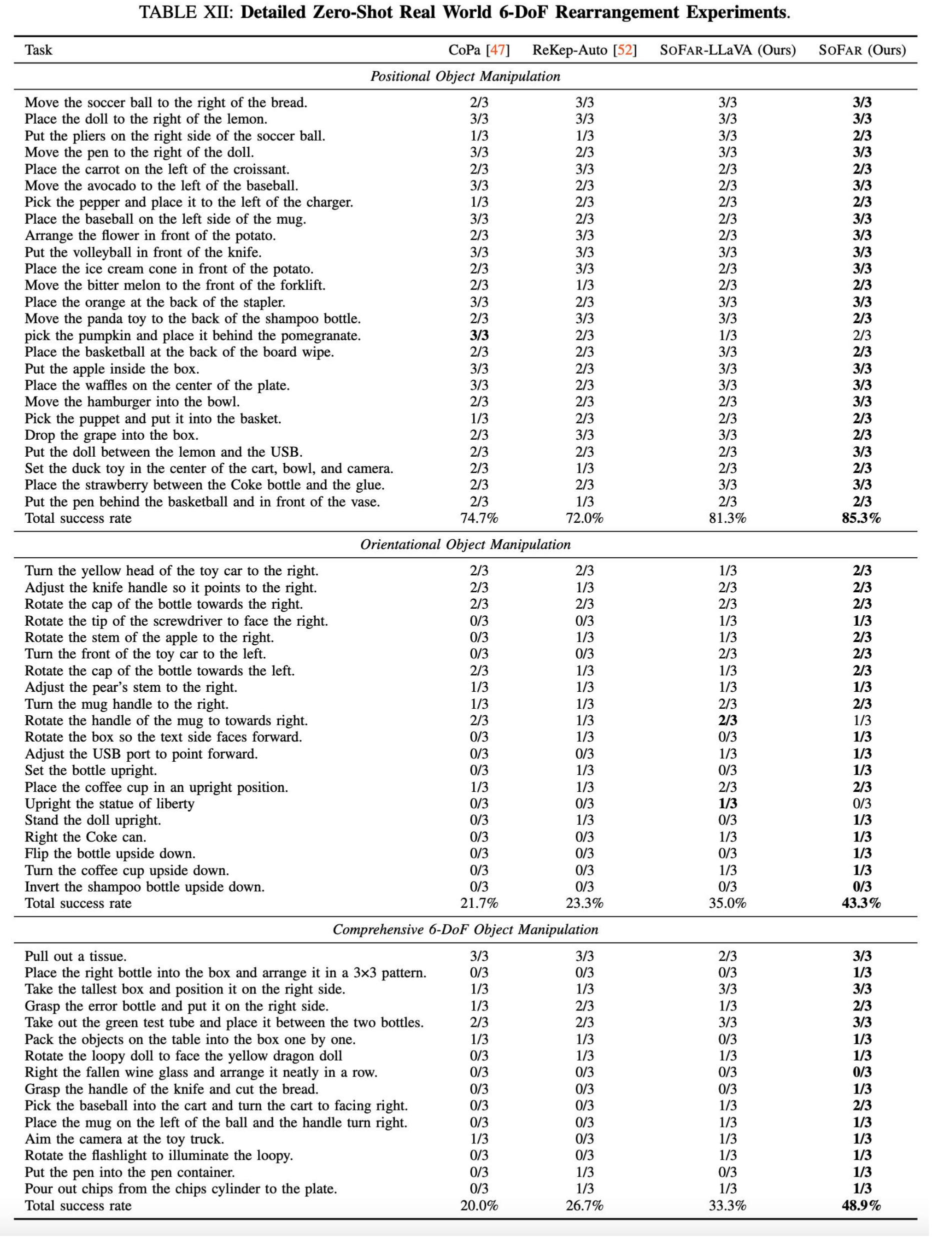


TABLE XII: **Detailed Zero-Shot Real World 6-DoF Rearrangement Experiments**.

Task	CoPa [47]	ReKep-Auto [52]	SoFar-LLaVA (Ours)	SoFar (Ours)
Positional Object Manipulation				
Move the soccer ball to the right of the bread.	2/3	3/3	3/3	**3/3**
Place the doll to the right of the lemon.	3/3	3/3	3/3	**3/3**
Put the pliers on the right side of the soccer ball.	1/3	1/3	3/3	**2/3**
Move the pen to the right of the doll.	3/3	2/3	3/3	**3/3**
Place the carrot on the left of the croissant.	2/3	3/3	2/3	**2/3**
Move the avocado to the left of the baseball.	3/3	2/3	2/3	**3/3**
Pick the pepper and place it to the left of the charger.	1/3	2/3	2/3	**2/3**
Place the baseball on the left side of the mug.	3/3	2/3	2/3	**3/3**
Arrange the flower in front of the potato.	2/3	3/3	2/3	**3/3**
Put the volleyball in front of the knife.	3/3	3/3	3/3	**3/3**
Place the ice cream cone in front of the potato.	2/3	3/3	2/3	**3/3**
Move the bitter melon to the front of the forklift.	2/3	1/3	2/3	**2/3**
Place the orange at the back of the stapler.	3/3	2/3	3/3	**3/3**
Move the panda toy to the back of the shampoo bottle.	2/3	3/3	3/3	**2/3**
pick the pumpkin and place it behind the pomegranate.	**3/3**	2/3	1/3	2/3
Place the basketball at the back of the board wipe.	2/3	2/3	3/3	**2/3**
Put the apple inside the box.	3/3	2/3	3/3	**3/3**
Place the waffles on the center of the plate.	3/3	2/3	3/3	**3/3**
Move the hamburger into the bowl.	2/3	2/3	2/3	**3/3**
Pick the puppet and put it into the basket.	1/3	2/3	2/3	**2/3**
Drop the grape into the box.	2/3	3/3	3/3	**2/3**
Put the doll between the lemon and the USB.	2/3	2/3	2/3	**3/3**
Set the duck toy in the center of the cart, bowl, and camera.	2/3	1/3	2/3	**2/3**
Place the strawberry between the Coke bottle and the glue.	2/3	2/3	3/3	**3/3**
Put the pen behind the basketball and in front of the vase.	2/3	1/3	2/3	**2/3**
Total success rate	74.7%	72.0%	81.3%	**85.3%**
Orientational Object Manipulation				
Turn the yellow head of the toy car to the right.	2/3	2/3	1/3	**2/3**
Adjust the knife handle so it points to the right.	2/3	1/3	2/3	**2/3**
Rotate the cap of the bottle towards the right.	2/3	2/3	2/3	**2/3**
Rotate the tip of the screwdriver to face the right.	0/3	0/3	1/3	**1/3**
Rotate the stem of the apple to the right.	0/3	1/3	1/3	**2/3**
Turn the front of the toy car to the left.	0/3	0/3	2/3	**2/3**
Rotate the cap of the bottle towards the left.	2/3	1/3	1/3	**2/3**
Adjust the pear's stem to the right.	1/3	1/3	1/3	**1/3**
Turn the mug handle to the right.	1/3	1/3	2/3	**2/3**
Rotate the handle of the mug to towards right.	2/3	1/3	**2/3**	1/3
Rotate the box so the text side faces forward.	0/3	1/3	0/3	**1/3**
Adjust the USB port to point forward.	0/3	0/3	1/3	**1/3**
Set the bottle upright.	0/3	1/3	0/3	**1/3**
Place the coffee cup in an upright position.	1/3	1/3	2/3	**2/3**
Upright the statue of liberty	0/3	0/3	**1/3**	0/3
Stand the doll upright.	0/3	1/3	0/3	**1/3**
Right the Coke can.	0/3	0/3	1/3	**1/3**
Flip the bottle upside down.	0/3	0/3	0/3	**1/3**
Turn the coffee cup upside down.	0/3	0/3	1/3	**1/3**
Invert the shampoo bottle upside down.	0/3	0/3	0/3	**0/3**
Total success rate	21.7%	23.3%	35.0%	**43.3%**
Comprehensive 6-DoF Object Manipulation				
Pull out a tissue.	3/3	3/3	2/3	**3/3**
Place the right bottle into the box and arrange it in a 3×3 pattern.	0/3	0/3	0/3	**1/3**
Take the tallest box and position it on the right side.	1/3	1/3	3/3	**3/3**
Grasp the error bottle and put it on the right side.	1/3	2/3	1/3	**2/3**
Take out the green test tube and place it between the two bottles.	2/3	2/3	3/3	**3/3**
Pack the objects on the table into the box one by one.	1/3	1/3	0/3	**1/3**
Rotate the loopy doll to face the yellow dragon doll	0/3	1/3	1/3	**1/3**
Right the fallen wine glass and arrange it neatly in a row.	0/3	0/3	0/3	**0/3**
Grasp the handle of the knife and cut the bread.	0/3	0/3	0/3	**1/3**
Pick the baseball into the cart and turn the cart to facing right.	0/3	0/3	1/3	**2/3**
Place the mug on the left of the ball and the handle turn right.	0/3	0/3	1/3	**1/3**
Aim the camera at the toy truck.	1/3	0/3	1/3	**1/3**
Rotate the flashlight to illuminate the loopy.	0/3	0/3	1/3	**1/3**
Put the pen into the pen container.	0/3	1/3	0/3	**1/3**
Pour out chips from the chips cylinder to the plate.	0/3	1/3	1/3	**1/3**
Total success rate	20.0%	26.7%	33.3%	**48.9%**

Simulation Experiments

TABLE I: **6-DoF object rearrangement evaluation** on Our Proposed Open6DOR V2 Benchmark.

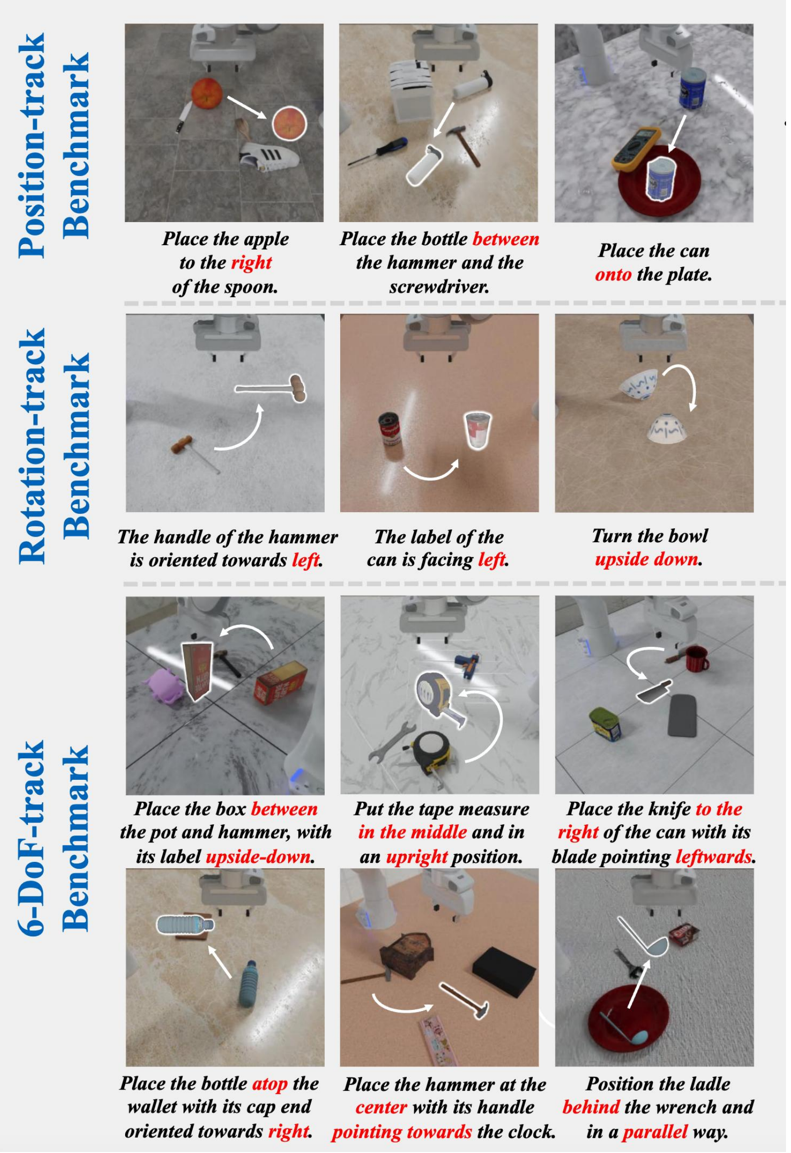
Method	Position Track				Rotation Track				6-DoF Track			Time Cost (s)
	Level 0	Level 1	Level 2	Overall	Level 0	Level 1	Level 2	Overall	Position	Rotation	Overall	
Perception Tasks on Issac Sim [90]												
GPT-4V [93]	46.8	39.1	50.0	45.2	9.1	6.9	11.7	9.2	-	-	-	-
Dream2Real [58]	17.2	11.0	0.0	15.9	37.3	27.6	26.2	31.3	26.2	18.7	13.5	358.3s
VoxPoser [50]	35.6	21.7	0.0	32.6	-	-	-	-	-	-	-	-
Open6DOR-GPT [26]	78.6	60.3	80.0	74.9	45.7	32.5	49.8	41.1	84.8	40.0	35.6	126.3 s
SoFar-LLaVA	86.3	57.9	**100.0**	78.7	62.5	30.2	67.1	48.6	83.0	48.2	40.3	9.6s
SoFar	**96.0**	**81.5**	**100.0**	**93.0**	**68.6**	**42.2**	**70.1**	**57.0**	**92.7**	**52.7**	**48.7**	**8.5s**
Execution Tasks on Libero [72]												
Octo [121]	51.2	32.1	0.0	47.2	10.7	18.3	29.9	17.2	45.6	8.0	8.0	-
OpenVLA [60]	51.6	32.4	0.0	47.6	11.0	18.5	30.6	17.6	46.2	8.2	8.2	-
SoFar	**75.3**	**65.6**	**50.0**	**72.4**	**46.6**	**29.7**	**45.8**	**34.6**	**70.1**	**33.8**	**25.6**	**40s**

TABLE XI: **Ablation study of open vocabulary detection module** on Open6DOR [26] perception tasks.

Method	Position Track				Rotation Track				6-DoF Track			Time Cost (s)
	Level 0	Level 1	Level 2	Overall	Level 0	Level 1	Level 2	Overall	Position	Rotation	Overall	
YOLO-World [15]	59.0	37.7	50.0	53.3	36.3	24.1	50.0	32.9	53.4	32.6	19.8	**7.4s**
Grounding DINO [77]	92.2	71.5	100.0	86.7	58.7	33.1	**61.8**	47.5	87.2	43.6	38.6	9.2s
Florence-2 [142]	**96.9**	**80.0**	**100.0**	**92.4**	**59.9**	**33.3**	58.2	**47.6**	**92.7**	**45.0**	**41.6**	8.5s

TABLE XV: **Statistics of Open6DOR V2 Benchmark.** The entire benchmark comprises three independent tracks, each featuring diverse tasks with careful annotations. The tasks are divided into different levels based on instruction categories, with statistics demonstrated above.

Track	Position-track								Rotation-track			6-DoF-track	Totel
Level	Level 0					Level 1		Level 2	Level 0	Level 1	Level 2	-	-
Task Catog.	Left	Right	Top	Behind	Front	Between	Center	Customized	Geometric	Directional	Semantic	-	-
Task Stat.	296	266	209	297	278	193	159	10	318	367	134	1810	4535
Benchmark Stat.	1708								1027			1810	4535


Position-track Benchmark
Place the apple to the right of the spoon.
Place the bottle between the hammer and the screwdriver.
Place the can onto the plate.
Rotation-track Benchmark
The handle of the hammer is oriented towards left.
The label of the can is facing left.
Turn the bowl upside down.
6-DoF-track Benchmark
Place the box between the pot and hammer, with its label upside-down.
Put the tape measure in the middle and in an upright position.
Place the knife to the right of the can with its blade pointing leftwards.
Place the bottle atop the wallet with its cap end oriented towards right.
Place the hammer at the center with its handle pointing towards the clock.
Position the ladle behind the wrench and in a parallel way.

Simulation Experiments

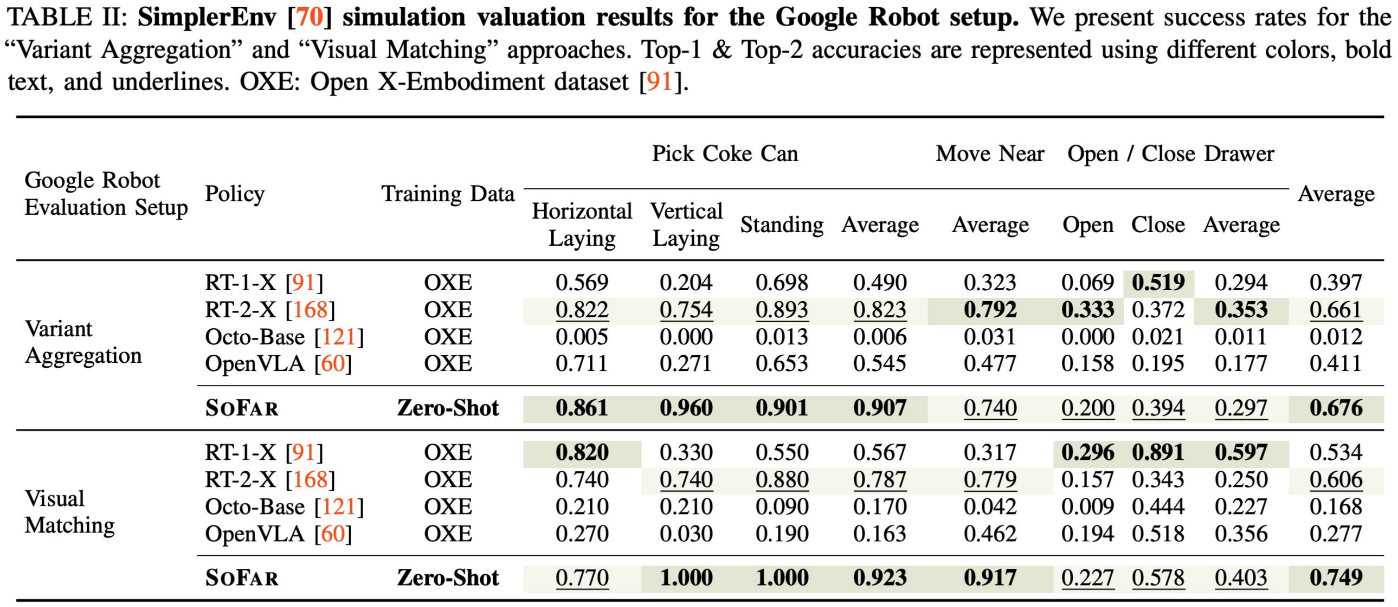


TABLE II: **SimplerEnv [70] simulation valuation results for the Google Robot setup.** We present success rates for the "Variant Aggregation" and "Visual Matching" approaches. Top-1 & Top-2 accuracies are represented using different colors, bold text, and underlines. OXE: Open X-Embodiment dataset [91].

Google Robot Evaluation Setup	Policy	Training Data	Pick Coke Can				Move Near	Open / Close Drawer			Average
			Horizontal Laying	Vertical Laying	Standing	Average	Average	Open	Close	Average	
Variant Aggregation	RT-1-X [91]	OXE	0.569	0.204	0.698	0.490	0.323	0.069	**0.519**	0.294	0.397
	RT-2-X [168]	OXE	0.822	0.754	0.893	0.823	**0.792**	**0.333**	0.372	**0.353**	0.661
	Octo-Base [121]	OXE	0.005	0.000	0.013	0.006	0.031	0.000	0.021	0.011	0.012
	OpenVLA [60]	OXE	0.711	0.271	0.653	0.545	0.477	0.158	0.195	0.177	0.411
	SoFar	**Zero-Shot**	**0.861**	**0.960**	**0.901**	**0.907**	0.740	0.200	0.394	0.297	**0.676**
Visual Matching	RT-1-X [91]	OXE	**0.820**	0.330	0.550	0.567	0.317	**0.296**	**0.891**	**0.597**	0.534
	RT-2-X [168]	OXE	0.740	0.740	0.880	0.787	0.779	0.157	0.343	0.250	0.606
	Octo-Base [121]	OXE	0.210	0.210	0.090	0.170	0.042	0.009	0.444	0.227	0.168
	OpenVLA [60]	OXE	0.270	0.030	0.190	0.163	0.462	0.194	0.518	0.356	0.277
	SoFar	**Zero-Shot**	0.770	**1.000**	**1.000**	**0.923**	**0.917**	0.227	0.578	0.403	**0.749**

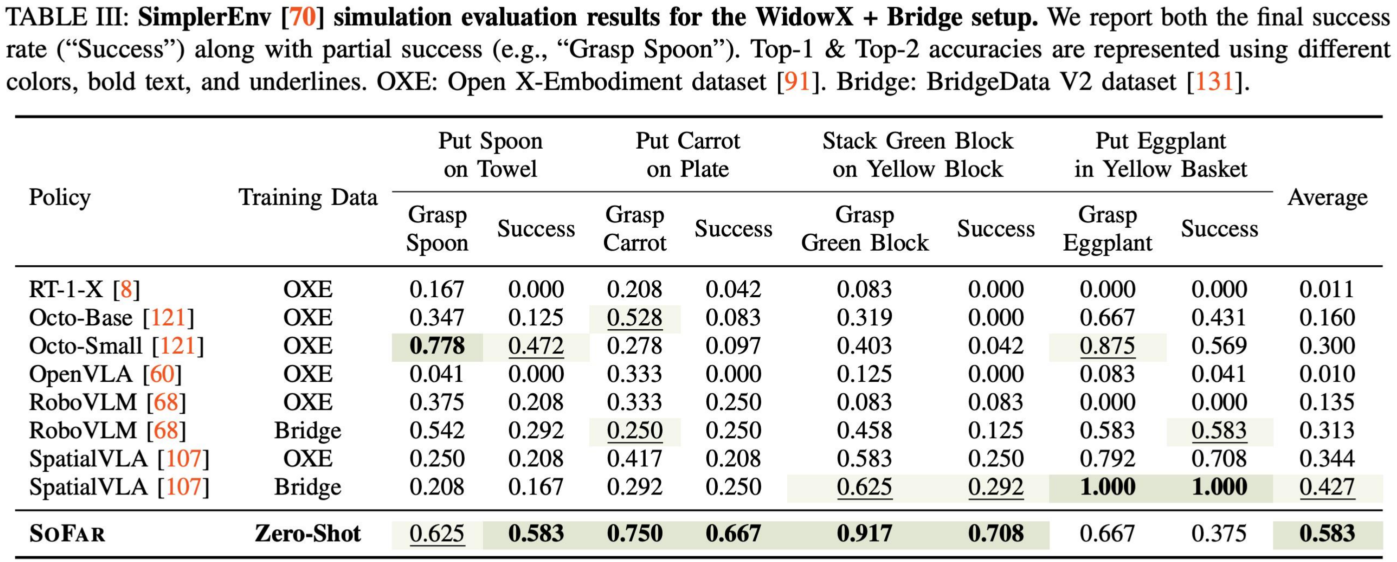
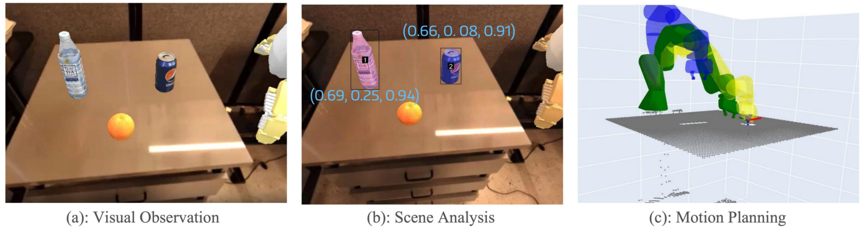


TABLE III: **SimplerEnv [70] simulation evaluation results for the WidowX + Bridge setup.** We report both the final success rate ("Success") along with partial success (e.g., "Grasp Spoon"). Top-1 & Top-2 accuracies are represented using different colors, bold text, and underlines. OXE: Open X-Embodiment dataset [91]. Bridge: BridgeData V2 dataset [131].

Policy	Training Data	Put Spoon on Towel		Put Carrot on Plate		Stack Green Block on Yellow Block		Put Eggplant in Yellow Basket		Average
		Grasp Spoon	Success	Grasp Carrot	Success	Grasp Green Block	Success	Grasp Eggplant	Success	
RT-1-X [8]	OXE	0.167	0.000	0.208	0.042	0.083	0.000	0.000	0.000	0.011
Octo-Base [121]	OXE	0.347	0.125	0.528	0.083	0.319	0.000	0.667	0.431	0.160
Octo-Small [121]	OXE	**0.778**	0.472	0.278	0.097	0.403	0.042	0.875	0.569	0.300
OpenVLA [60]	OXE	0.041	0.000	0.333	0.000	0.125	0.000	0.083	0.041	0.010
RoboVLM [68]	OXE	0.375	0.208	0.333	0.250	0.083	0.083	0.000	0.000	0.135
RoboVLM [68]	Bridge	0.542	0.292	0.250	0.250	0.458	0.125	0.583	0.583	0.313
SpatialVLA [107]	OXE	0.250	0.208	0.417	0.208	0.583	0.250	0.792	0.708	0.344
SpatialVLA [107]	Bridge	0.208	0.167	0.292	0.250	0.625	0.292	**1.000**	**1.000**	0.427
SoFar	**Zero-Shot**	0.625	**0.583**	**0.750**	**0.667**	**0.917**	**0.708**	0.667	0.375	**0.583**



(a): Visual Observation (b): Scene Analysis (c): Motion Planning

(a) Pick the Coke can.

(b) Pick the Coke can.

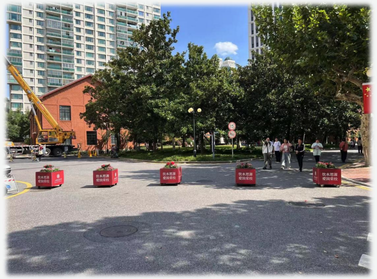
Spatial Understanding

[Task Type: Position Question Type: Absolute]

[Question]: Count from right to left and start at 1, which two of the red flower pots are the group of people in the middle of?

[A]: "4 and 5"
[B]: "2 and 3"
[C]: "1 and 2"
[D]: "3 and 4"

[Answer]: C

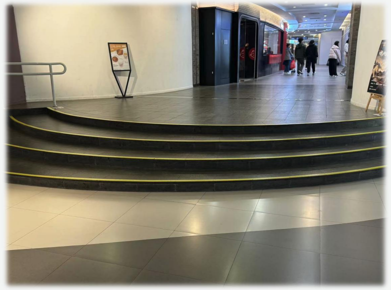


[Task Type: Position Question Type: Relative]

[Question]: Which side of the steps is narrower?

[A]: "the left"
[B]: "the right"
[C]: "the middle"
[D]: "the same"

[Answer]: B

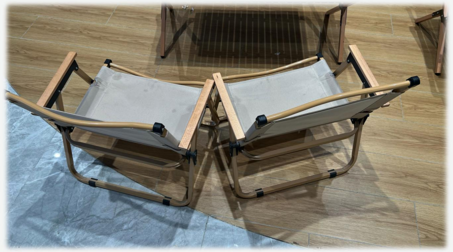


[Task Type: Orientation Question Type: Absolute]

[Question]: If you want to align the orientations of the two chairs, what is the minimum angle you need to rotate the chair on the right?

[A]: "75°"
[B]: "55°"
[C]: "35°"
[D]: "15°"

[Answer]: C

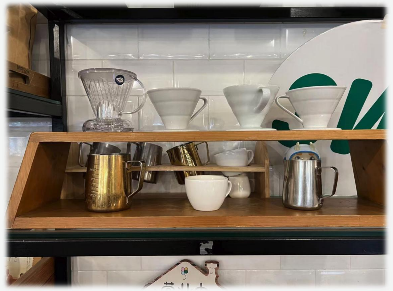


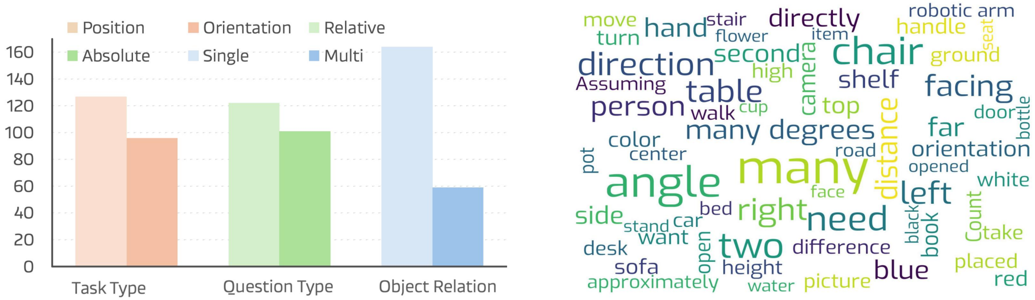
[Task Type: Orientation Question Type: Relative]

[Question]: Which direction does the handle of the cup in the upper right corner point to?

[A]: "left"
[B]: "right"
[C]: "front"
[D]: "back"

[Answer]: A

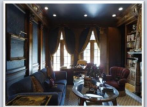




(a): Statistical Analysis (b): Word Cloud Graph

Fig. 19: **6-DoF SpatialBench Statistic**

Question: How are curtain and shelves positioned in relation to each other in the image?

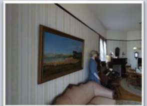


Options:
- **A. The curtain is left of the shelves.**
- B. The curtain is under the shelves.
- C. The curtain is right of the shelves.
- D. The curtain is out of the shelves.

Question: What is the spatial arrangement of pan and pepper shaker in the image concerning each other?

Options:
- **A. The pan is left of the pepper shaker.**
- B. The pan is blocking the pepper shaker.
- C. The pan is inside the pepper shaker.
- D. The pan is right of the pepper shaker.

Question: From your perspective, which object in the image is at the shortest distance?



Options:
- A. table.
- B. chair.
- **C. sculpture.**
- D. fireplace.

Question: Which object from the list is situated at the largest distance from your point of view within the image?

Options:
- A. potato.
- B. bowl.
- C. pot.
- **D. bin.**

Spatial Understanding

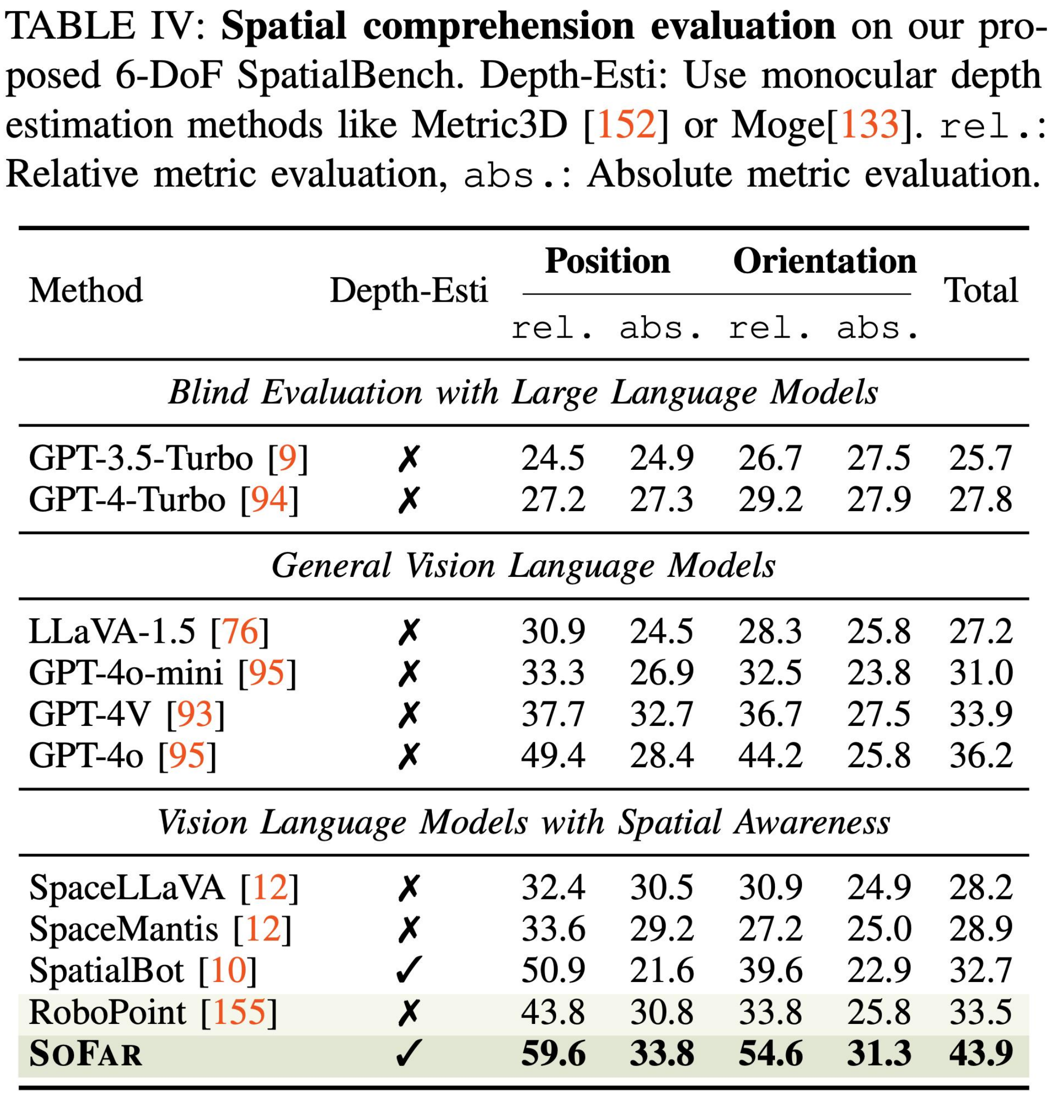


TABLE IV: **Spatial comprehension evaluation** on our proposed 6-DoF SpatialBench. Depth-Esti: Use monocular depth estimation methods like Metric3D [152] or Moge[133]. `rel.`: Relative metric evaluation, `abs.`: Absolute metric evaluation.

Method	Depth-Esti	**Position**		**Orientation**		Total
		`rel.`	`abs.`	`rel.`	`abs.`	
Blind Evaluation with Large Language Models						
GPT-3.5-Turbo [9]	✗	24.5	24.9	26.7	27.5	25.7
GPT-4-Turbo [94]	✗	27.2	27.3	29.2	27.9	27.8
General Vision Language Models						
LLaVA-1.5 [76]	✗	30.9	24.5	28.3	25.8	27.2
GPT-4o-mini [95]	✗	33.3	26.9	32.5	23.8	31.0
GPT-4V [93]	✗	37.7	32.7	36.7	27.5	33.9
GPT-4o [95]	✗	49.4	28.4	44.2	25.8	36.2
Vision Language Models with Spatial Awareness						
SpaceLLaVA [12]	✗	32.4	30.5	30.9	24.9	28.2
SpaceMantis [12]	✗	33.6	29.2	27.2	25.0	28.9
SpatialBot [10]	✓	50.9	21.6	39.6	22.9	32.7
RoboPoint [155]	✗	43.8	30.8	33.8	25.8	33.5
SoFar	✓	**59.6**	**33.8**	**54.6**	**31.3**	**43.9**

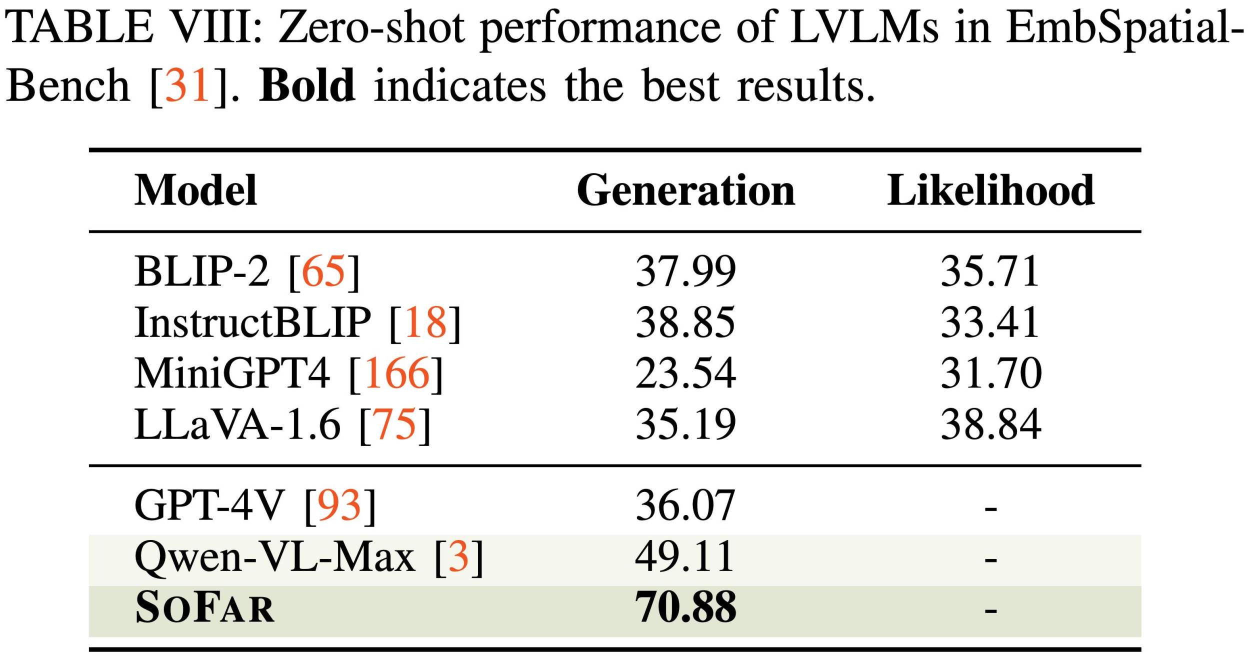


TABLE VIII: Zero-shot performance of LVLMs in EmbSpatial-Bench [31]. **Bold** indicates the best results.

Model	**Generation**	**Likelihood**
BLIP-2 [65]	37.99	35.71
InstructBLIP [18]	38.85	33.41
MiniGPT4 [166]	23.54	31.70
LLaVA-1.6 [75]	35.19	38.84
GPT-4V [93]	36.07	-
Qwen-VL-Max [3]	49.11	-
SoFar	**70.88**	-

Open Question

Should we do End2End learning or a modular design?

Thank you

Robot Setups