3D Vision-Language-Action Model

Building the 3D Generative World Model

Haoyu Zhen, SJTU

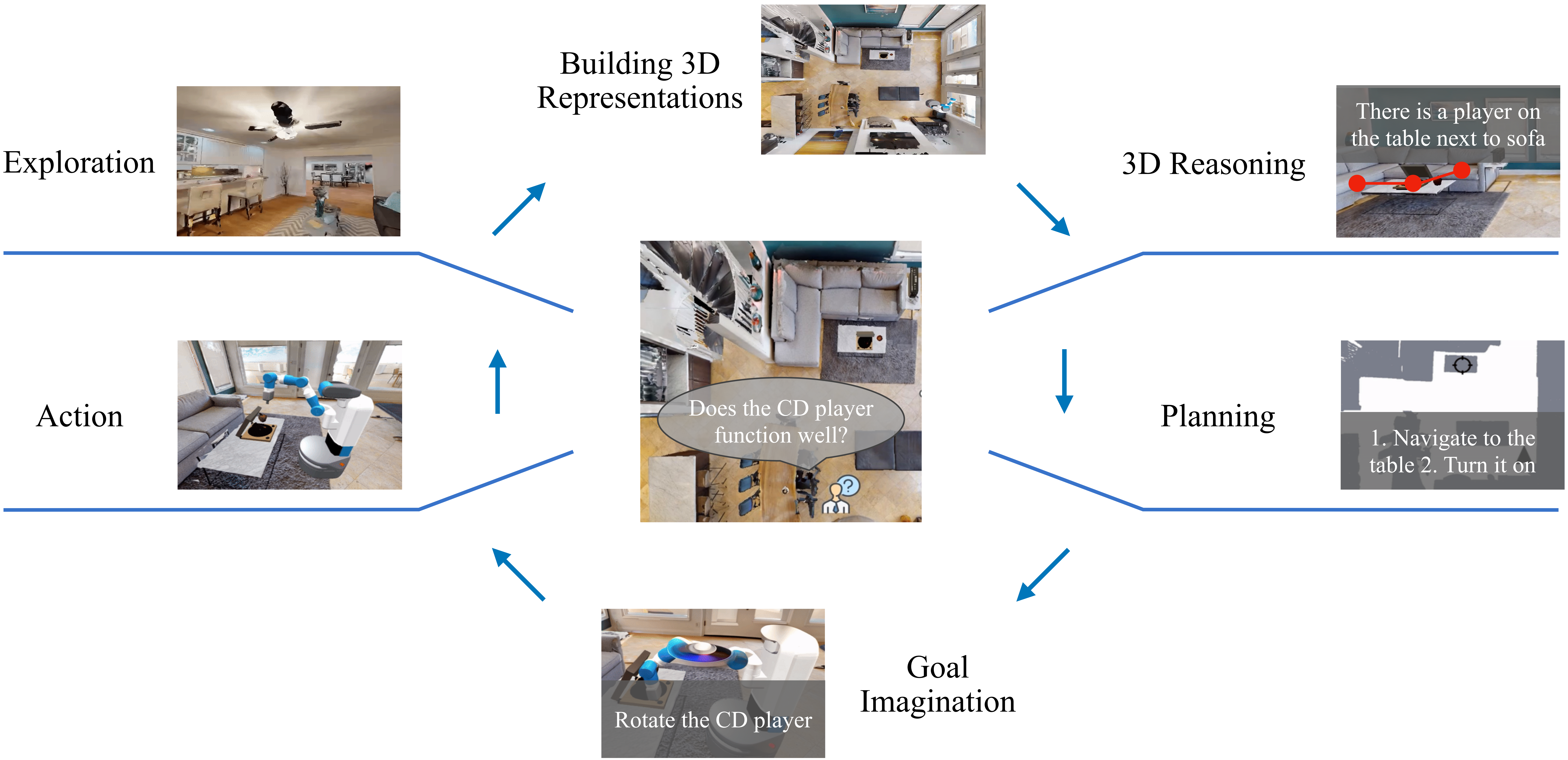
Motivation

Large Language Model

Vision Language Model

[1] OpenAI [2] Mistral AI. [3] Meta, LLaMa. [4] Stanford, Alpaca. [5] Google, Flamingo.
[6] Google, RT-2. [7] Salesforce, LAVIS. [8] LLaVa. [9] ScanNet. [10] RLbench

How Human Interact with the 3D World?



3D Vision-Language-Action Generative World Model

World Model as Foundation Model

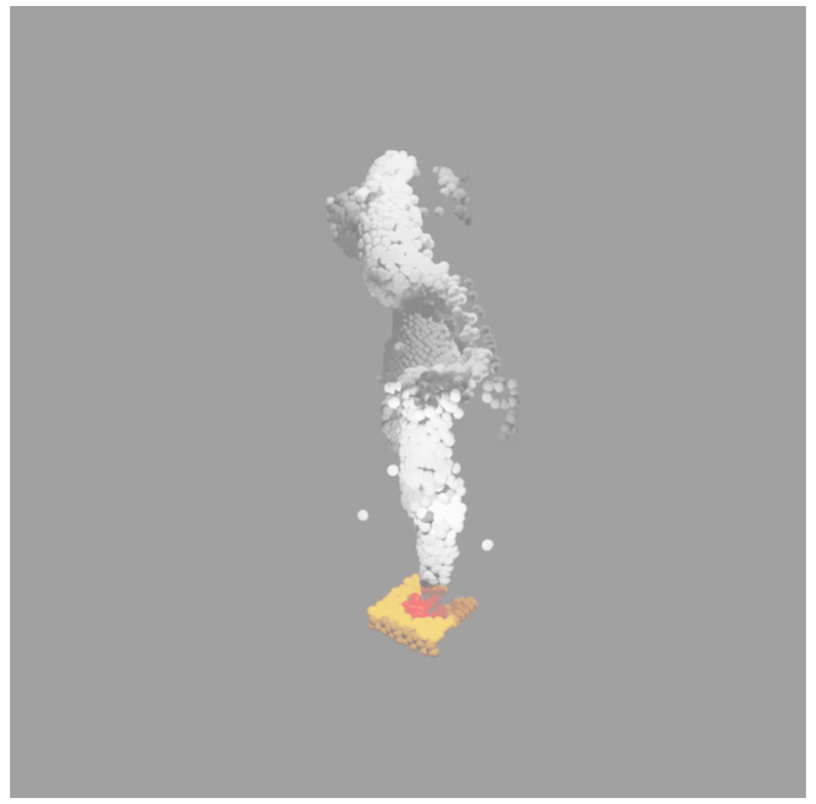
Road Map

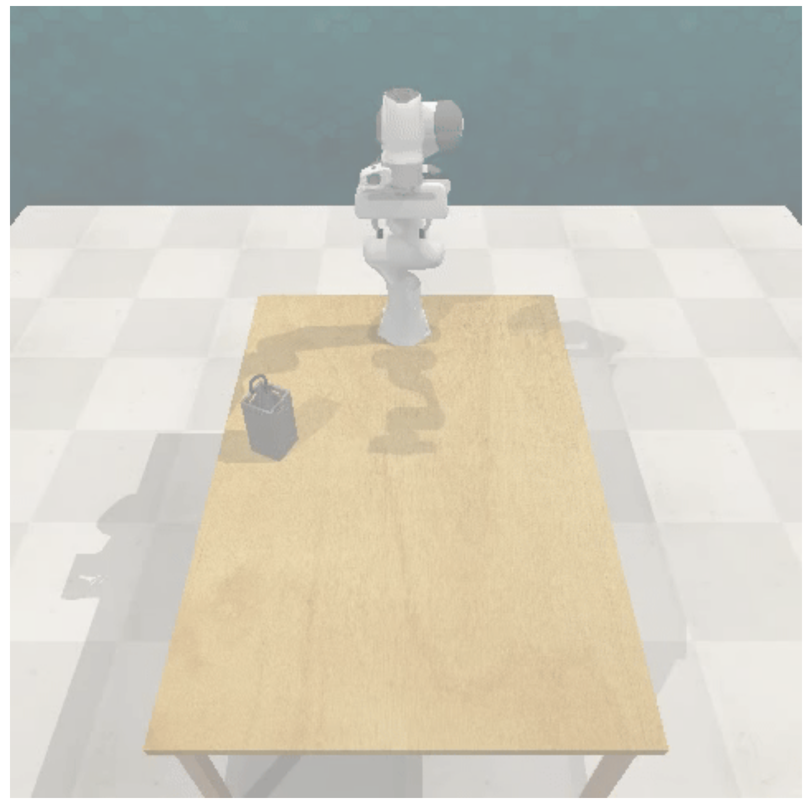
Reasoning and Planning with Embodied Foundation Models

3D-LLM, NeurIPS 2023 Spotlight

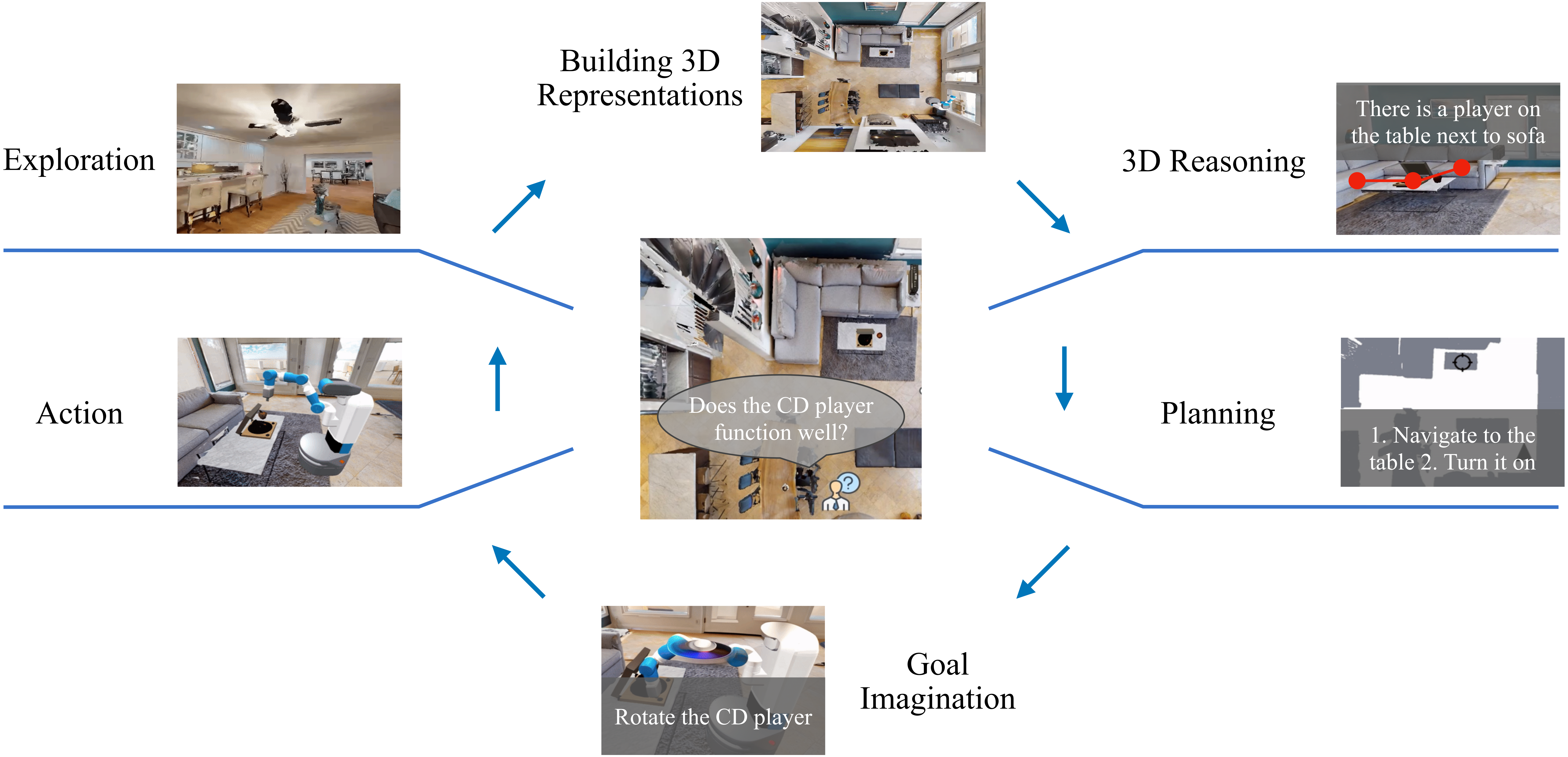
Bridging Interaction and Dynamics with Generative World Model

3D-VLA, ICML 2024

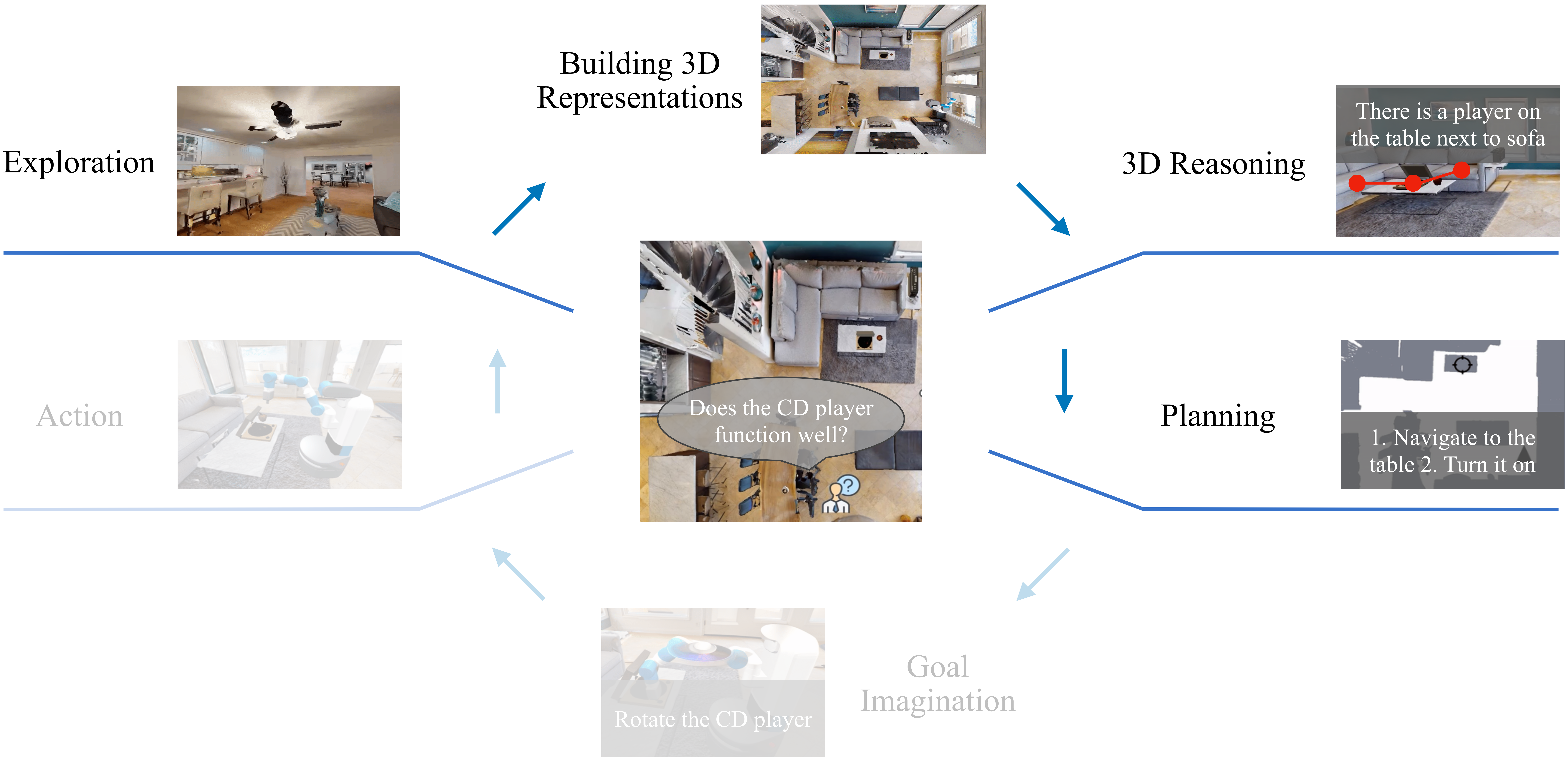




Recall: How Human Interact with the 3D World?

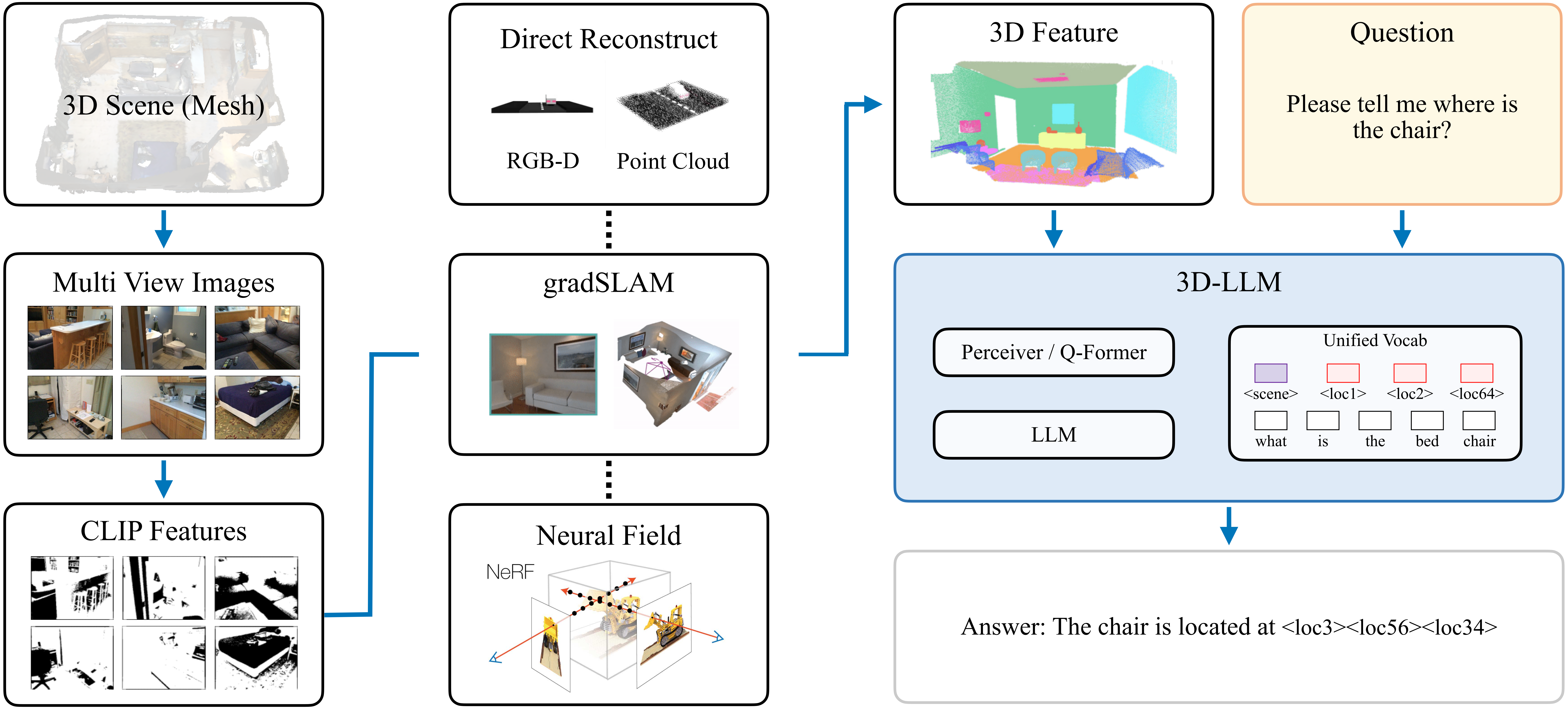


How **3D-LLM** Interact with the 3D World?



How **3D-LLM** Interact with the 3D World?

3D-LLM Framework



Limitations

1. **Overfitting** on current room datasets and object datasets.

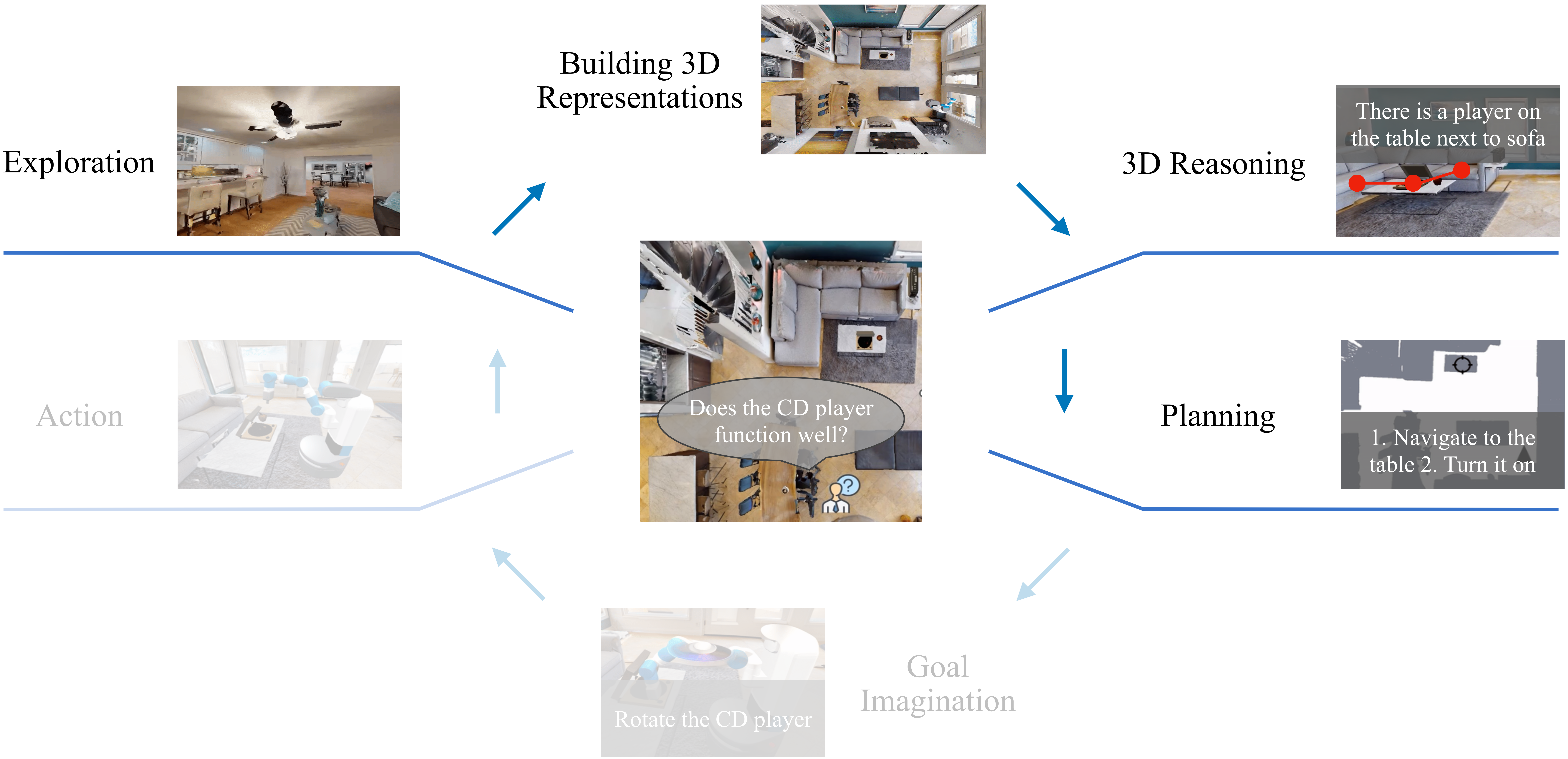
2. **Hallucination** is severe

3. Performance is much worse than traditional methods in tasks such as **localization** and **navigation**.

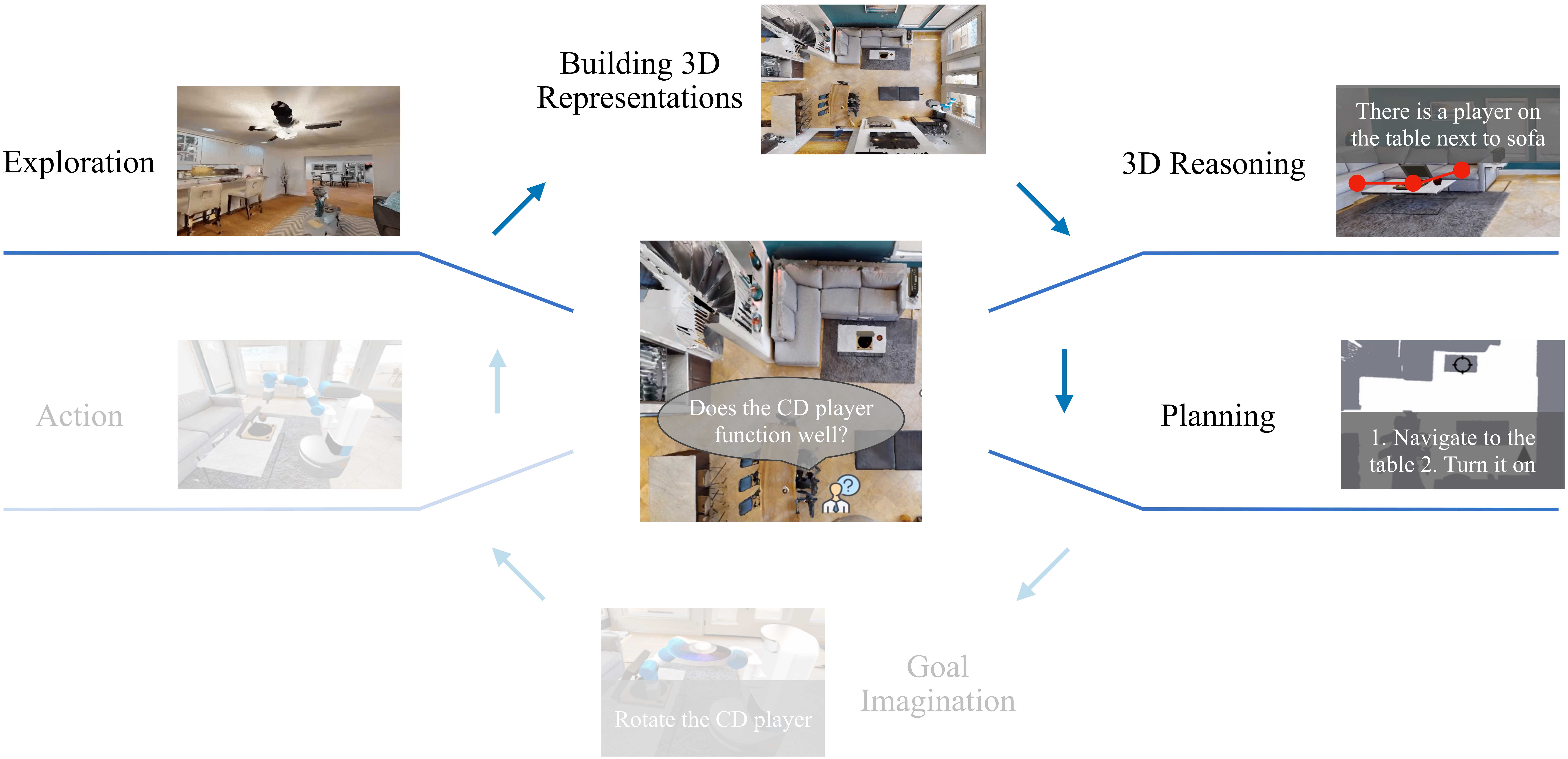
4. **Black-Box**

5. **For Robotics / Embodied AI**

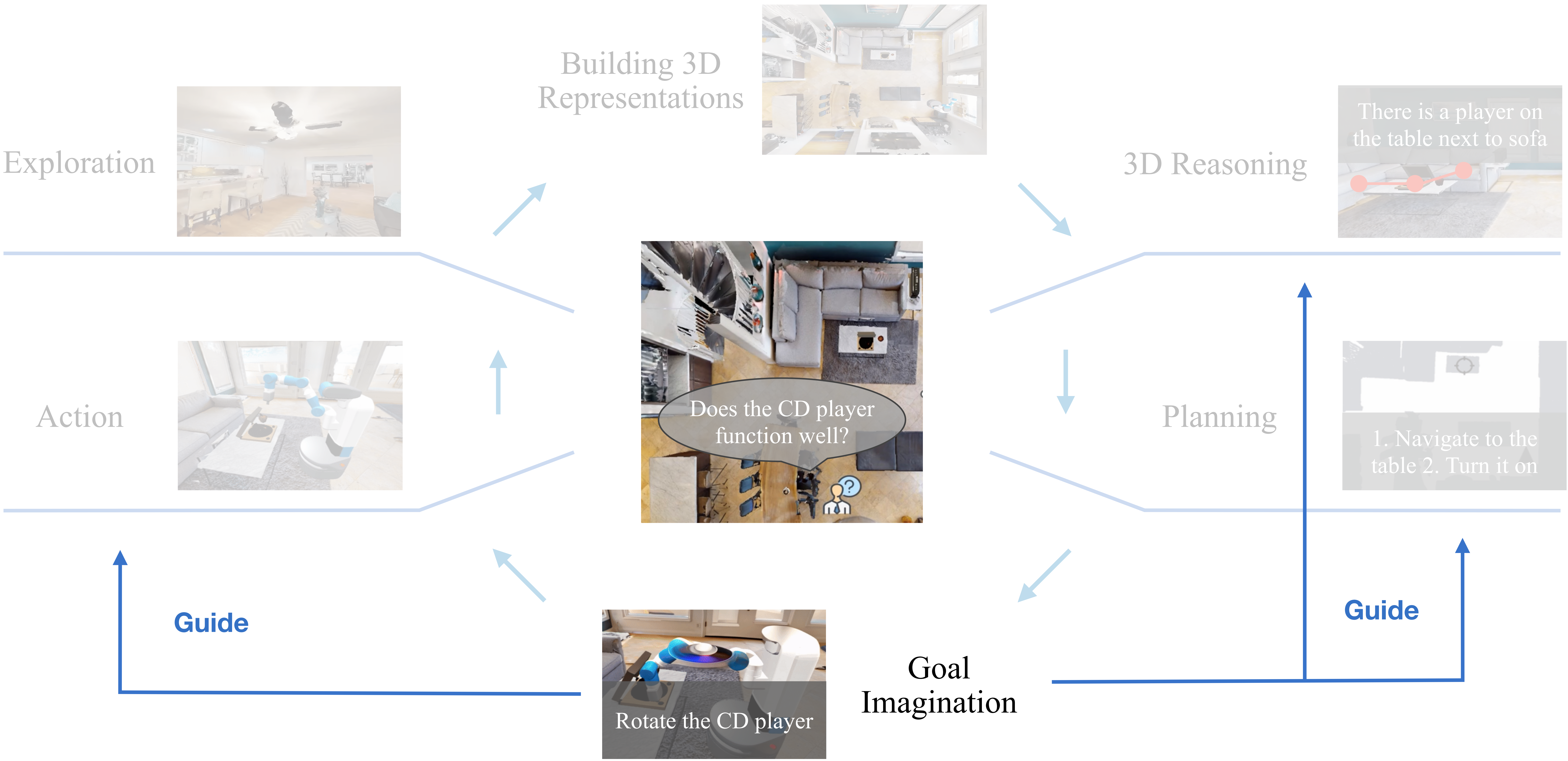
Recall: How 3D-LLM Interact with the 3D World?



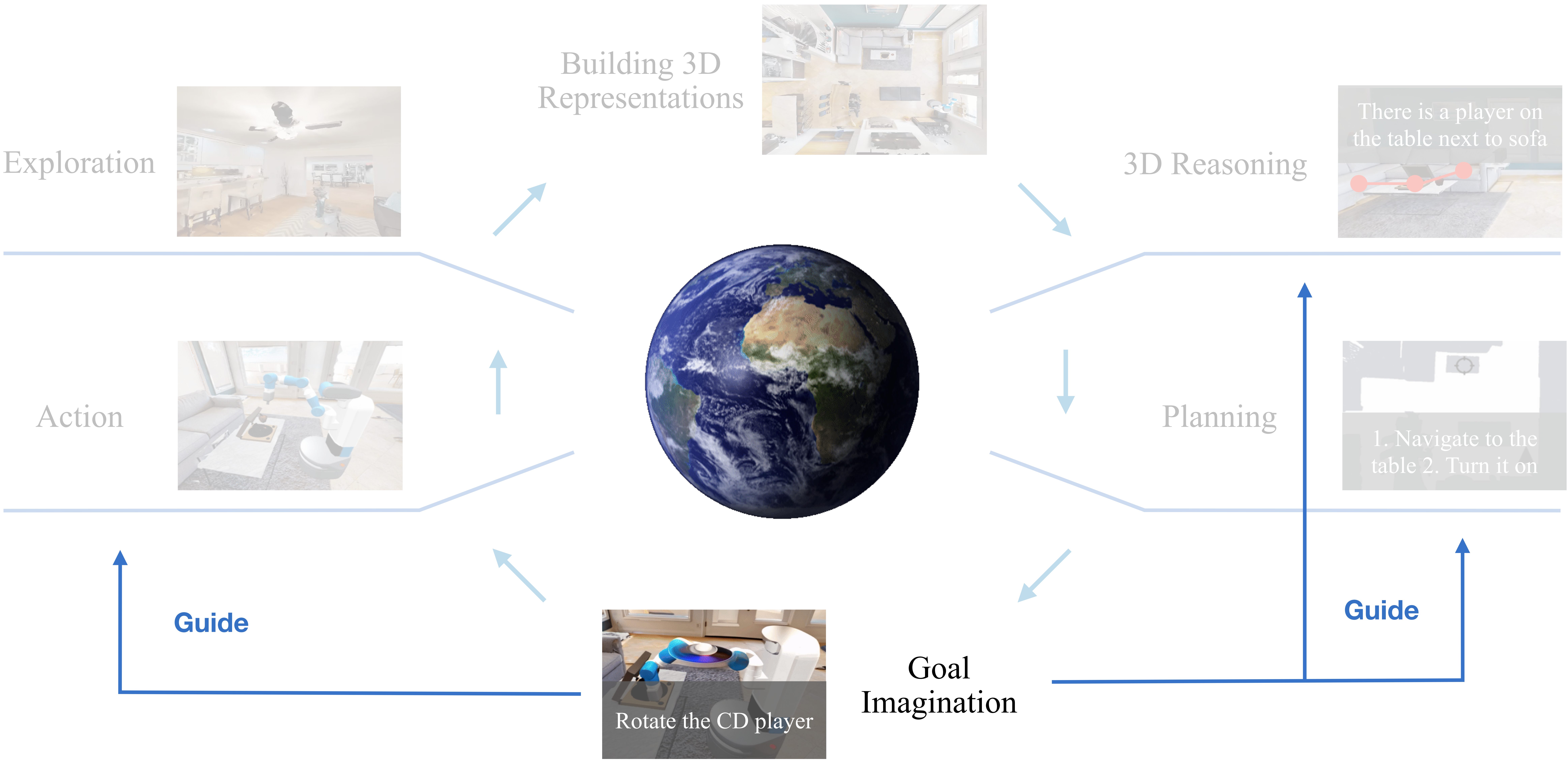
How **3D-VLA** Interact with the 3D World?



How **3D-VLA** Interact with the 3D World?



How **3D-VLA** Interact with the 3D World?



World Models

Possible Definition

Build internal **Representations** of the 3D world

Predict and simulate future events within the internal representation

Reasoning and planning: governed by our brain's prediction of the future based on our internal world model

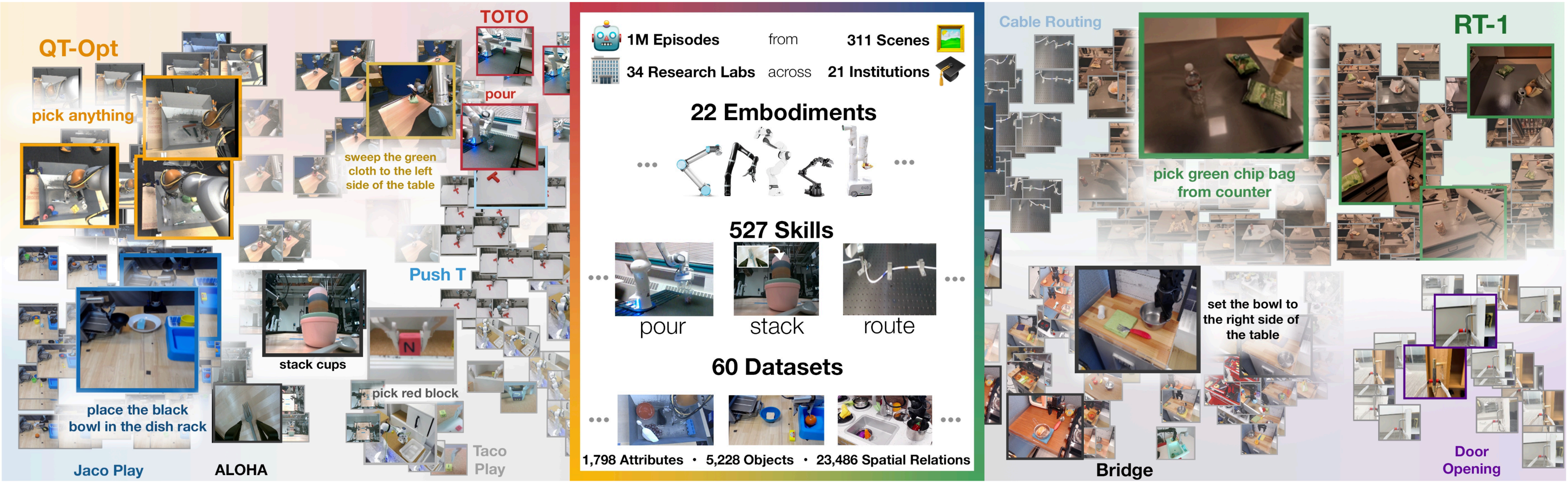
[1] Primary Visual Cortex Represents the Difference Between Past and Present. N. Nortmann et al. 2015

[2] Counterintuitive behavior of social systems. J.W. Forrester. 1971.

[3] Motion-Dependent Representation of Space in Area MT+. M. Gerrit et al. 2013

Embodied Instruction Tuning Dataset

OpenX Embodiment was released



What we have: 2D-instruction pairs

However, where do the 3D information and language data come from?

[1] Padalkar, Abhishek, et al. "Open x-embodiment”

Embodied Instruction Tuning Dataset

Lift 2D to 3D

ZoeDepth + RAFT + Grounded SAM + GPT4-V

ZoeDepth: state-of-the-art Depth estimator.

RAFT: compute the Optical Flow. To obtain the background, the moving robotic arm and the manipulated object.

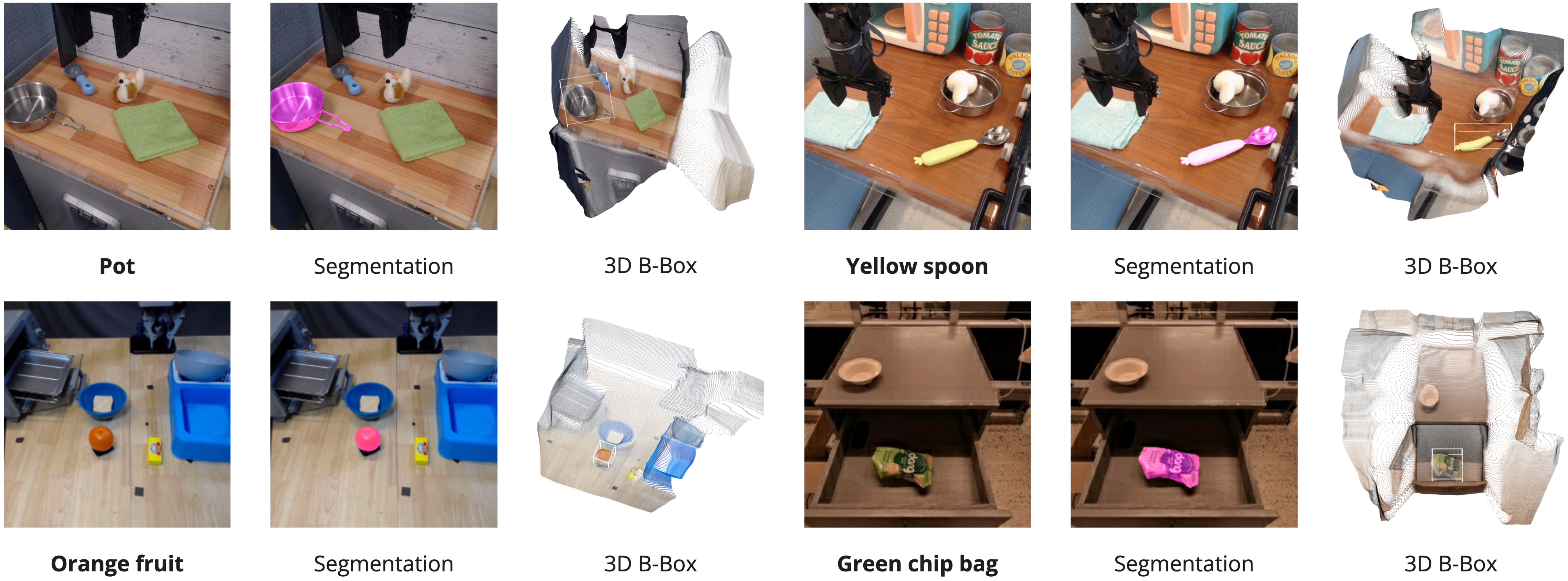
Grounded SAM: detect and segment anything with text inputs. To get the mask of the object.

GPT4-V: diverse language data.

Embodied Instruction Tuning Dataset

Lift 2D to 3D

ZoeDepth + RAFT + Grounded SAM + GPT4-V



Embodied Instruction Tuning Dataset

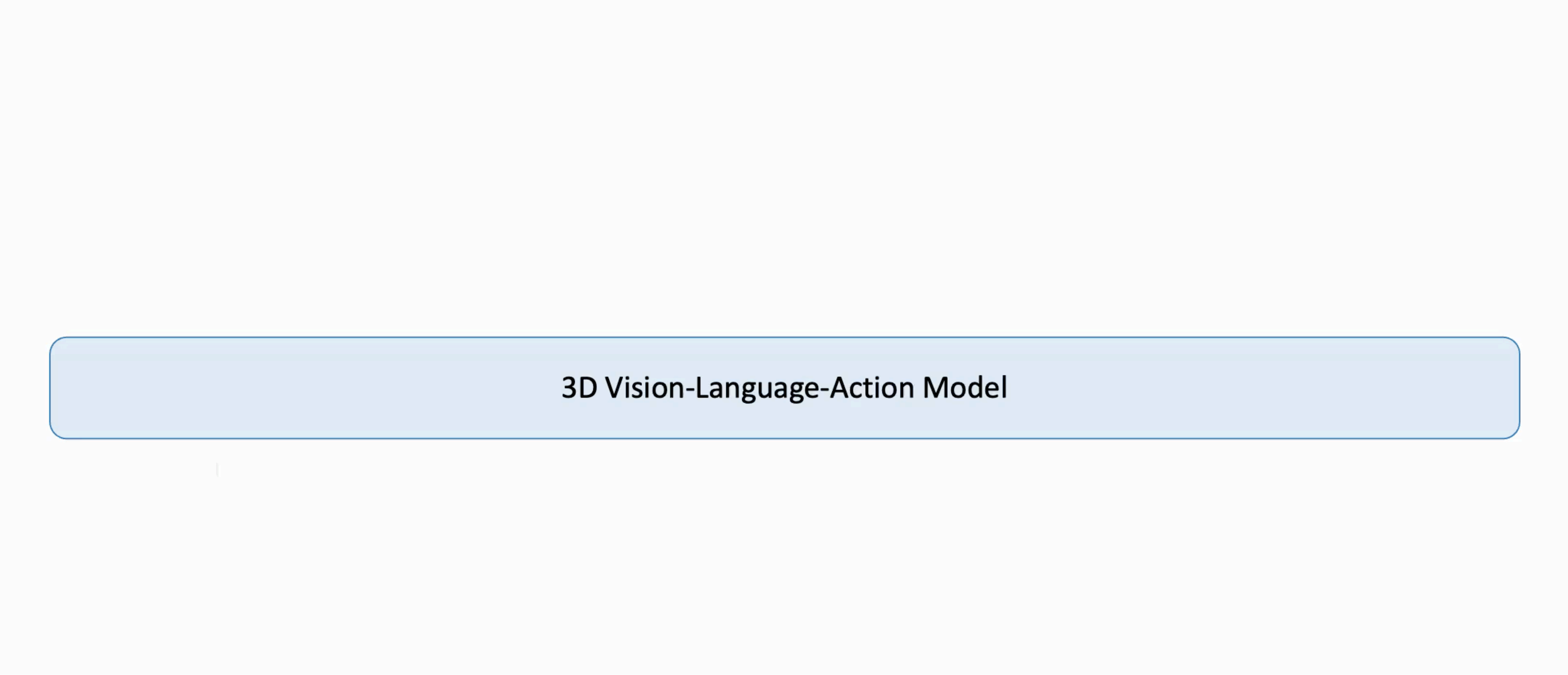
Datasets Statistics

Dataset	# of Used Episodes	Reasoning and Perception					Goal Generation			Decision Making
		Embodied QA What-if QA	Task Caption (w/ Object Grounding)	Dense Caption	Verification	Detection	Image	Depth	Point Cloud	Action Prediction
Robotics Datasets	305k	✓	✓	✓	✓	✓	✓	✓	✓	✓
BC-Z	40k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Bridge	25k	✓	✓	✓	✓	✓	✓	✓	✓	✓
CALVIN	10k	-	-	-	✓	✓	✓	✓	✓	✓
Dobb-E	20k	✓	✓	-	✓	-	✓	✓	✓	✓
Fractal	70k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Jaco Play	0.9k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Lang Table	13k	✓	✓	-	-	-	✓	✓	✓	✓
Mutex	1.5k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Pick&Place	1.3k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Play Fusion	0.5k	✓	✓	✓	✓	✓	✓	✓	✓	✓
Playing Food	4.2k	✓	✓	-	✓	-	✓	✓	✓	✓
RH20T	2.0k	✓	✓	✓	✓	✓	✓	✓	✓	✓
RLBench	50k	-	-	-	✓	✓	✓	✓	✓	✓
Roboturk	2.0k	-	-	-	✓	-	✓	✓	✓	✓
RoboVQA	61k	✓	-	-	-	-	-	-	-	-
Taco Play	3.2k	✓	✓	✓	✓	✓	✓	✓	✓	✓
HOI Datasets	11k	-	-	-	-	-	✓	✓	✓	-
Epic Kitchen	6k	-	-	-	-	-	✓	✓	✓	-
HOI4D	5k	-	-	-	-	-	✓	✓	✓	-
All Datasets	316k	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 8. Datasets used in our paper. We categorize them into four categories: Robotics, HOI, and Room datasets.

Bridging Interaction and Dynamics

3D-VLA Architecture



Bridging Interaction and Dynamics

Interactive Tokens

User: The initial scene is <scene> [init embed] </scene> Find some snacks for me.

Robot: Sure! I should <img pcd> pick up <obj> the chip bag </obj> [loc tokens] </img /pcd>

User: <scene> [goal embed] </scene> Execute now.

Robot: Actions are: [action tokens]

<scene> </scene>: to separate the 3D features and word embeddings in an LLM

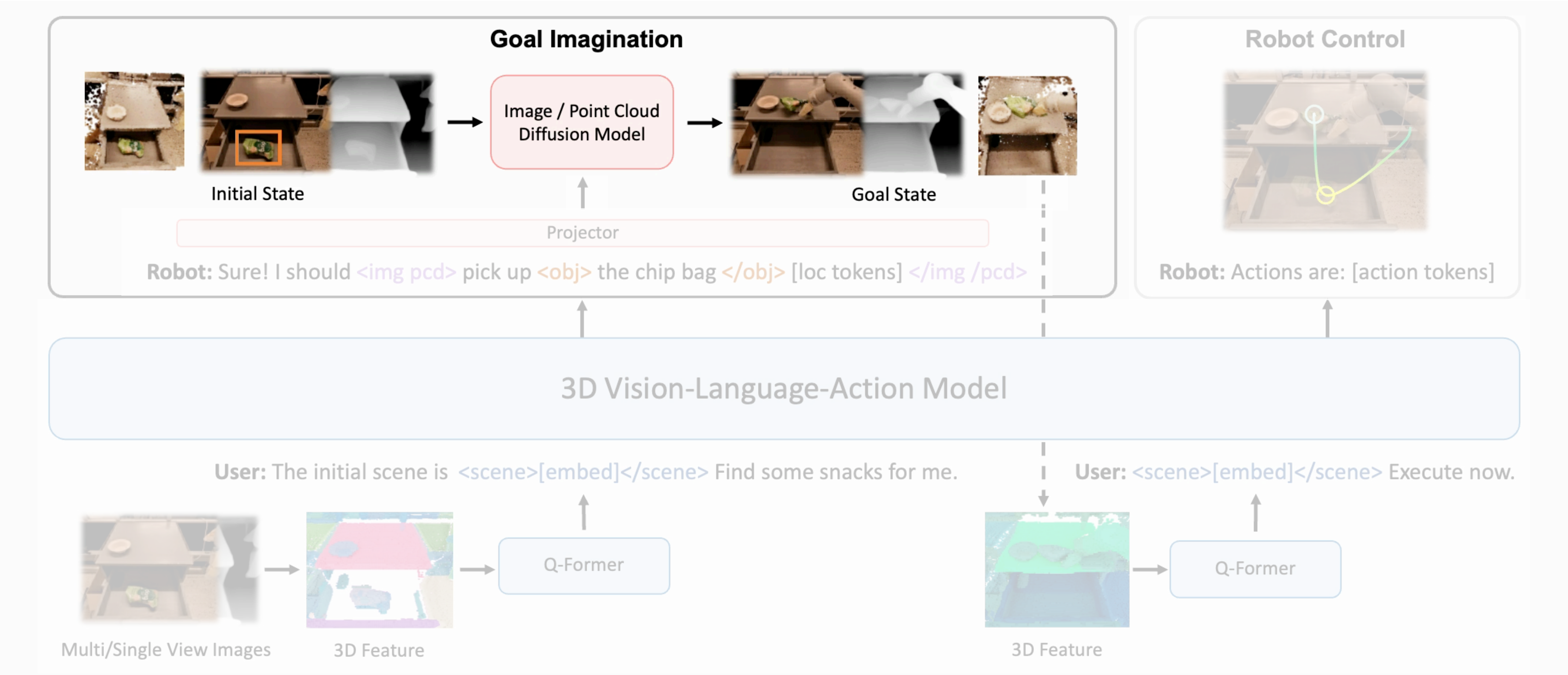
<obj> </obj>: to enclose the object nouns and followed by the location tokens.

[loc tokens], <loc0-255>: to locate objects

<img> </img> <pcd> </pcd>: as a signal to prompt the decoding side to output a certain modality

[action tokens]: to represent the 7 DoF state of a robotic arm

Training Stage

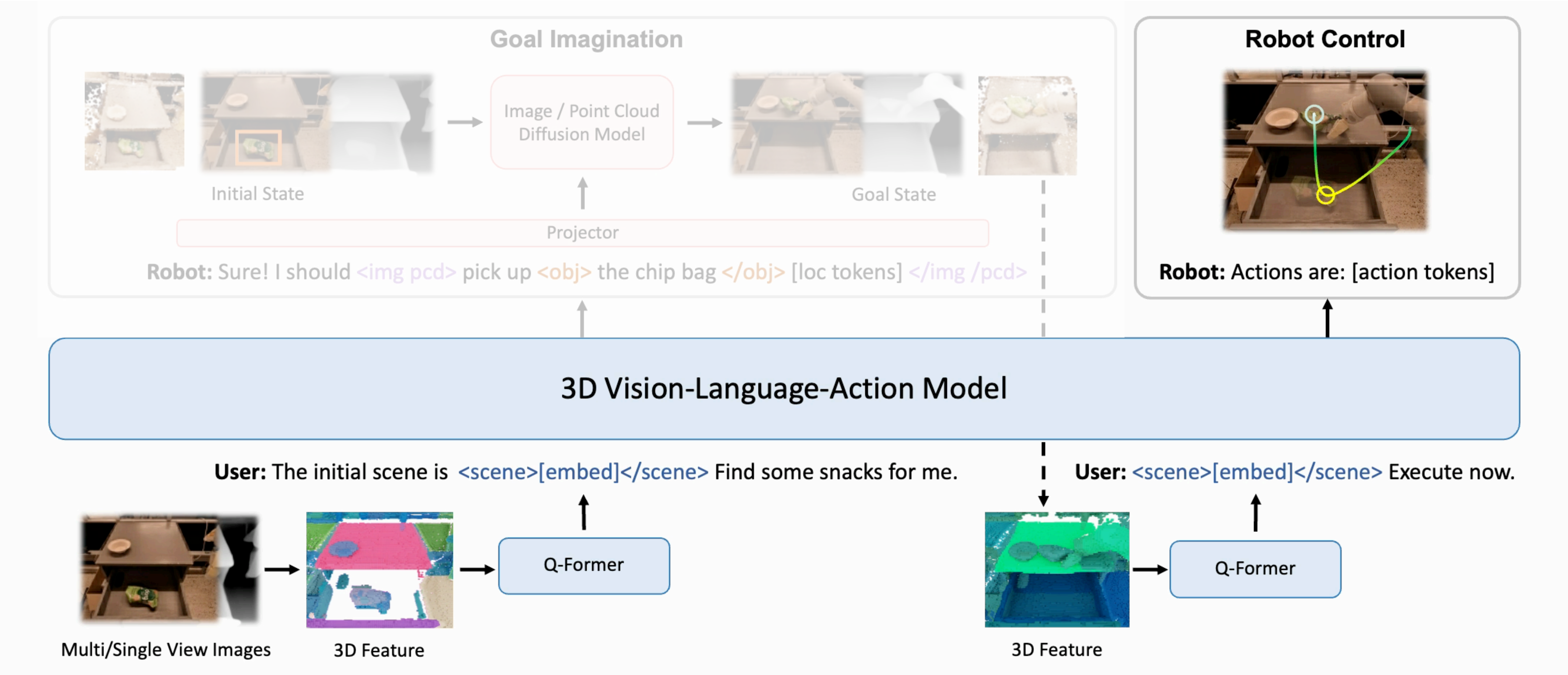


1. Embodied DM

2. Robotics LLM

3. Bridge LLM and DM

Training Stage

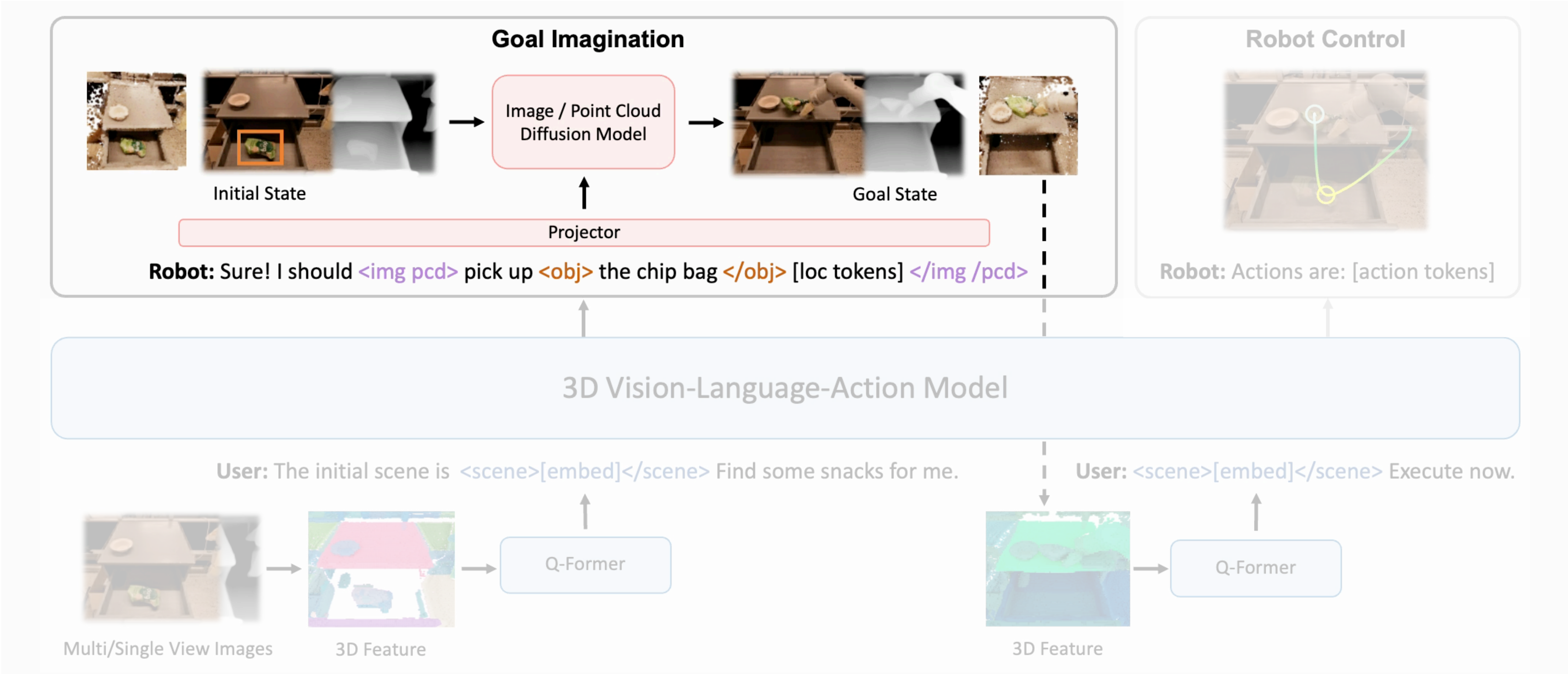


1. Embodied DM

2. Robotics LLM

3. Bridge LLM and DM

Training Stage



1. Embodied DM

2. Robotics LLM

3. Bridge LLM and DM

Language-related Tasks

Tasks	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH-L	EM@1
Embodied QA	3D-LLM*	1.05	0.38	0.15	0.02	12.96	0.91	0.00
	BLIP2 $OPT_{2.7B}$*	7.39	3.17	0.03	0.02	3.87	7.40	3.03
	BLIP2 $FlanT5_{XL}$*	22.84	16.17	12.50	10.11	11.41	32.01	10.31
	$OpenFlamingo_{4B}$*	9.50	6.51	5.14	4.29	6.84	10.40	1.21
	$LLaVA_{7B}$*	11.66	8.06	6.01	4.58	12.59	14.17	5.67
	BLIP2 $FlanT5_{XL}$	37.31	27.20	20.32	15.48	17.80	38.92	15.35
	3D-VLA	**48.34**	**38.55**	**31.72**	**26.80**	**23.72**	**49.33**	**24.53**
Task Caption	3D-LLM*	0.78	0.16	0.07	0.05	0.57	1.33	0.00
	BLIP2 $FlanT5_{XL}$*	8.50	2.07	0.35	0.00	3.40	8.45	0.00
	$OpenFlamingo_{4B}$*	7.61	1.64	0.37	0.00	4.74	9.36	0.00
	$LLaVA_{7B}$*	2.63	0.69	0.16	0.00	2.63	4.65	0.00
	BLIP2 $FlanT5_{XL}$	22.05	11.40	5.72	3.16	8.72	26.12	7.75
	3D-VLA	**55.69**	**45.88**	**39.39**	**34.88**	**27.57**	**62.01**	**29.34**
What-if QA	BLIP2 $FlanT5_{XL}$	28.23	11.47	4.49	0.06	8.27	28.41	5.85
	3D-VLA	**53.09**	**40.94**	**34.34**	**29.38**	**26.83**	**52.82**	**14.7**
Dense Caption	3D-LLM*	0.52	0.22	0.16	0.13	0.34	0.64	0.00
	BLIP2 $FlanT5_{XL}$	36.17	24.72	18.06	13.96	17.83	40.56	13.10
	3D-VLA	**51.90**	**42.83**	**38.11**	**34.62**	**25.25**	**55.91**	**39.49**

Table 1. Evaluation on reasoning ability using held-in data. $*$ denotes zero-shot transfer results without training on our pre-train datasets.

Ablation Studies on Generation Tasks

Method	PSNR ↑	CLIP Sim ↑	SSIM ↑	FID ↓
Instruct-P2P	14.41	0.909	0.389	0.309
SuSIE	15.20	0.898	0.549	0.182
NeXT-GPT	8.86	0.199	0.153	0.432
Instruct-P2P*	16.67	**0.941**	0.628	0.178
3D-VLA w/o Pred BBox	17.02	0.919	0.632	**0.173**
3D-VLA	**17.21**	0.920	**0.636**	0.177

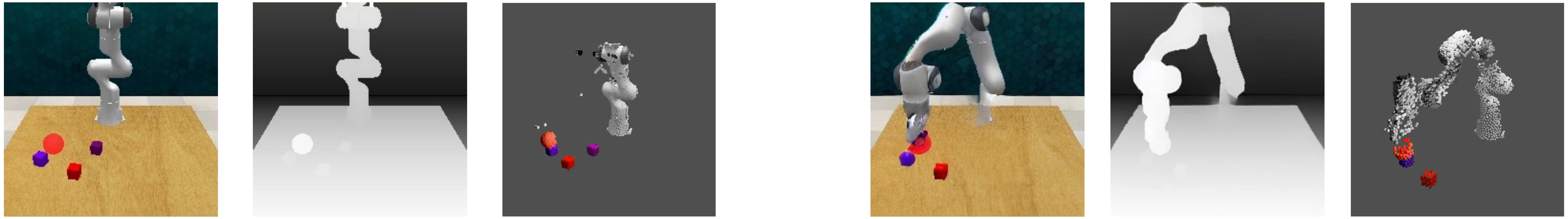
Table 3. RGB image goal generation results. $*$ denotes the model is trained on our pretrained dataset.

Models	P-FID ↓	Chamfer-L_1 ↓
Point-E*	5.241	0.159
3D-VLA w/o Pred BBox	4.914	0.143
3D-VLA	**4.796**	**0.139**

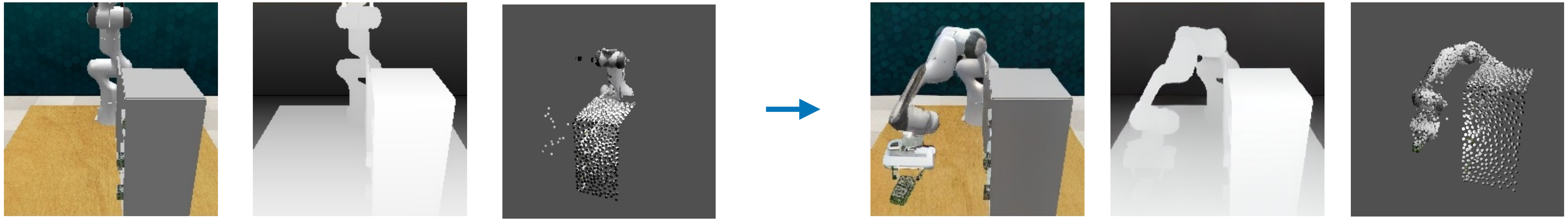
Table 4. Point Cloud goal generation results. $*$ denotes the model is trained on our pretrained dataset.

Goal Image, Depth and Point Cloud Generation on RLBench

Grasping the purple block to the target



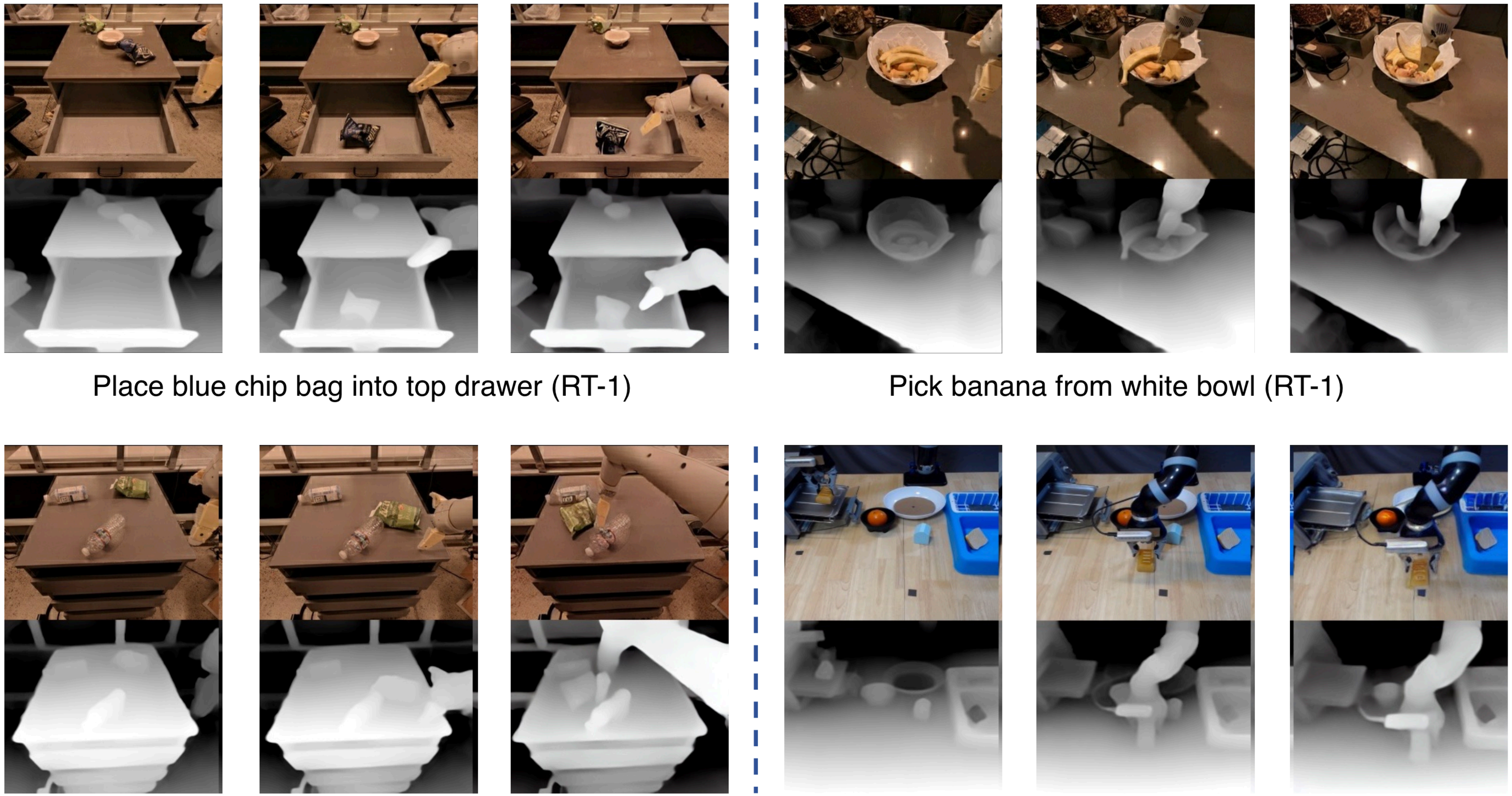
Taking the stack of money and placing it on the table



Picking up the red cup

Goal Generation on Real-world Scenes

Input | GT Goal | Pred Goal | Input | GT Goal | Pred Goal



Place blue chip bag into top drawer (RT-1)

Pick banana from white bowl (RT-1)

Move green chip bag near water bottle (RT-1)

Place the long bread on the table (Jaco Play)

Possible Emergent Abilities

Zero-shot Results

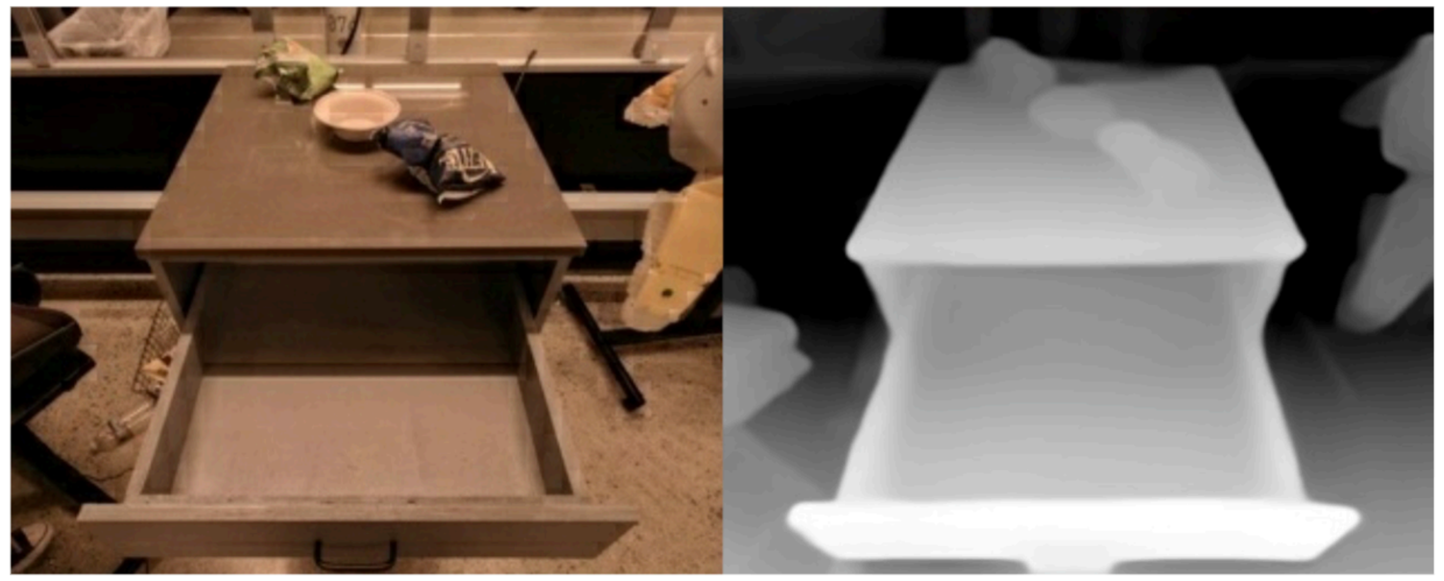
Generated by DALL-E 3

Pick up the plate

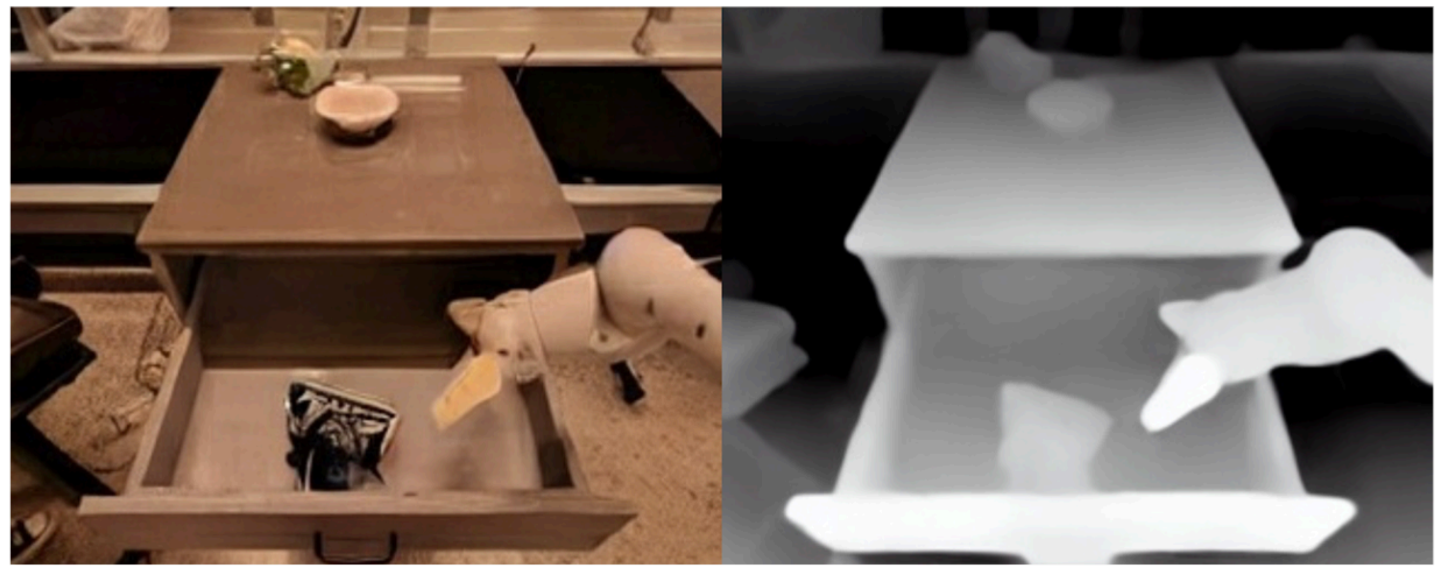
Open the drawer

Pick up the bottle and place on the plate

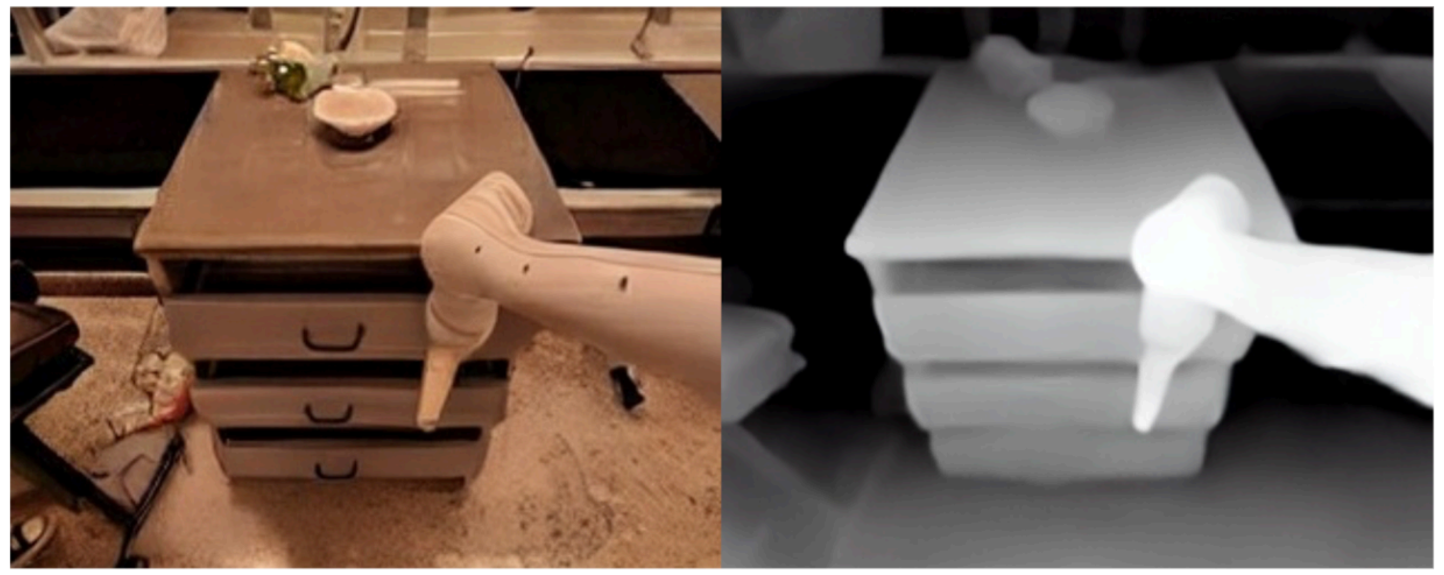
Long Horizon Task



Initial Image

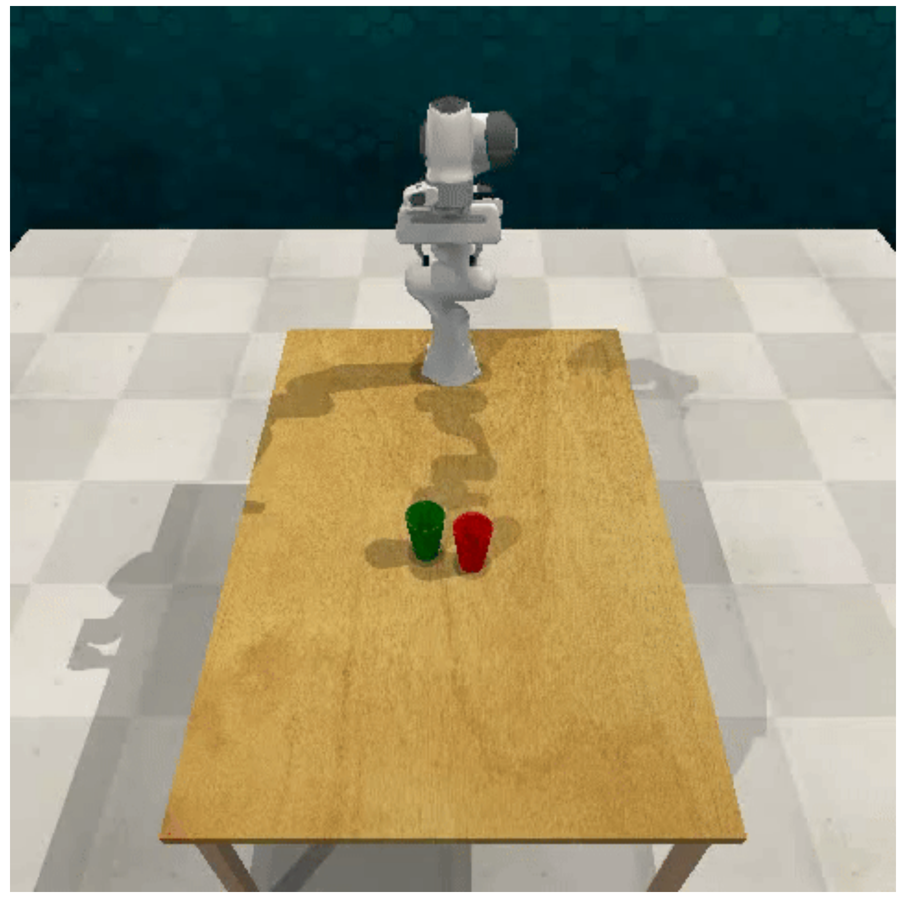


First place chip bag into top drawer

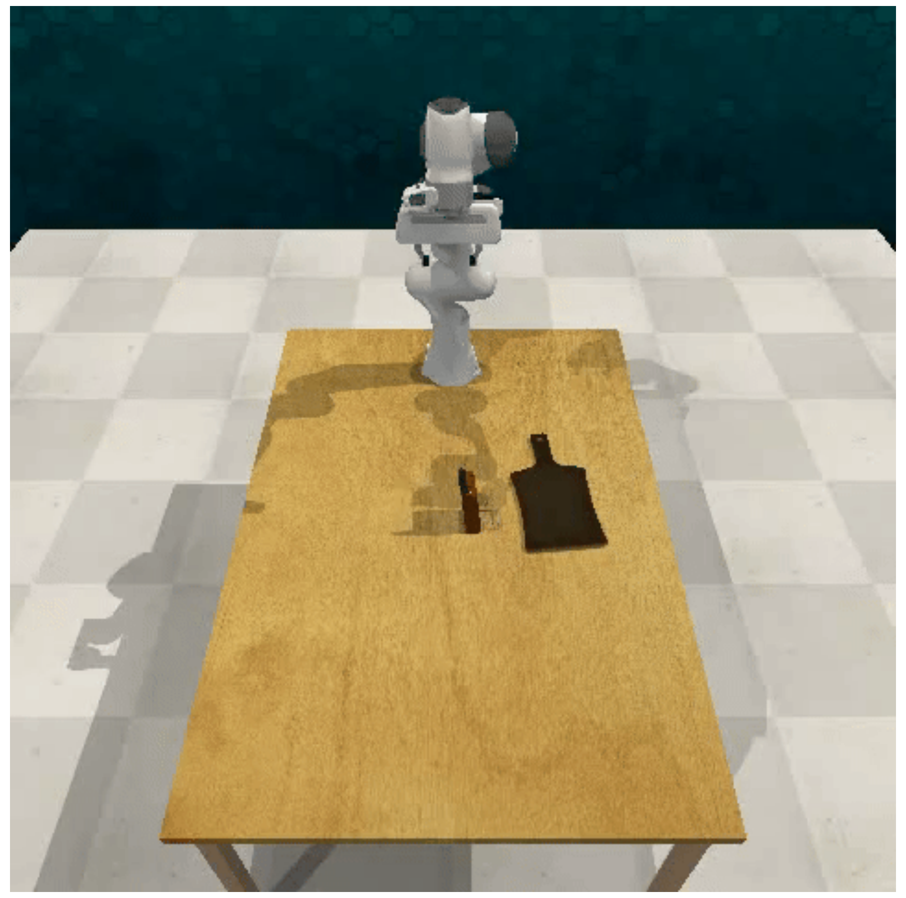


Then close the drawer

Manipulation Results

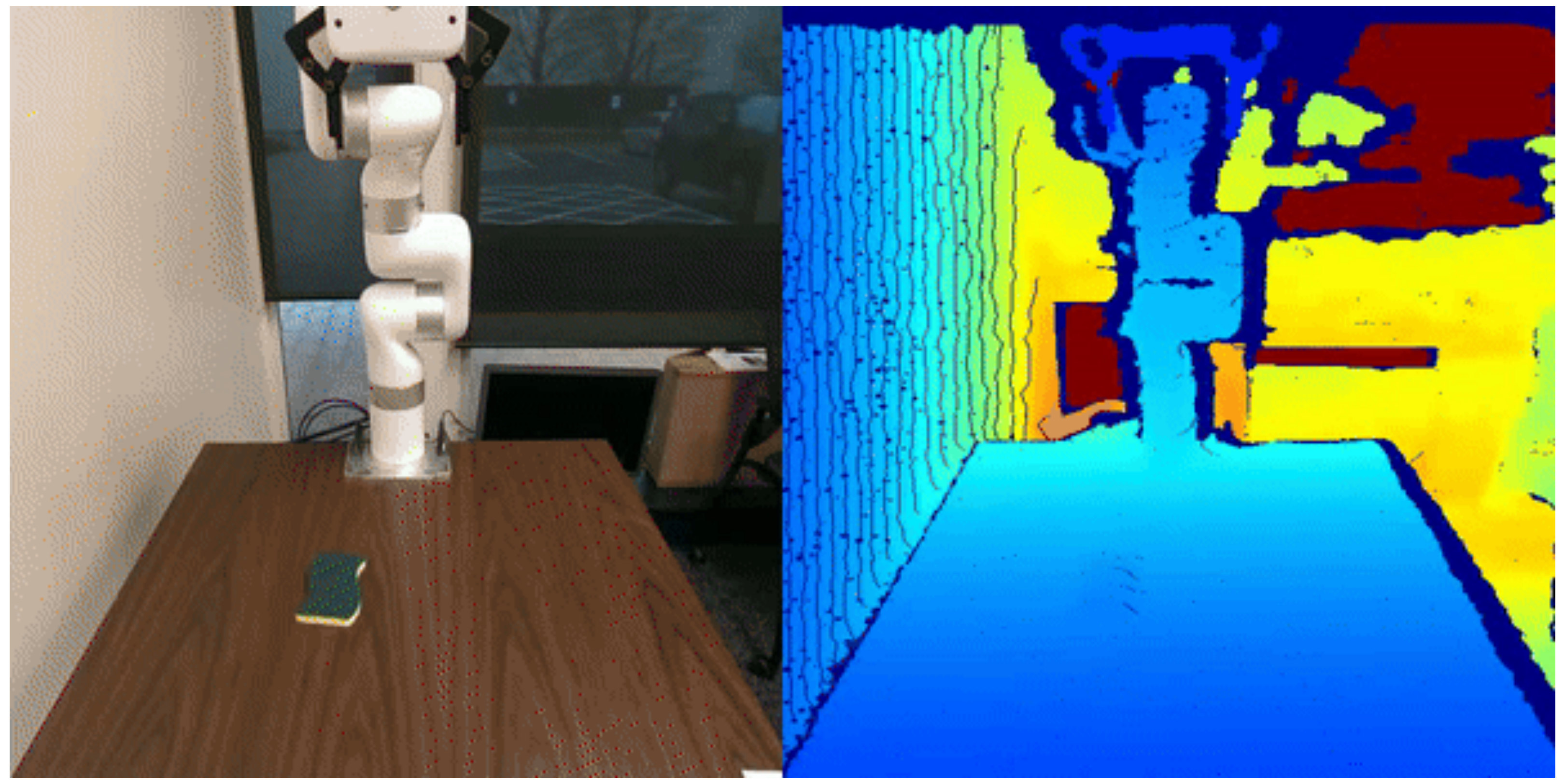


Pick up the green cup

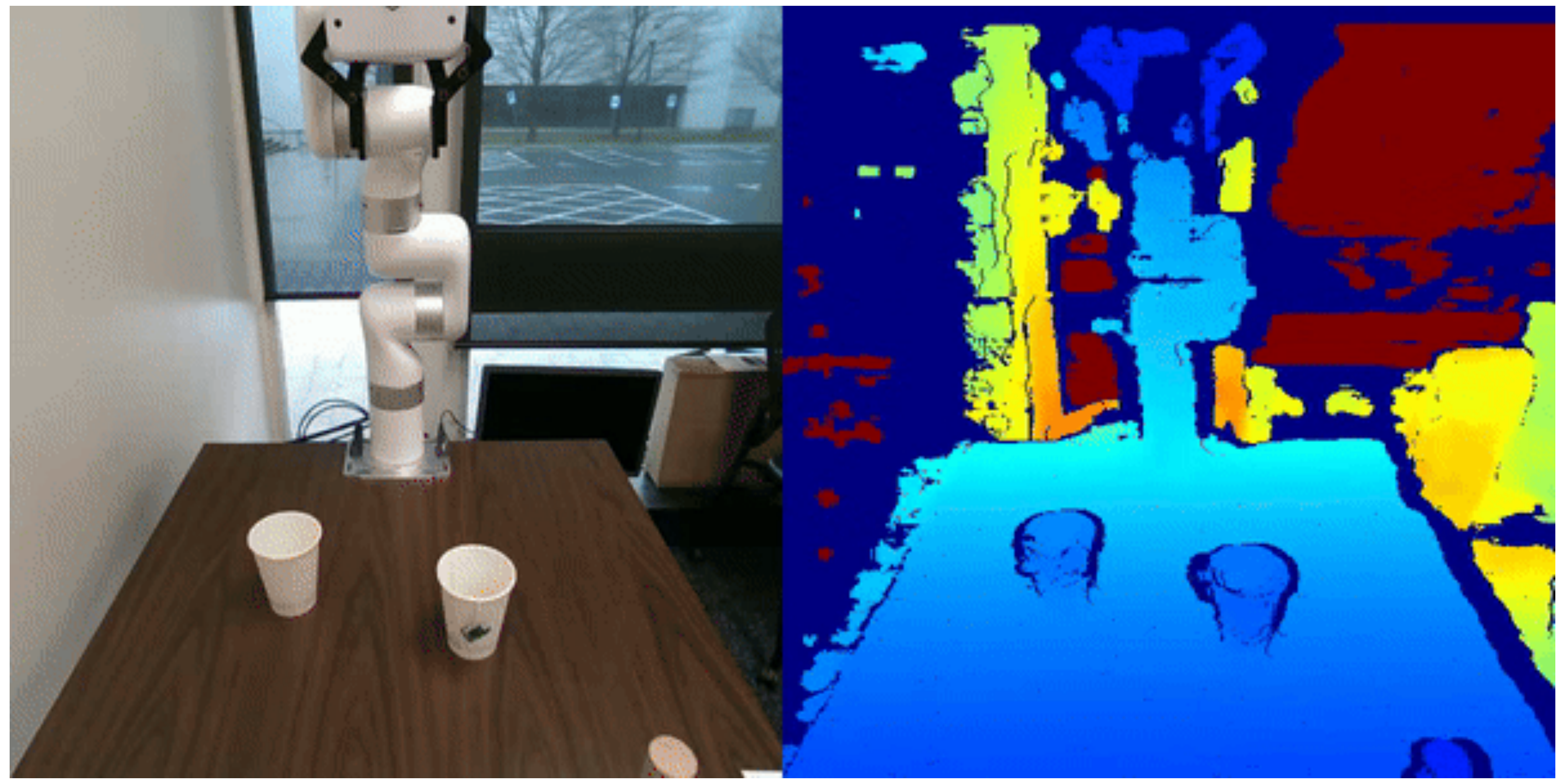


Put knife on chopping board

Take umbrella



Swipe the table



Pick up the cup

Can goal generation guide the better execution of other tasks?

Pretrained	Goal Gen	B-1	B-2	B-3	B-4	M	R	EM
✗	✗	42.4	30.9	25.3	21.1	22.2	45.4	6.8
✗	✓	42.7	31.0	25.2	20.6	22.0	45.6	7.9
✓	✗	43.9	32.7	26.3	22.1	22.4	42.0	9.2
✓	✓	48.6	37.5	31.2	26.9	24.1	46.2	12.0

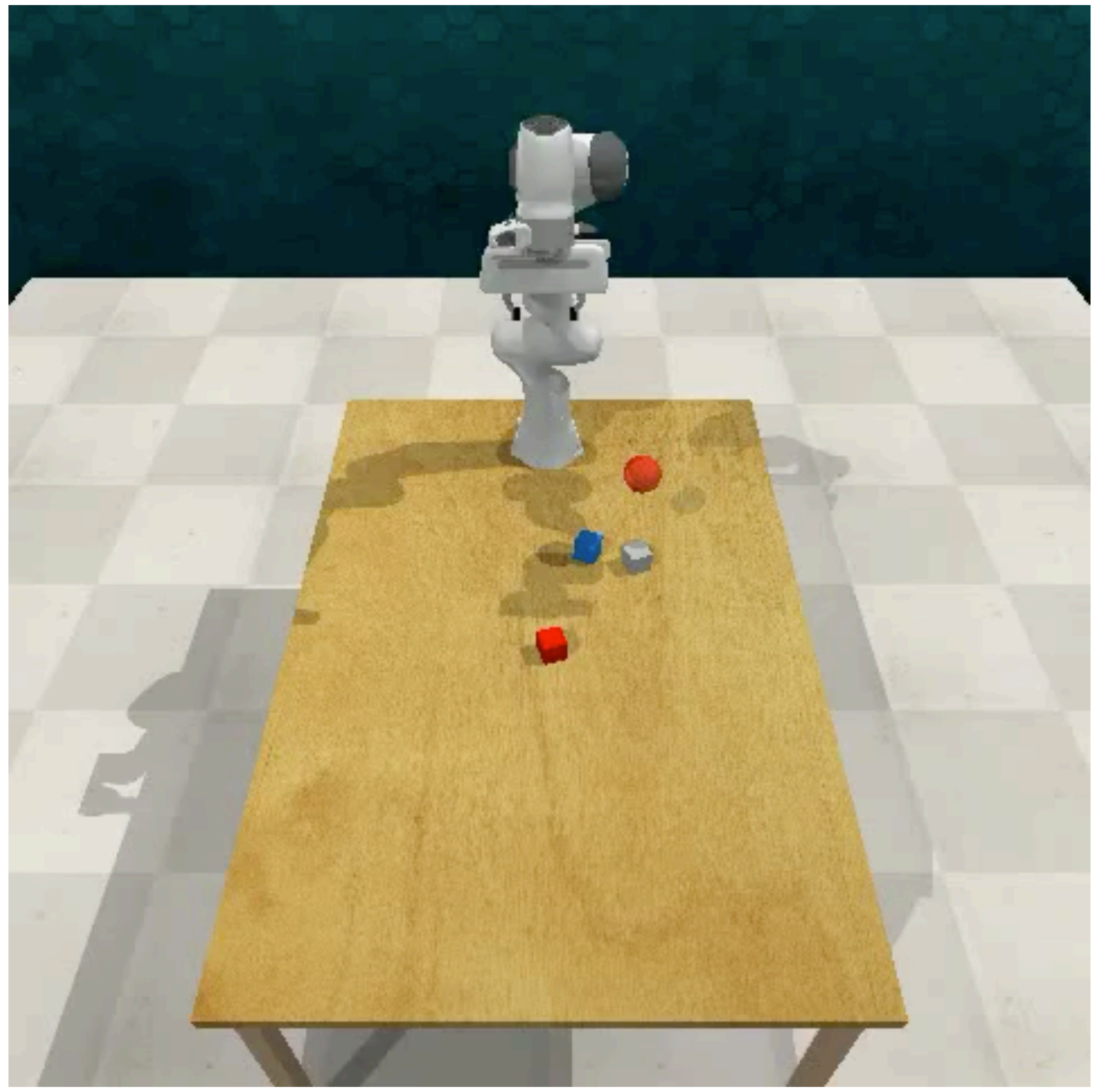
What-if QA

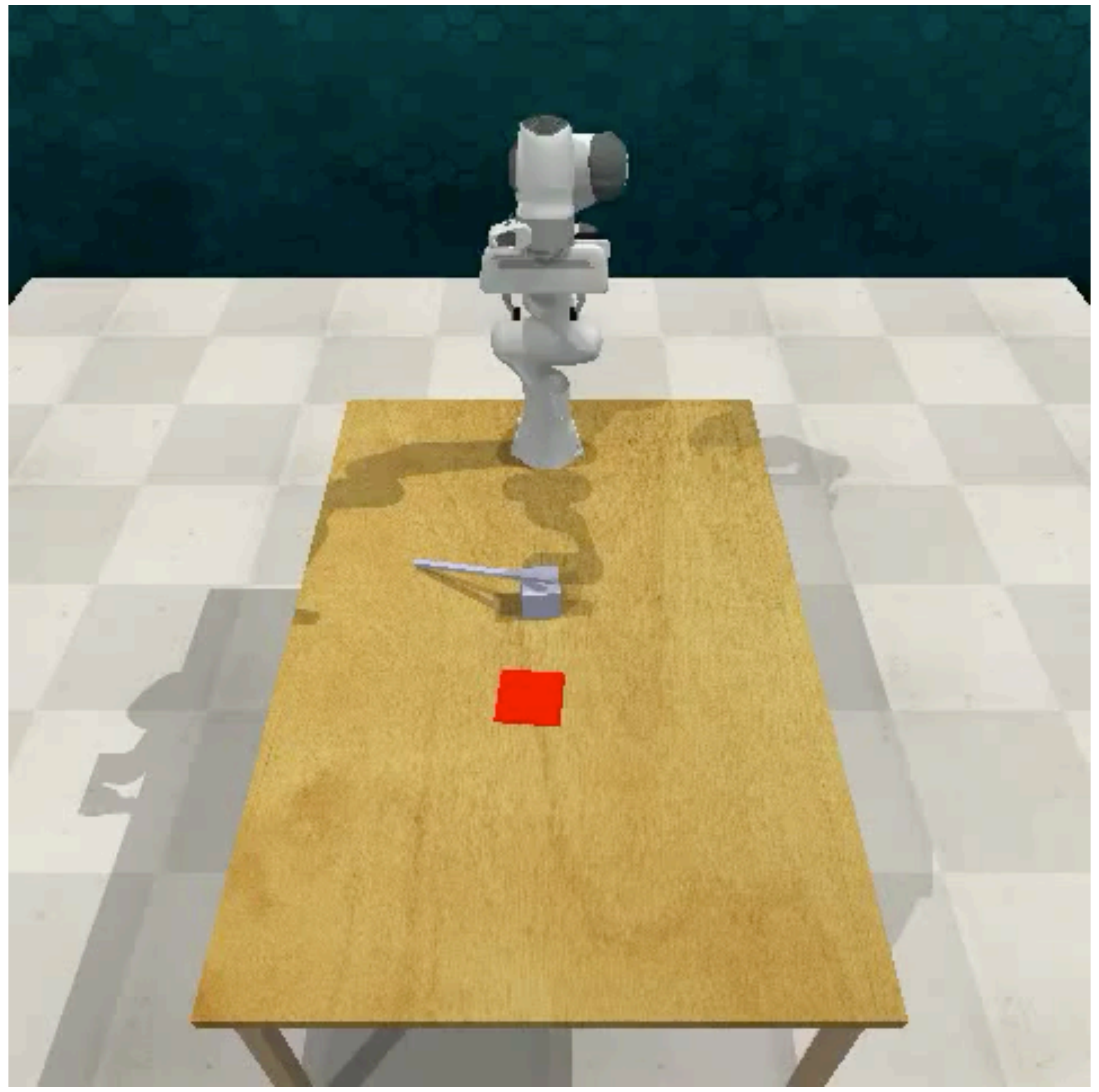
Pretrained	Goal Gen	Put Knife	Take Um	Cup
✗	✗	52	62	28
✗	✓	56	62	24
✓	✗	68	58	34
✓	✓	68	80	40

Manipulation

Limitations

1. Difficulty in **precise** control



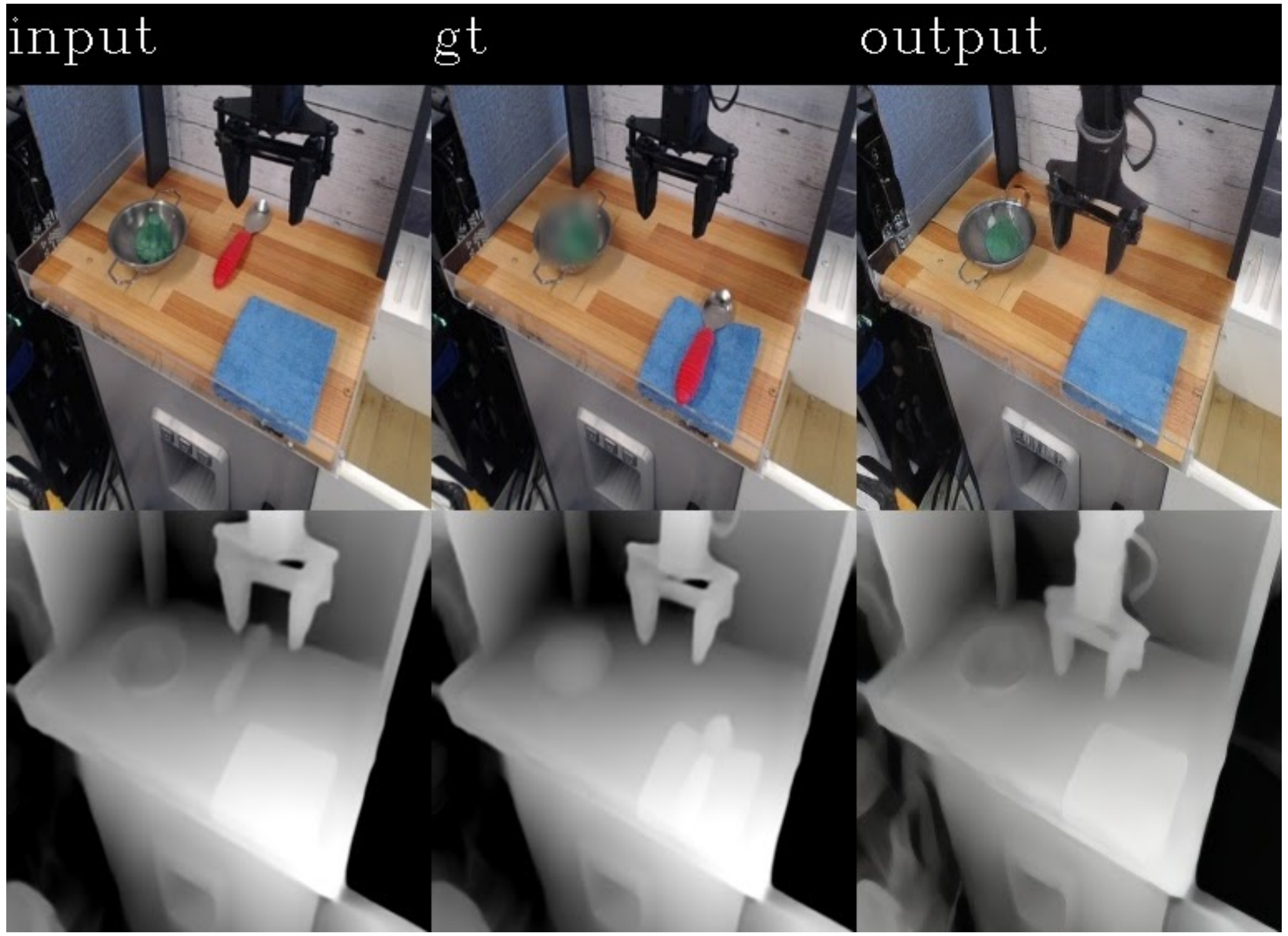


Limitations

1. Difficulty in **precise** control

2. **Hallucination** of the diffusion model

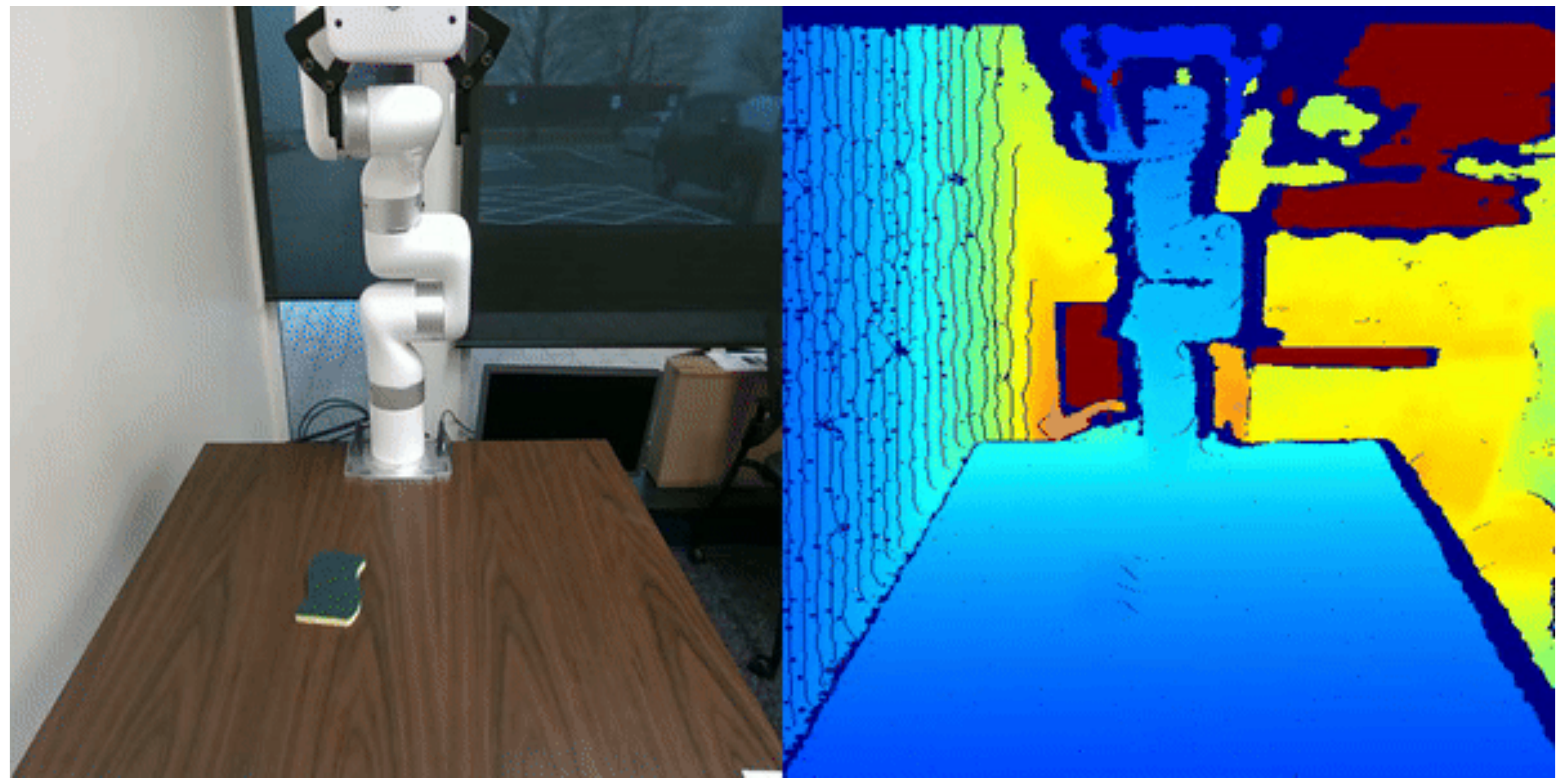
Move spoon on to blue towel



Where is spoon?

Limitations

1. Difficulty in **precise** control
2. **Hallucination** of the diffusion model
3. Issues with depth and point clouds in the real world



Limitations

1. Difficulty in **precise** control
2. **Hallucination** of the diffusion model
3. Issues with depth and point clouds in the real world
4. The **long-tail** distribution
5. Datasets with high variance in quality

[1] BC_Z dataset, Google

Future Works

Humanoid + Mobile Robot + Real World + Video Diffusion + Agent ……

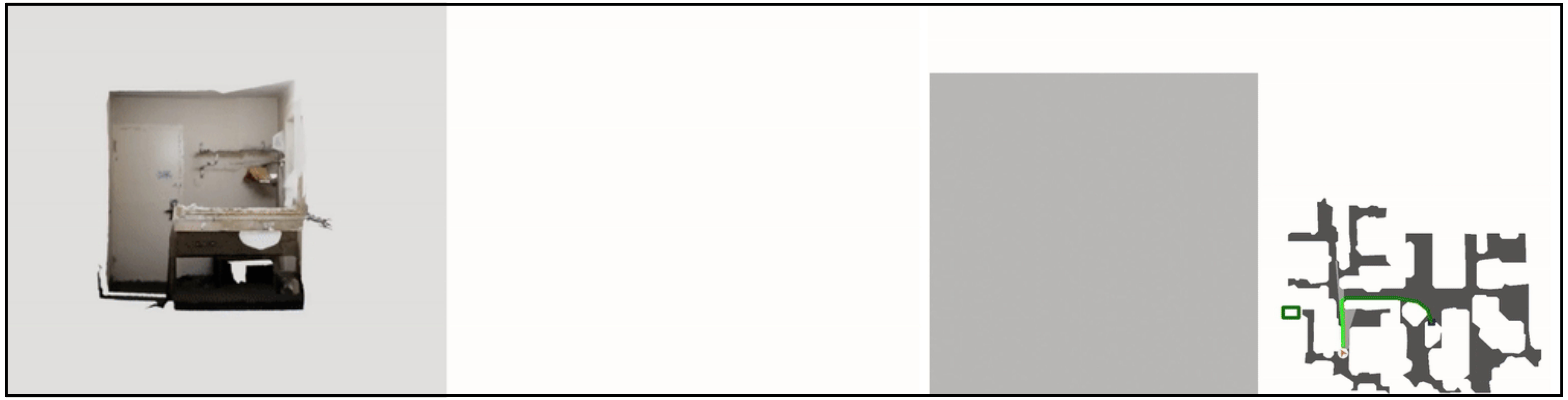
Acknowledgement

Some material in the slides is borrowed from Yining Hong.

Q&A / Discussion

Reasoning and Planning with Embodied Foundation Models

3D-LLM, NeurIPS 2023 Spotlight



Bridging Interaction and Dynamics with Generative World Model

3D-VLA, ICML 2024