北京大學 PEKING UNIVERSITY

BAAI 智源研究院

MobileVLA-R1: Reinforcing Vision-Language-Action for Mobile Robots

Presenter: Ting Huang

Supervisor: Hao Tang

目录

北京大學
PEKING UNIVERSITY

PART 1

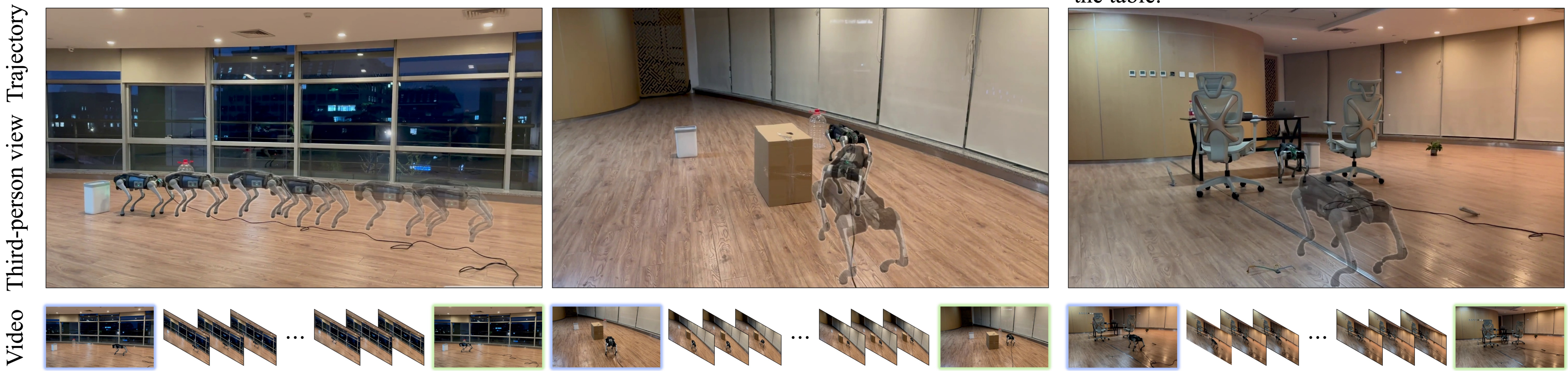
Background & Motivation

Background: Vision-Language-Action for Mobile Robots

✓ **Vision-Language-Action (VLA)** aims to translate natural-language instructions into grounded perception, reasoning, and executable robot actions.

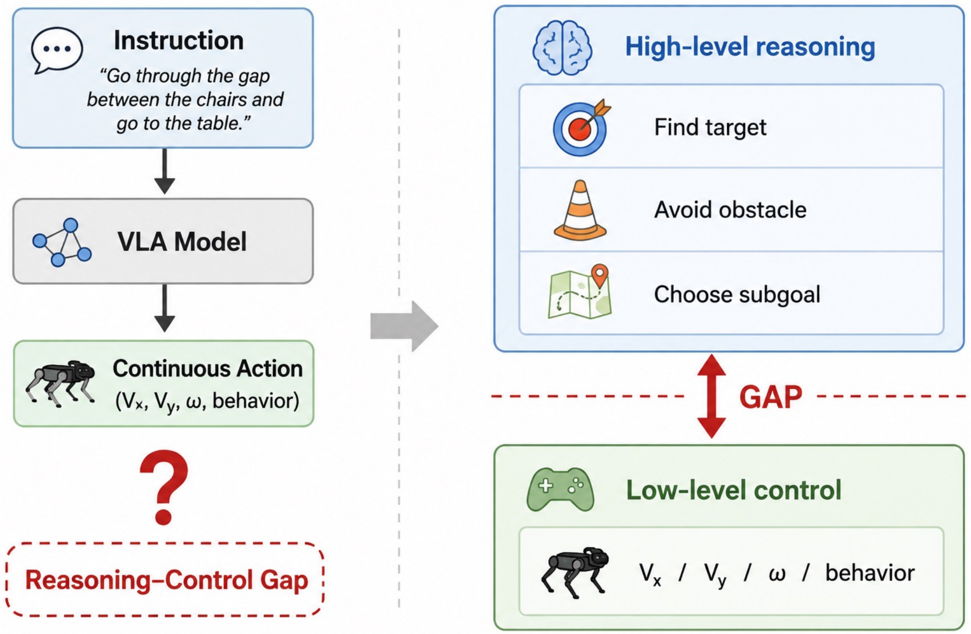
✓ Embodied VLA is a fundamental capability for intelligent mobile robots operating in real-world environments.

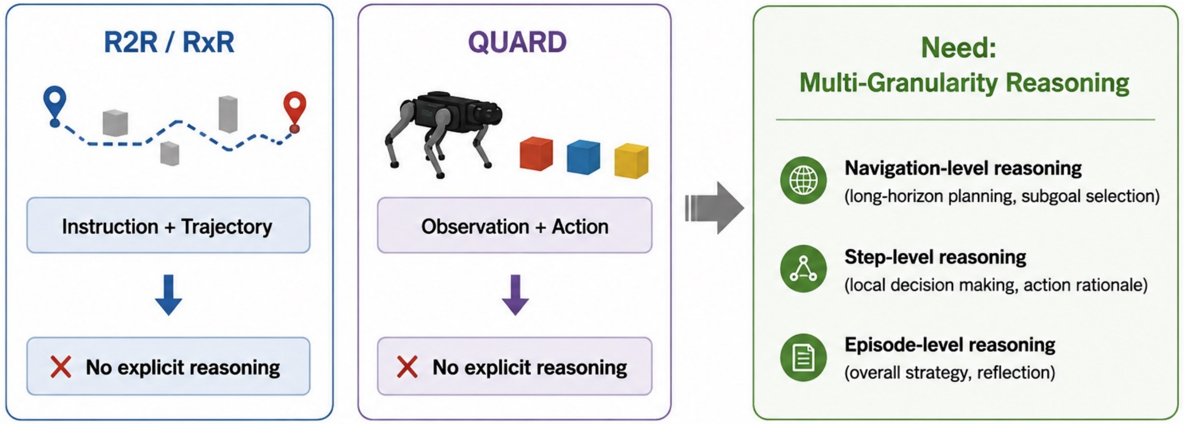
Go to the trashcan and distinguish the water barrel. Go to the water barrel and avoid the box. Go through the gap between the chairs and go to the table.



Motivation

- Existing VLA models often directly map language and observations to actions, making high-level reasoning and low-level motor control **difficult to disentangle**.
- Existing embodied datasets mainly provide navigation trajectories or control targets, but **lack explicit intermediate reasoning supervision**.





We need a VLA model that can **explicitly** reason before acting, and **align** reasoning with executable control.

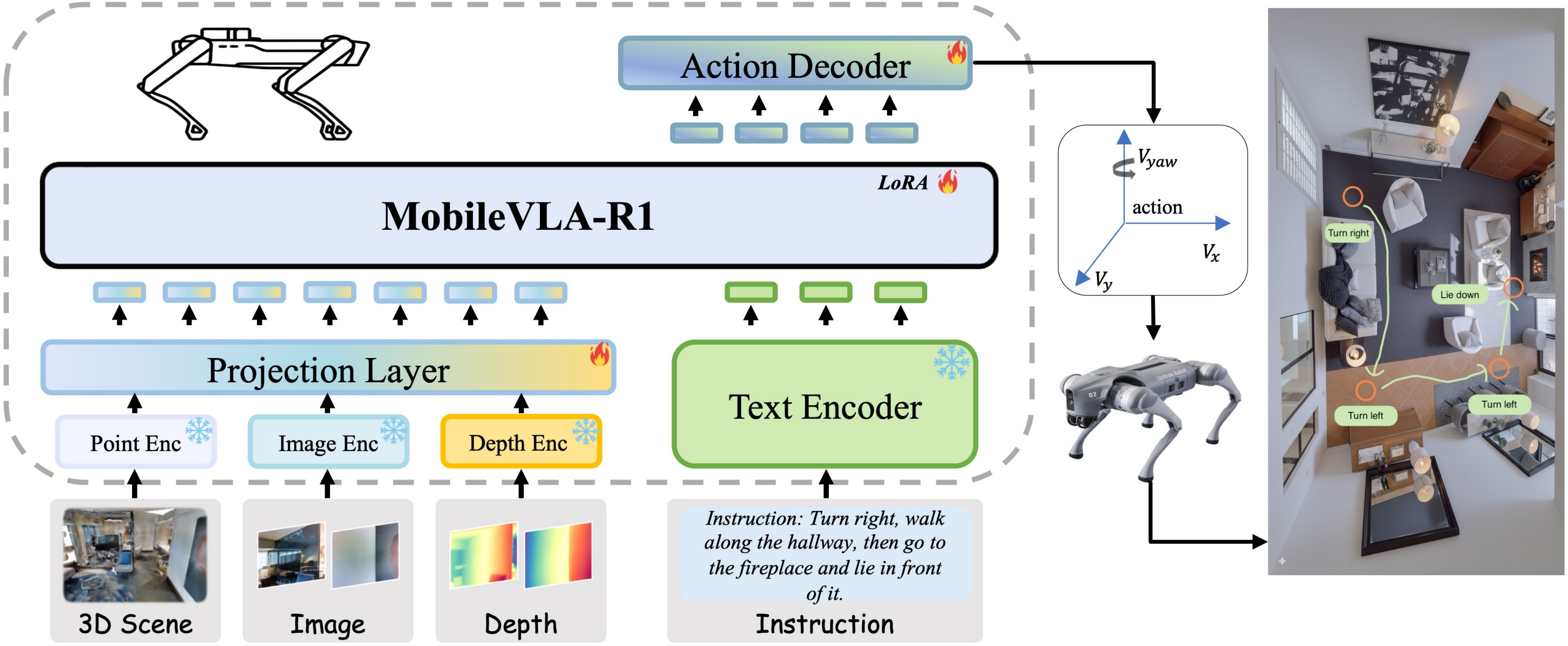
北京大学
PEKING UNIVERSITY

PART 2

Method

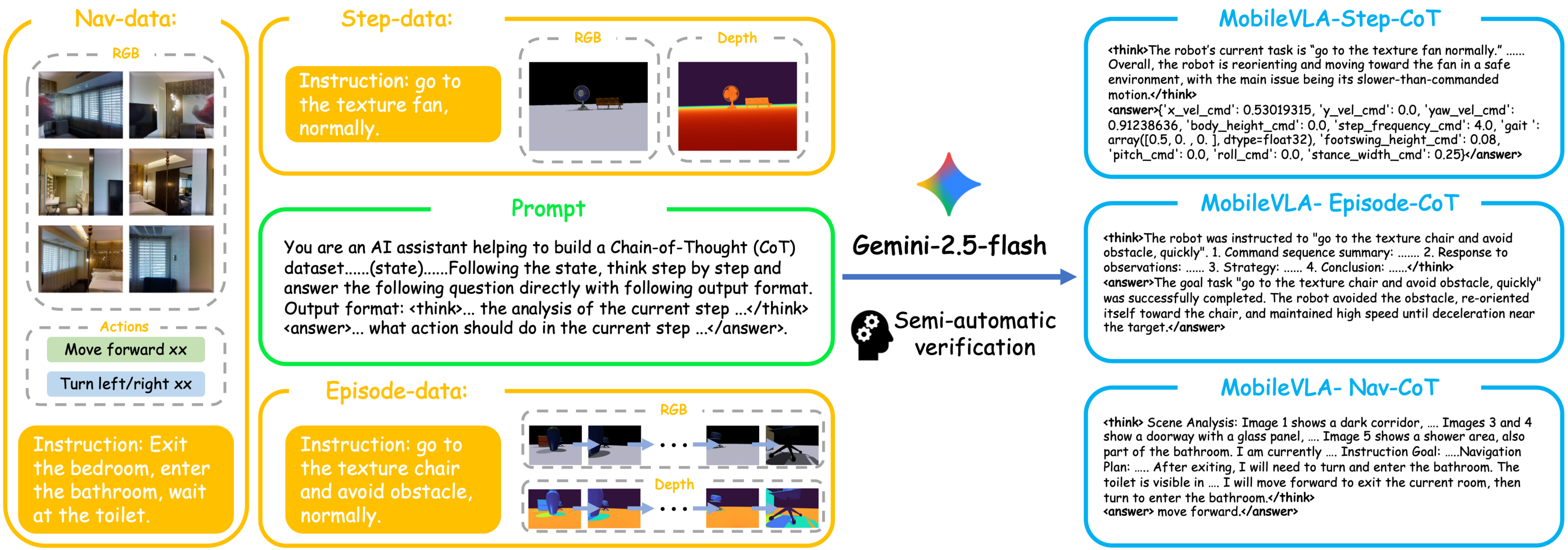
MobileVLA-R1: Overall Framework

Multimodal Perception → Structured Reasoning → Executable Control



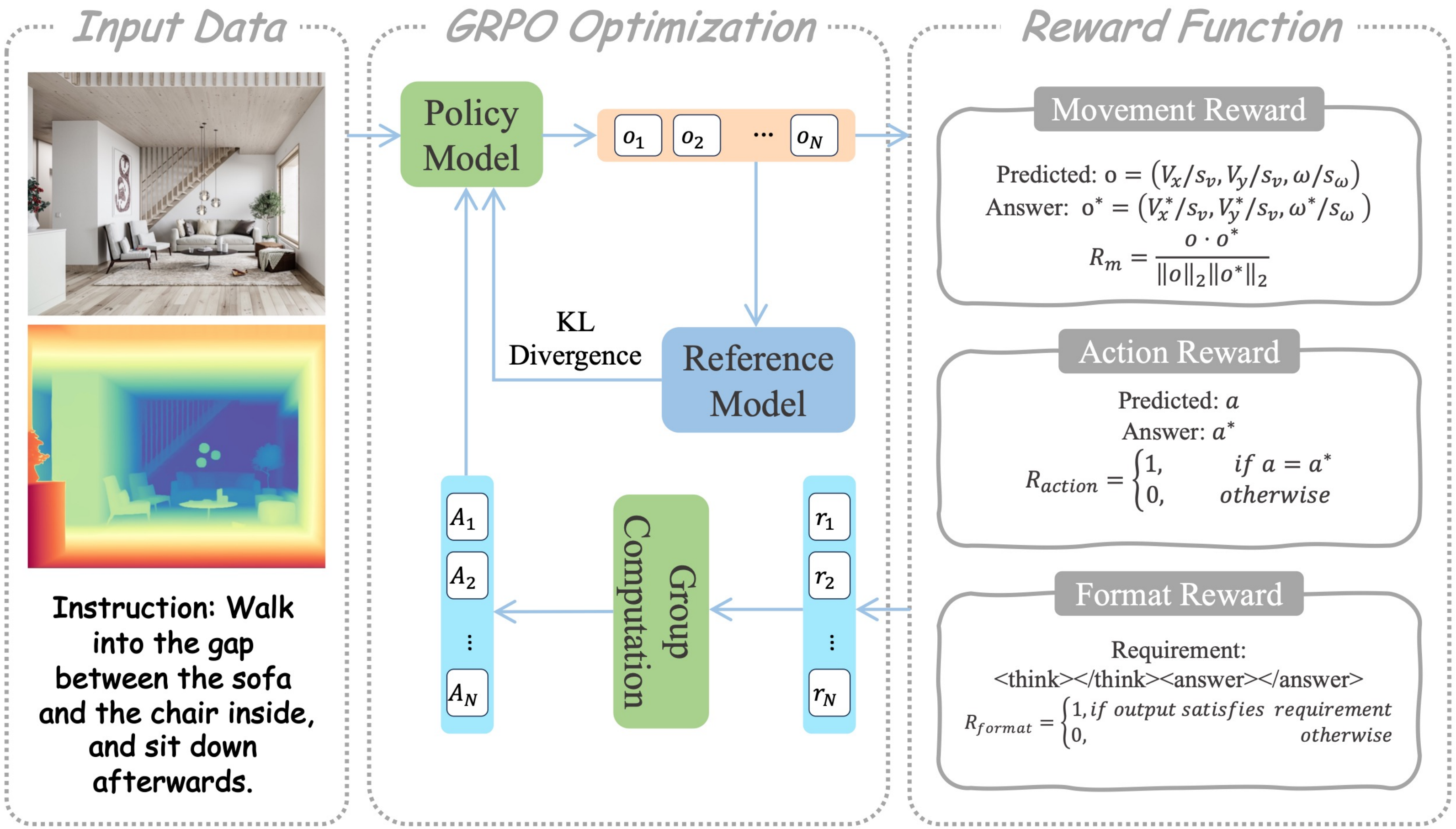
MobileVLA-CoT: Multi-Granularity Reasoning Supervision

Synthesize reasoning traces at multiple temporal scales to bridge language instructions and executable actions.



GRPO for Reasoning-to-Control Alignment

Optimize structured reasoning outputs with action-level rewards for stable and executable robot control.



北京大學
PEKING UNIVERSITY

PART 3

Experiments

Main Results: Navigation & Quadruped Control

Method	Observation				R2R-CE Val-Unseen				RxR-CE Val-Unseen			
	S.RGB	Pano.	Depth	Odom.	NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑	nDTW↑
CMA* [21]	✗	✓	✓	✓	6.20	52.0	41.0	36.0	8.76	26.5	22.1	47.0
Sim2Sim* [31]	✗	✓	✓	✓	6.07	52.0	43.0	36.0	-	-	-	-
GridMM* [56]	✗	✓	✓	✓	5.11	61.0	49.0	41.0	-	-	-	-
Ego^2-Map* [23]	✗	✓	✓	✓	5.54	56.0	47.0	41.0	-	-	-	-
DreamWalker* [52]	✗	✓	✓	✓	5.53	59.0	49.0	44.0	-	-	-	-
Reborn* [2]	✗	✓	✓	✓	5.40	57.0	50.0	46.0	5.98	48.6	42.0	63.3
ETPNav* [1]	✗	✓	✓	✓	4.71	65.0	57.0	49.0	5.64	54.7	44.8	61.9
HNR* [55]	✗	✓	✓	✓	4.42	67.0	61.0	51.0	5.50	56.3	46.7	63.5
AG-CMTP [8]	✗	✓	✓	✓	7.90	39.0	23.0	19.0	-	-	-	-
R2R-CMTP [8]	✗	✓	✓	✓	7.90	38.0	26.0	22.0	-	-	-	-
InstructNav [36]	✗	✓	✓	✓	6.89	-	31.0	24.0	-	-	-	-
LAW [44]	✓	✗	✓	✓	6.83	44.0	35.0	31.0	10.90	8.0	8.0	38.0
CM2 [18]	✓	✗	✓	✓	7.02	41.0	34.0	27.0	-	-	-	-
WS-MGMap [9]	✓	✗	✓	✓	6.28	47.0	38.0	34.0	-	-	-	-
AO-Planner [7]	✗	✓	✓	✗	5.55	59.0	47.0	33.0	7.06	43.3	30.5	50.1
Seq2Seq [32]	✓	✗	✓	✗	7.77	37.0	25.0	22.0	12.10	13.9	11.9	30.8
CMA [32]	✓	✗	✓	✗	7.37	40.0	32.0	30.0	-	-	-	-
NaVid [65]	✓	✗	✗	✗	5.47	49.0	37.0	35.0	-	-	-	-
Uni-NaVid [64]	✓	✗	✗	✗	5.58	53.5	47.0	42.7	6.24	48.7	40.9	-
NaVILA [12]	✓	✗	✗	✗	5.22	62.5	54.0	49.0	6.77	49.3	44.0	58.8
VLN-R1 [41]	✓	✗	✗	✗	7.00	41.2	30.2	21.8	9.10	22.7	17.6	-
OctoNav [17]	✓	✗	✗	✗	-	42.9	37.1	33.6	-	-	-	-
StreamVLN [58]	✓	✗	✗	✗	4.98	64.2	56.9	51.9	6.22	52.9	46.0	61.9
CorrectNav [63]	✓	✗	✗	✗	4.24	67.5	65.1	62.3	4.09	69.3	63.3	75.2
MobileVLA-R1 (Ours)	✓	✗	✓	✗	**4.05**	**69.7**	**68.3**	**65.2**	**3.92**	**71.5**	**66.8**	**76.1**

VLN-CE: Vision language navigation results

Method	Easy	Medium	Hard				Average
	Distinguish	Go-to	Go-avoid	Go-through	Crawl	Unload	
CLIP [42]	0.44	0.43	0.45	0.19	0.00	0.00	0.25
VC-1 [38]	0.46	0.43	0.45	0.31	0.00	0.00	0.28
QUART [14]	0.66	0.60	0.53	0.41	0.32	0.12	0.44
MoRE [67]	0.82	0.80	0.59	0.57	0.49	0.33	0.60
MobileVLA-R1	**0.92**	**0.89**	**0.71**	**0.65**	**0.58**	**0.44**	**0.73**

QUARD: Quadruped control results

Navigation — R2R-CE

68.3 SR **65.2 SPL**

Navigation — RxR-CE

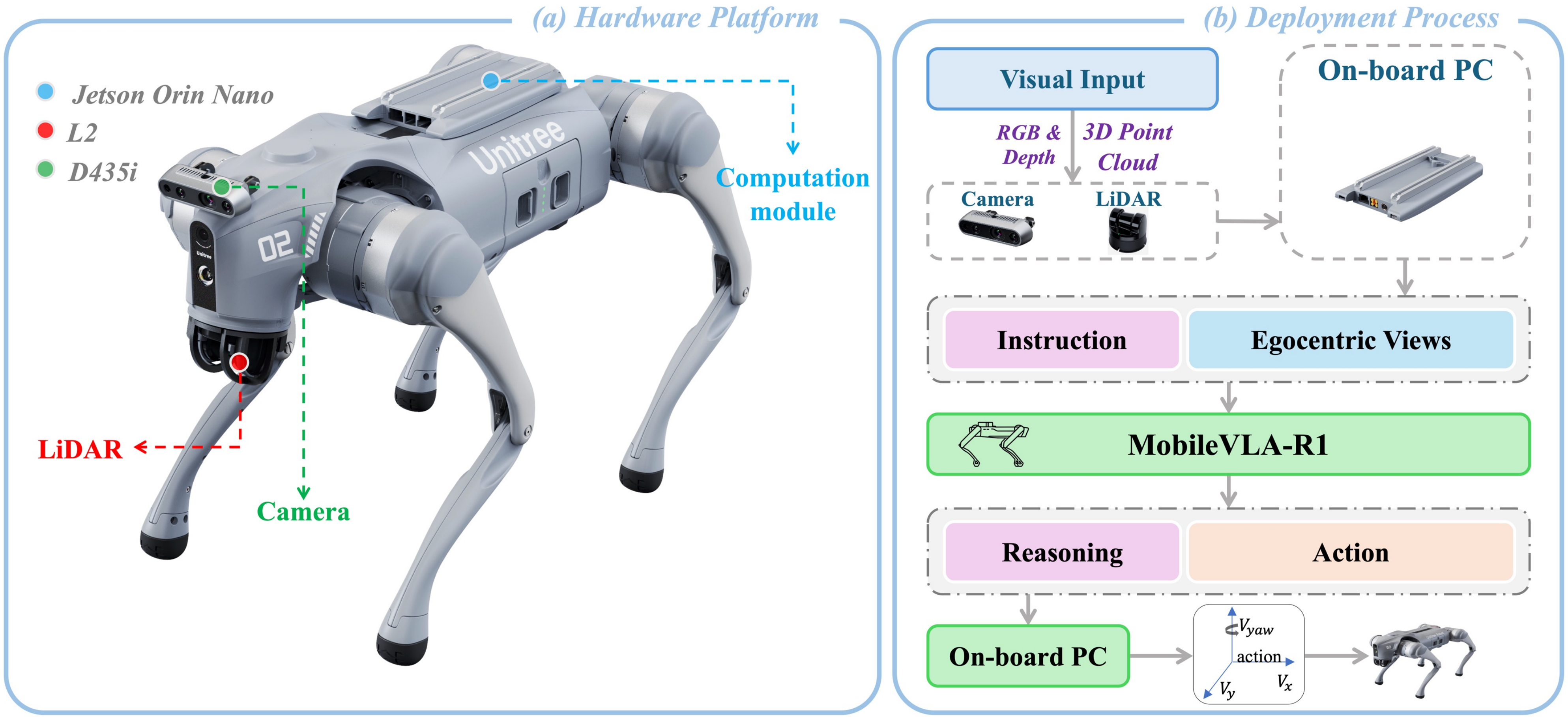
71.5 SR **66.8 SPL**

Quadruped Control — QUARD

0.73 Avg. Success

MobileVLA-R1 achieves strong performance across both navigation and executable quadruped control.

Real-World Deployment



Hybrid Deployment
On-board: sensing, mapping, action decoding
Remote H20: **8B MobileVLA-R1 inference**

Real-Time Control
205–245 ms end-to-end latency
5 Hz high-level control frequency

Real-World Evaluation
32 tasks · 160 episodes
Workspace · Corridor · Outdoor

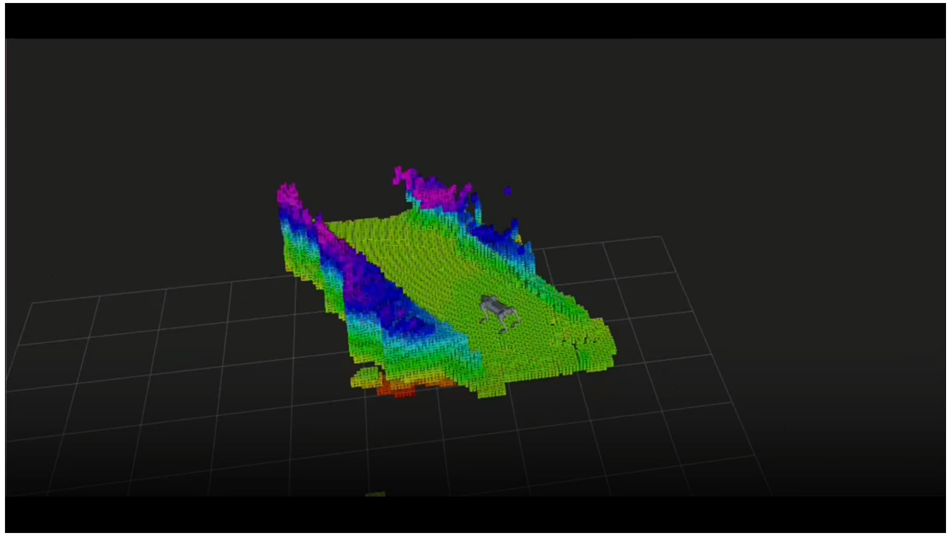
Real-World Results

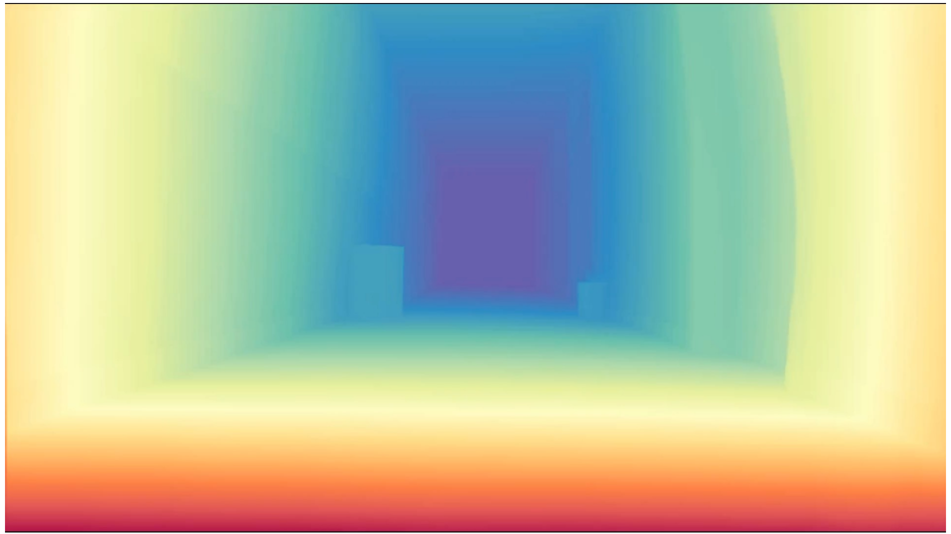
Method	Workspace				Corridor				Outdoor			
	Simple		Complex		Simple		Complex		Simple		Complex	
	NE↓	SR↑	NE↓	SR↑	NE↓	SR↑	NE↓	SR↑	NE	SR↑	NE	SR↑
GPT-4o [29]	2.01	0.67	2.38	0.33	1.49	0.53	3.00	0.00	–	0.67	–	0.50
NaVILA [12]	1.29	0.83	1.76	0.80	1.15	0.89	1.76	0.67	–	0.91	–	0.83
MobileVLA-R1	**1.03**	**0.93**	**1.23**	**0.91**	**0.96**	**1.00**	**1.23**	**0.86**	–	**1.00**	–	**0.96**

Consistent gains across all real-world environments, especially under complex long-horizon instructions.

Qualitative Real World Result——Indoor

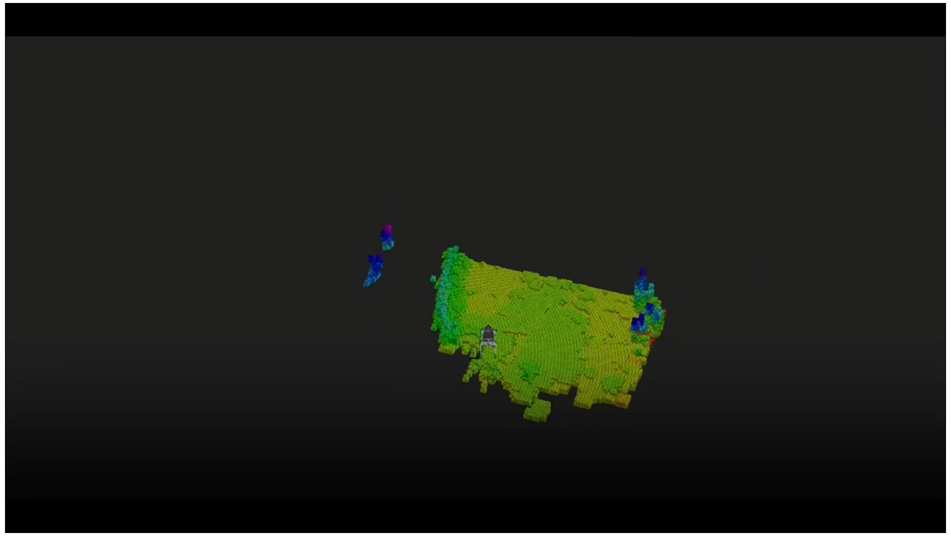
Starting from the initial position, walk forward to find a cardboard box, stop in front of it, and lie down.

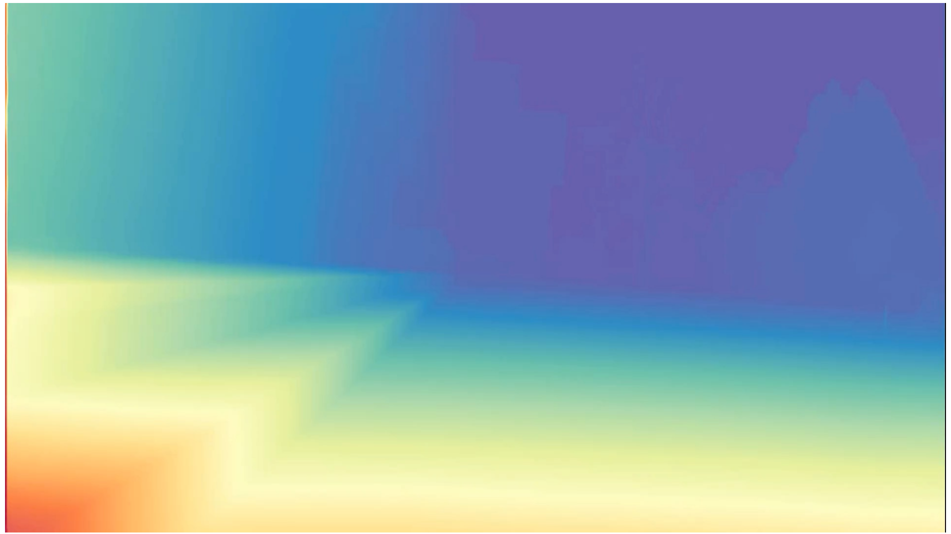




Qualitative Real World Result——Outdoor

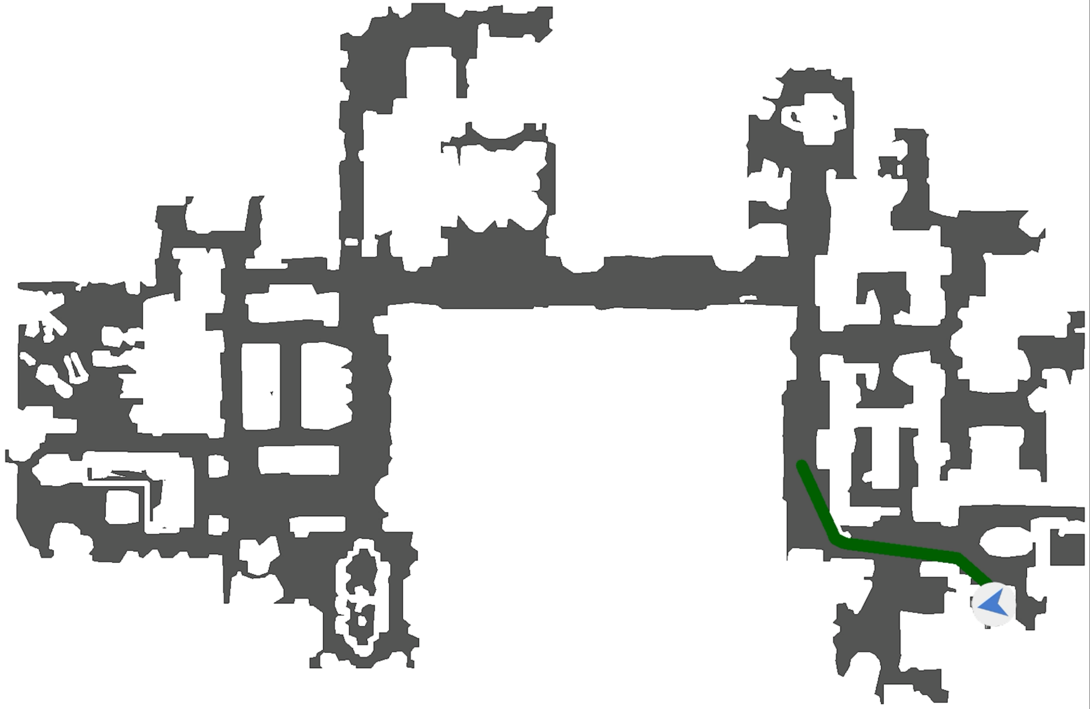
Turn left and slowly go up the stairs, then walk straight ahead for five seconds and stop. Turn right and descend the stairs smoothly.





Qualitative Simulator Result

You're starting in the toilet room of a master bathroom. Turn around to your right. Exit this room into the rest of the master bathroom, walking toward the white, large tub. Veer to the left of the tub and walk through the archway. And, you're now in a hallway. To the left, there is a TV. To the right, there's another large archway with a brown door that goes into what looks like a closet. Go straight ahead. You'll see another arched doorway with two doors that are open. Go straight through. This will take you to a wall with two framed pictures (I guess). Take a right, walk down the hall, and stand next to the piano, and you're done.



北京大學
PEKING UNIVERSITY

PART 4

Conclusion

Conclusion

01 PROBLEM

Reasoning–Control Gap

High-level reasoning

Find target · Avoid obstacle · Choose subgoal

GAP

Low-level control

02 METHOD

MobileVLA-R1

Multi-Granularity CoT

Navigation · Step · Episode

+

SFT + GRPO

03 RESULTS

Strong Generalization

68.3	71.5	0.73
R2R-CE SR	RxR-CE SR	QUARD Avg.

Real-world Unitree Go2

Reason Before Acting

Structured Reasoning → Executable Robot Control

MobileVLA-R1 bridges language understanding and physical action.

北京大学
PEKING UNIVERSITY

谢 谢！

请各位专家老师批评指正