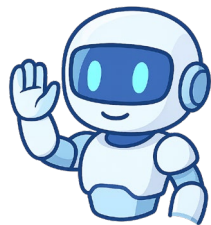


19:00开始, 欢迎观看~



MemoryVLA

Perceptual-Cognitive Memory in Vision-Language-Action Models for Robotic Manipulation

*Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang,
Erjin Zhou, Haoqiang Fan, Xiangyu Zhang and Gao Huang.*



Homepage



Paper



清华大学
Tsinghua University



Hao Shi (石昊)

Department of Automation, Tsinghua University

2025.09.15

CONTENTS

- 1 What Matters in VLA
- 2 MemoryVLA
Solution & Experiments
- 3 Deep Exploration:
Why Memory Important
- 4 Some Thoughts on
Memory & Reasoning



MemoryVLA Team



Hao Shi
(石昊)



Bin Xie
(谢斌)



Yingfei Liu
(刘迎飞)



Lin Sun
(孙林)



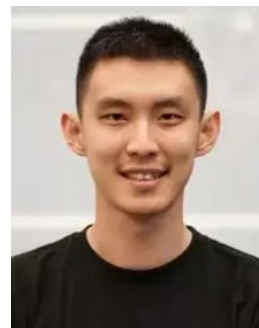
Fengrong Liu
(刘峰荣)



Tiancai Wang
(汪天才)



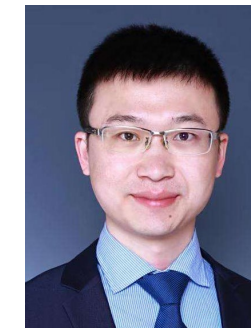
Erjin Zhou
(周而进)



Haoqiang Fan
(范浩强)



Xiangyu Zhang
(张祥雨)



Gao Huang
(黄高)





What Matters in VLA



Background of VLA

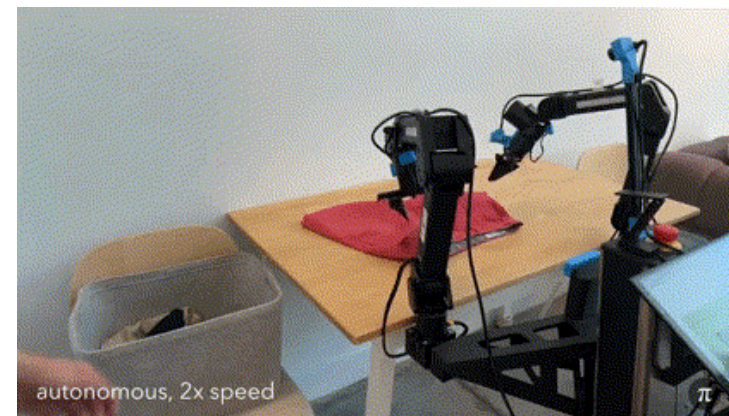
1 Background: Embodied AI



Digital AI

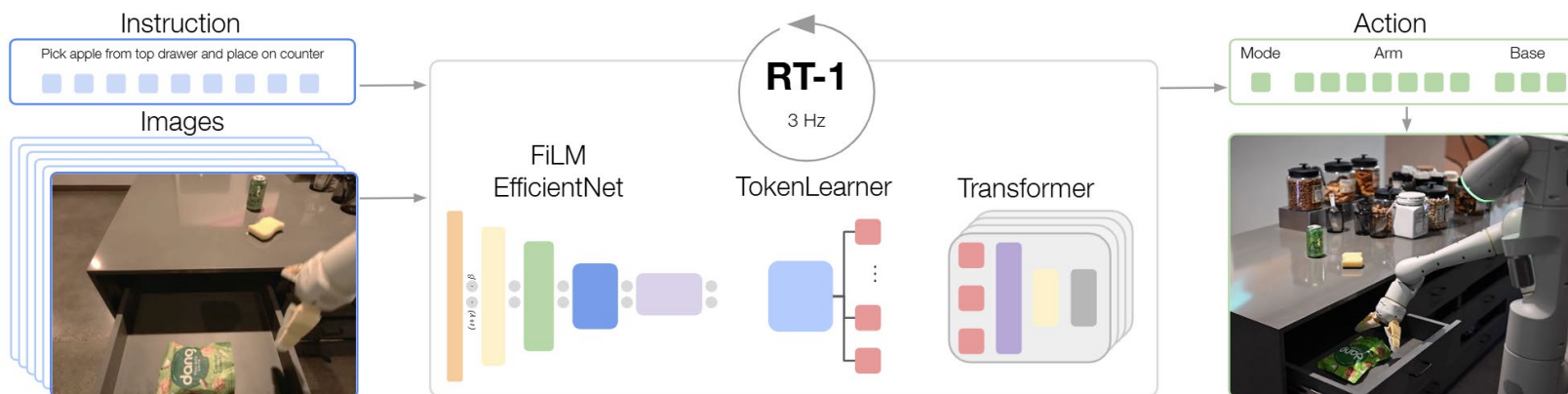


Embodied AI



RT-1 (2212)

Google DeepMind



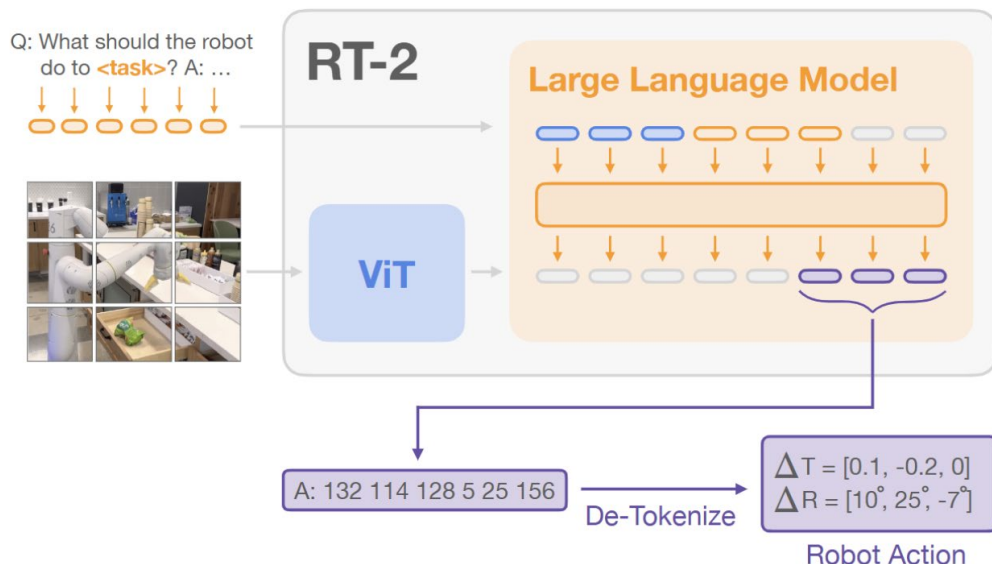
- ❑ Transformer-based
- ❑ Large Real-robot Data
 - 130K demos
 - 17 Month
- ❑ Small Model: 35 M

1 Background: Classic VLA



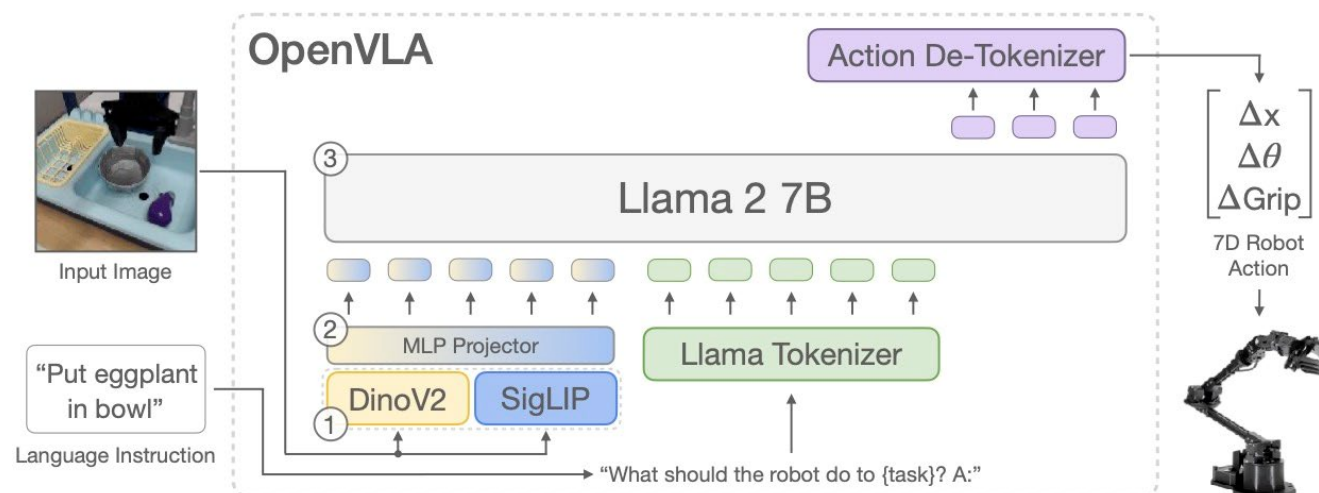
RT-2 (2307)

Google DeepMind



OpenVLA (2406)

Stanford Berkeley
UNIVERSITY OF CALIFORNIA



❑ RT-2 (55B): Pioneer of VLA, closed-source.

- Based on VLM, during de-tokenization, rare vocabulary tokens are replaced by discrete action tokens.

❑ OpenVLA (7B): First open-source VLA, reproduction of RT-2.

- Real-world performance remains limited.

[1] <https://robotics-transformer2.github.io>

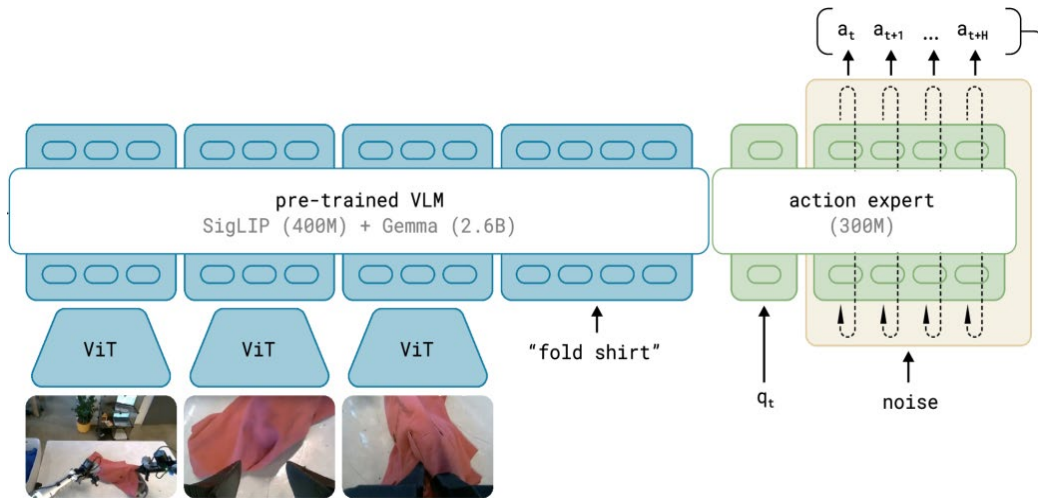
[2] <https://openvla.github.io>

1 Background: 大小脑VLA

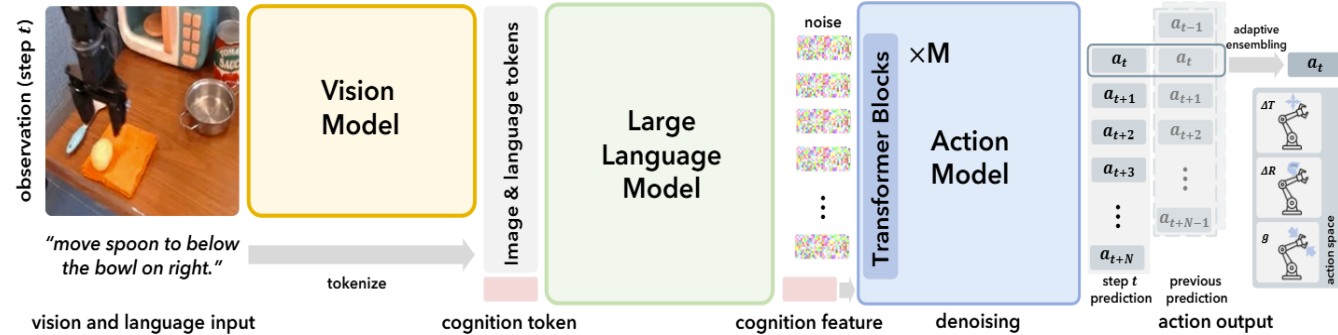


PI-0 (2410)

Physical Intelligence (π)



CogACT (2411)



[1] <https://www.physicalintelligence.company/blog/pi0>
[2] <https://cogact.github.io>

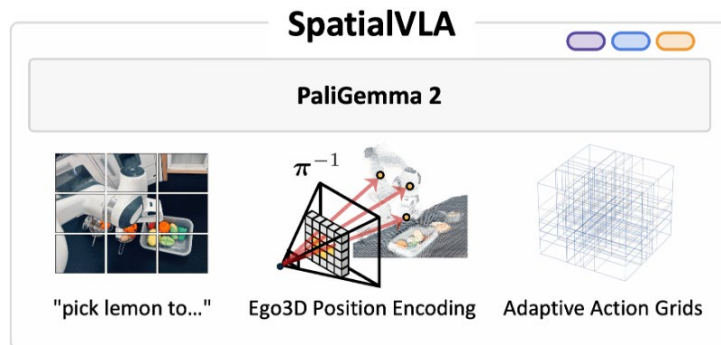
❑ VLM + Diffusion Policy (Brain + Cerebellum): current mainstream architecture

- Multi-modal Behavior Modeling
- Continuous Action Space: OpenVLA' s discrete action space has limited action resolution
- Action Chunking: Look ahead.

1 Background: VLA+Various Additional Inputs

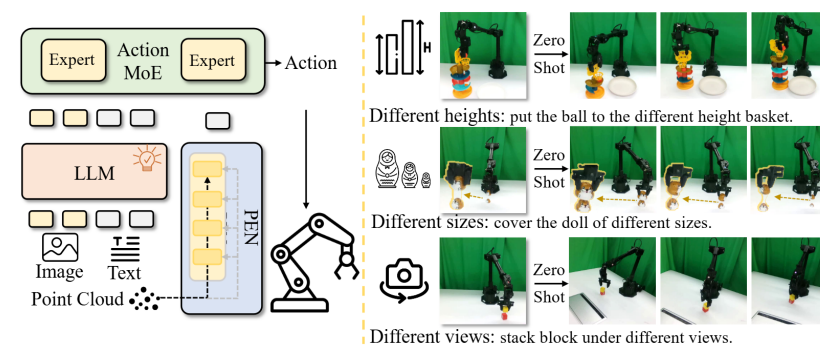


SpatialVLA

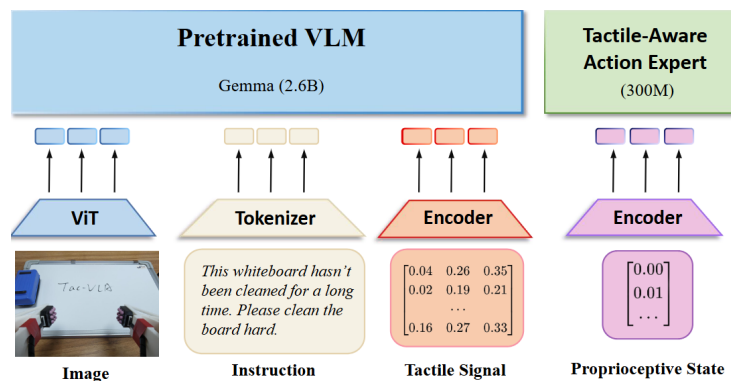


+ 3D

GeoVLA

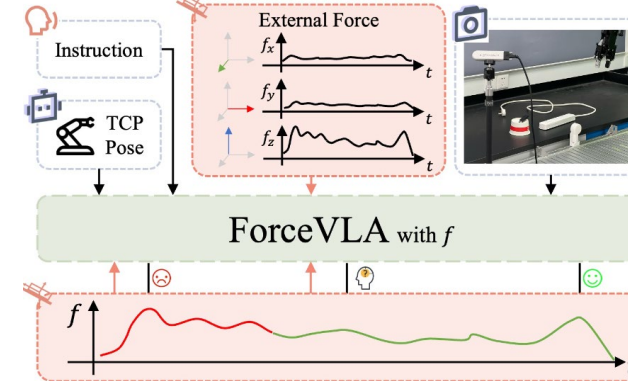


TactileVLA



+ Haptics

ForceVLA



+ Mask

+ Point

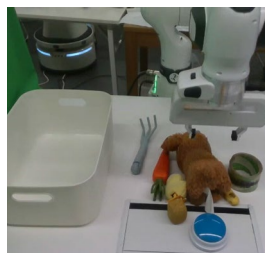
+ View

.....



What Abilities Are These VLAs Evolving?

2 Physical Grounding Matters



Clean table and count.

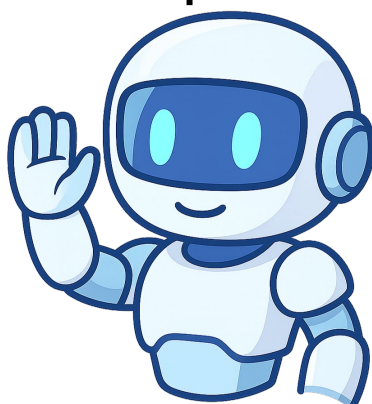
VLA

ΔX ΔY ΔZ $\Delta \theta_X$ $\Delta \theta_Y$ $\Delta \theta_Z$ Grip. Action

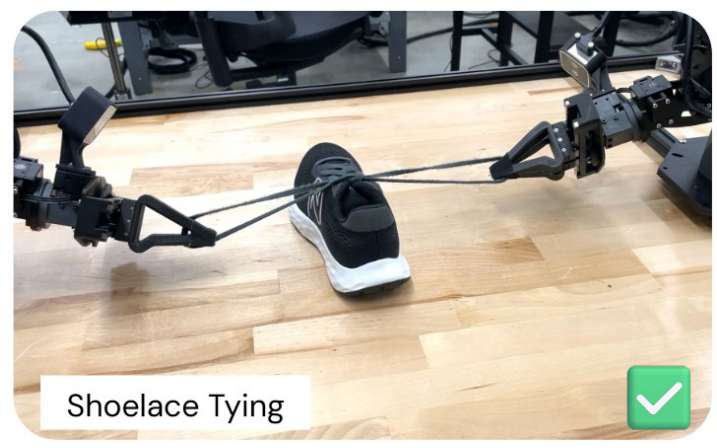


Spatial

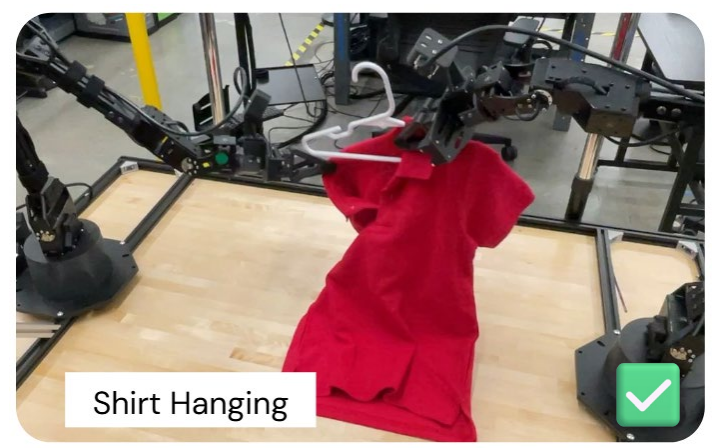
Physical Grounding
Mapping sensor observation and instruction into **6-DoF pose**, with physical-world properties in mind.



If physical grounding is good ...



Shoelace Tying



Shirt Hanging



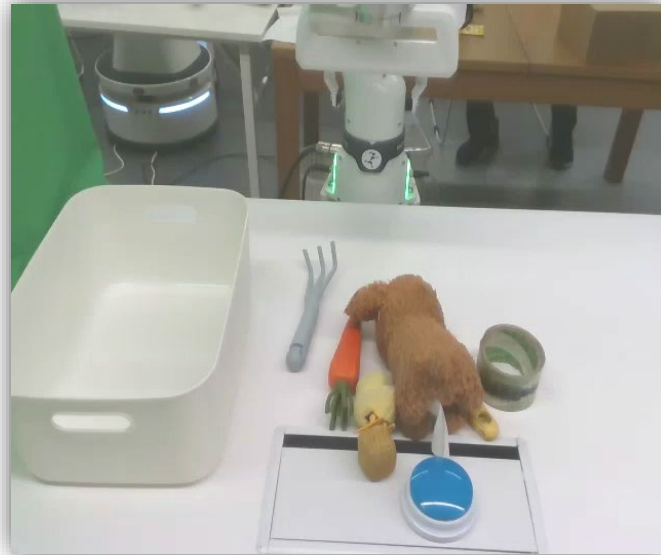


Is Good Physical Grounding Enough?

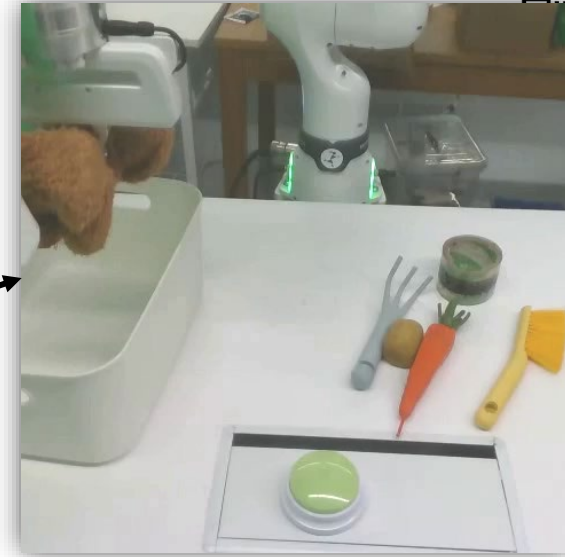
3 Is Physical Grounding Enough?



Hang Cup



Clean Table & Count



Miss



Repeat

Will I press the button, or have I just pressed it?

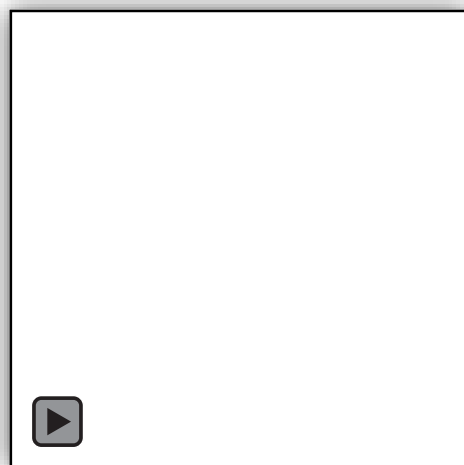


Have I pressed it before?

3 Is Physical Grounding Enough?

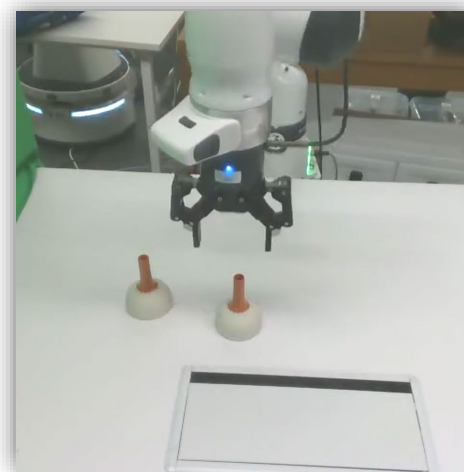
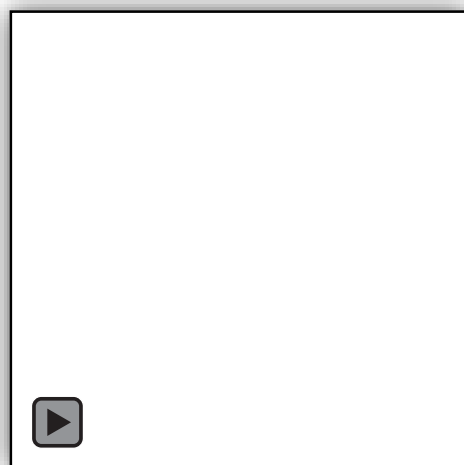


Change Food



What was first placed on the plate? Did I just put down the corn, or the carrot?

Guess Where



Which cup is the block really under? 🤖

3 Why Are Robots Confused?



Clean Table & Count



Change Food



Guess Where

Robotic manipulation tasks are inherently **non-Markovian**

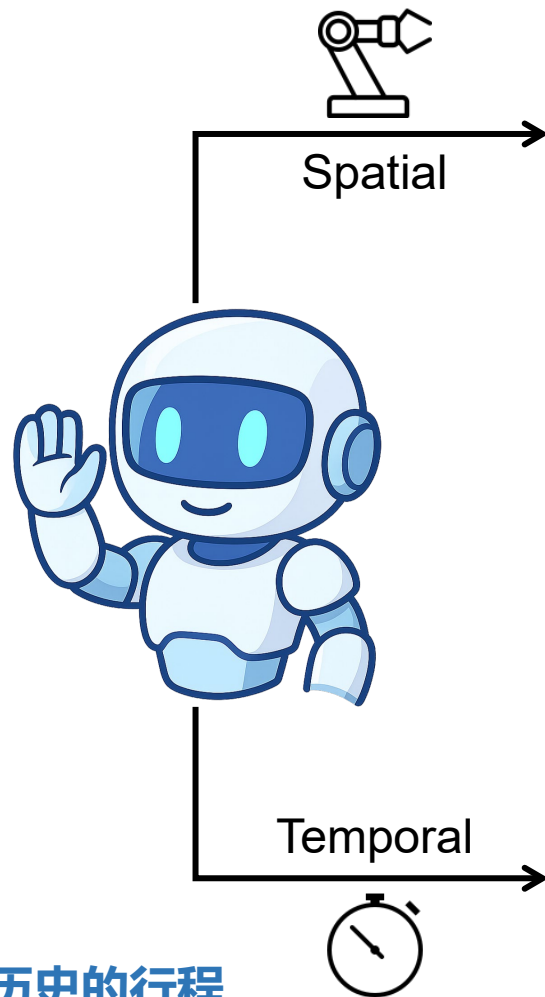
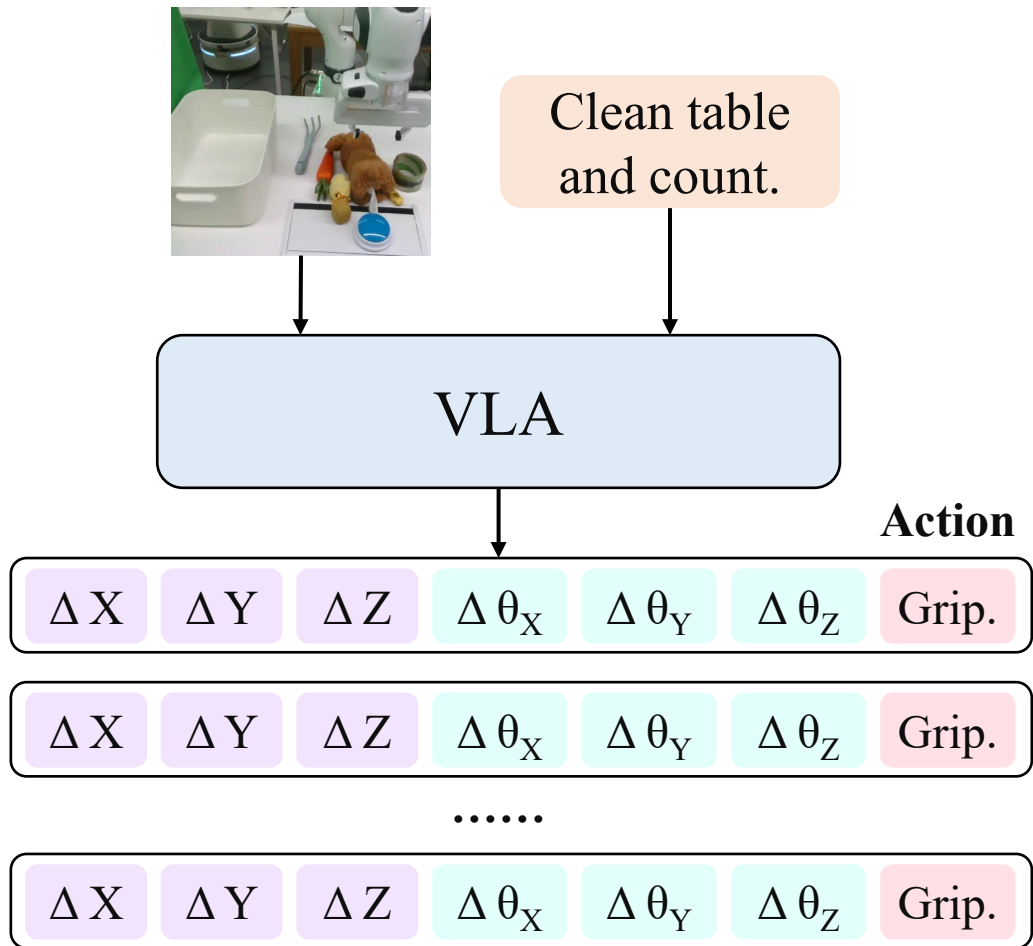
Current decision relies on historical state.

Mainstream VLAs (PI-0, OpenVLA, CogACT) are struggling with **temporally-dependent / long-horizon** manipulation tasks. 🤖



What Matters in VLA for Robotic Manipulation

4 What Matters in VLA



Physical Grounding

Mapping sensor observation and instruction into **6-DoF pose**, with physical-world properties in mind.

抓得准 Skill Generalization

Sequential Decision-Making

Making a series of decisions based on current and historical states to achieve long-term objectives.

抓得对 铭记历史 (Memory) 展望未来 (Reasoning)

一个人的命运，当然要靠自我奋斗，但也要考虑到历史的行程。

- w/o memory & reasoning: 鼠目寸光 (49年入国军)
- w/o physical grounding: 眼高手低 (挥泪斩马谡)



MemoryVLA Solution

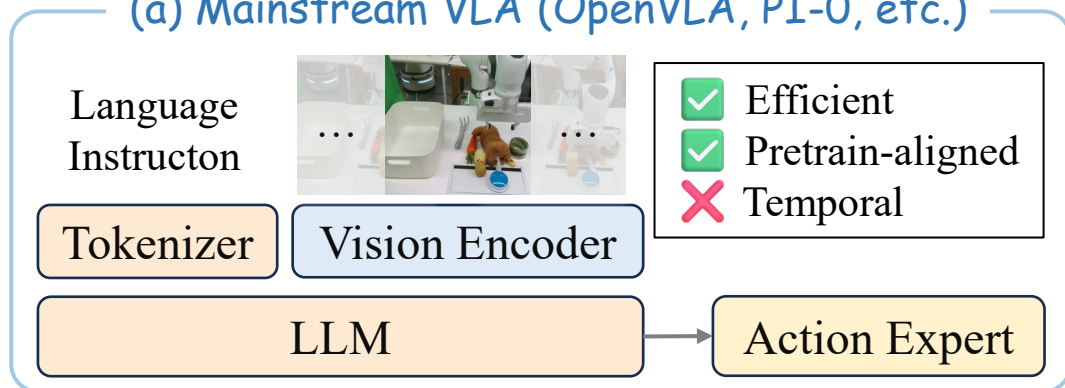


How to Capture Temporal Dependencies?

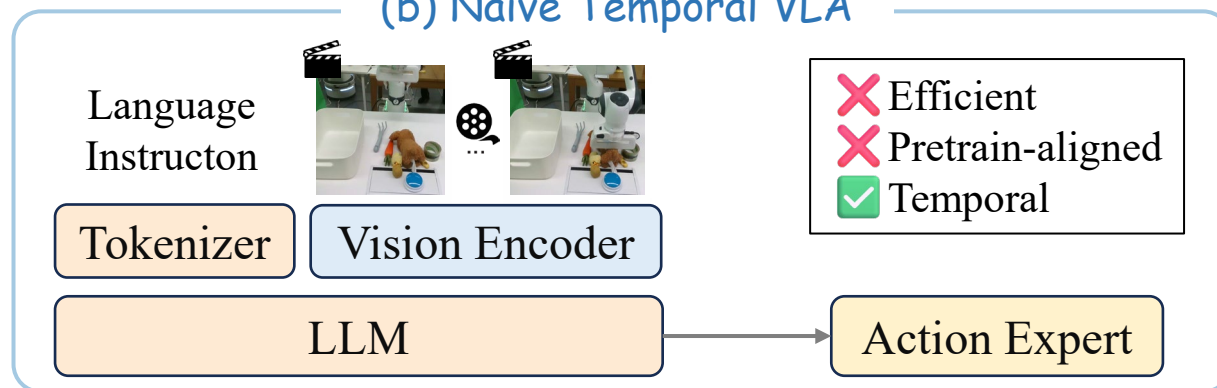
1 Human-inspired



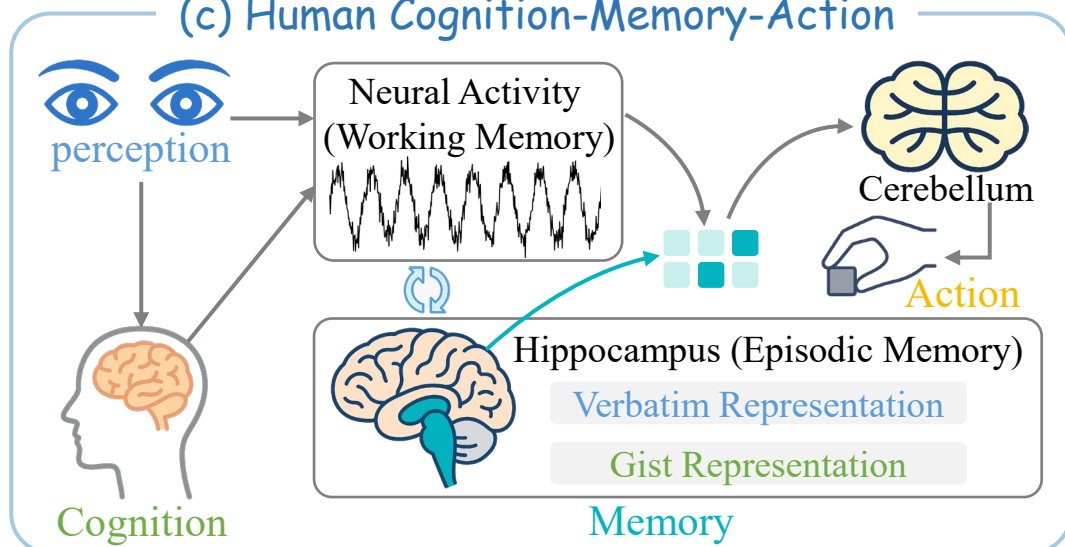
(a) Mainstream VLA (OpenVLA, PI-0, etc.)



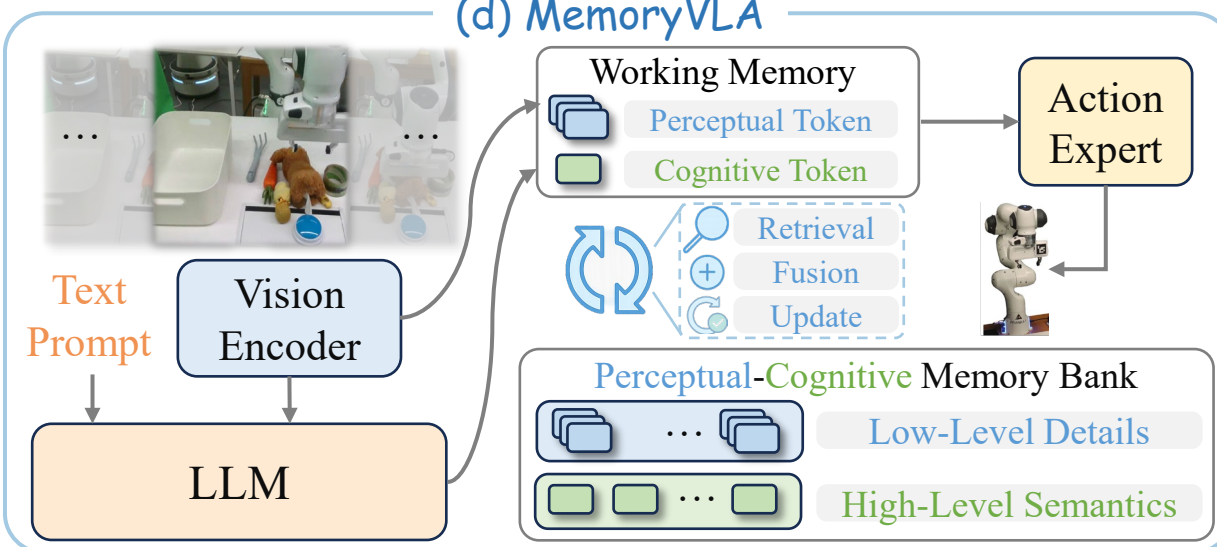
(b) Naive Temporal VLA



(c) Human Cognition-Memory-Action



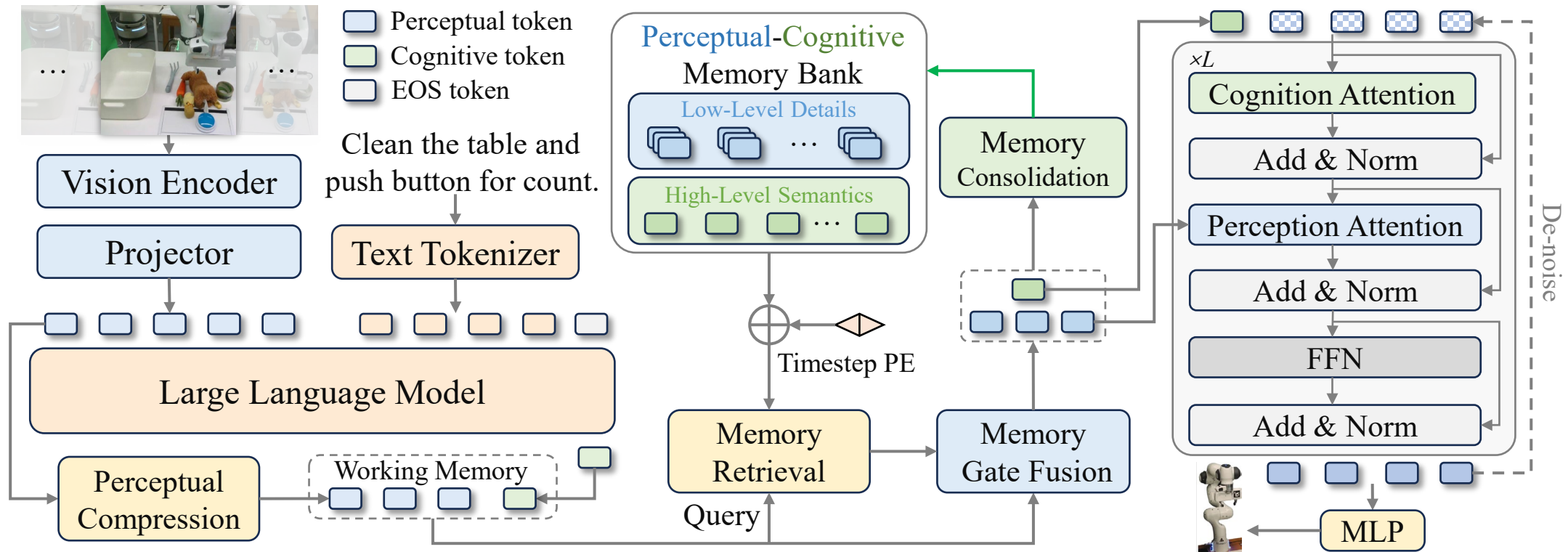
(d) MemoryVLA



Equip the robot with a Hippocampus (海马体).

[1] Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-Trace Theory: An Interim Synthesis. *Learning and Individual Differences*, 7(1), 1–75.
 [2] Baddeley, A. D., & Hitch, G. J. (1974). Working Memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press.
 [3] Tulving, E. (1972). Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381–403). Academic Press.

2 Method Overview

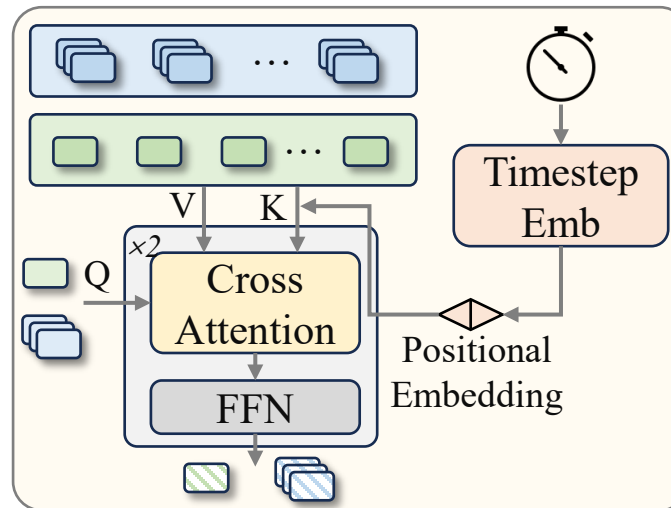
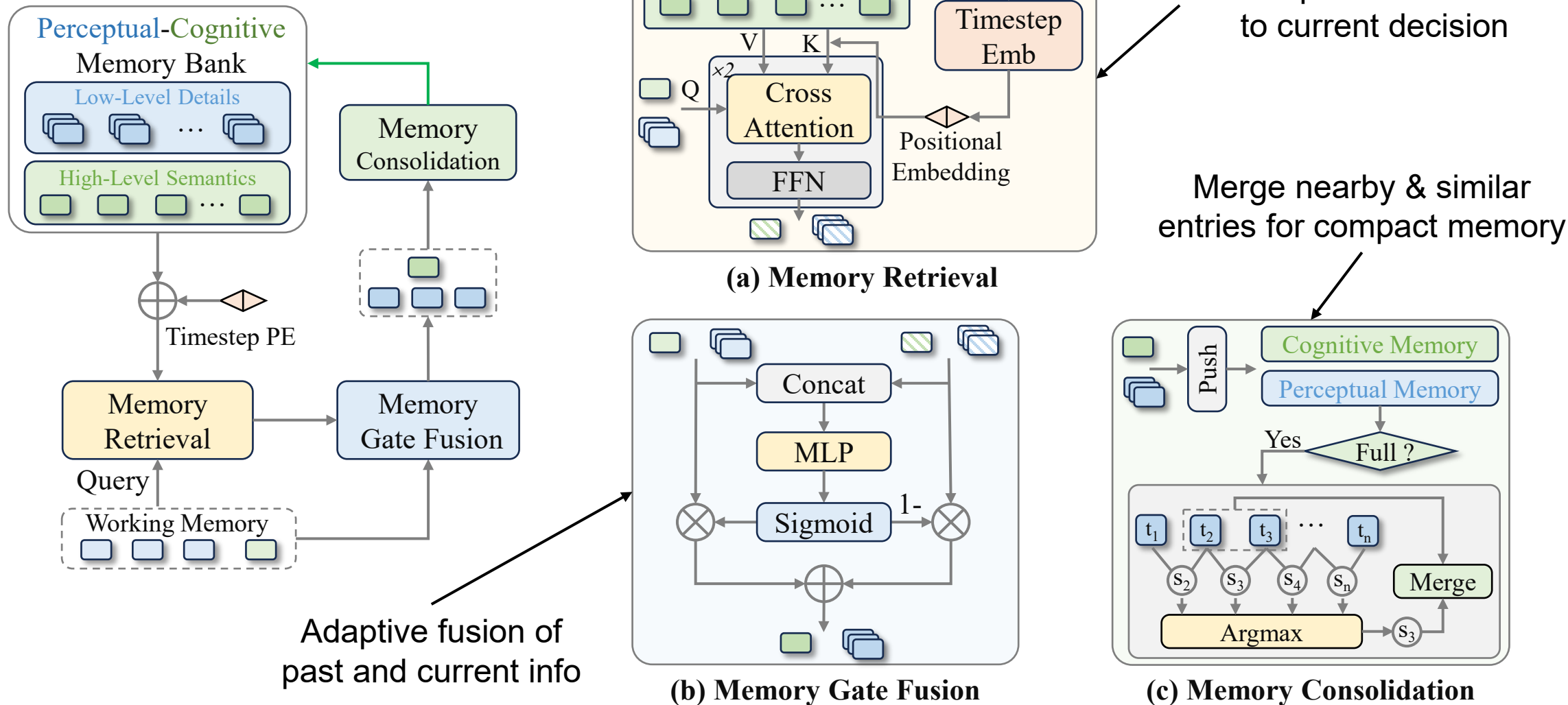


❑ LLM Backbone: Llama-7B, pretrained in Open-X Embodiment.

➤ We are training a new model based on Qwen2.5-7B and pretrained with self-collected data.

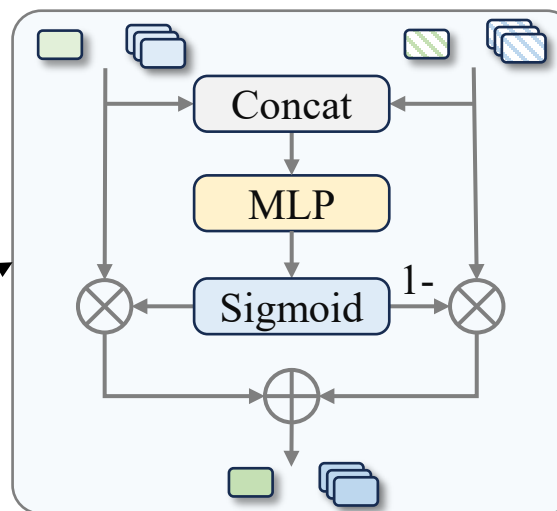
❑ Action Expert: DiT, 300 M.

2 Module Details

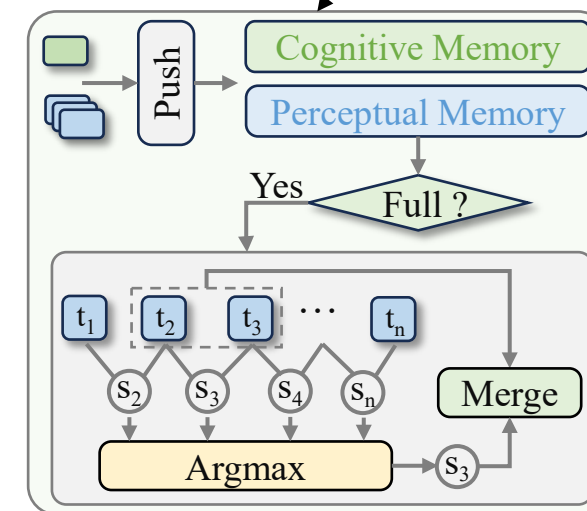


Select past info relevant to current decision

Merge nearby & similar entries for compact memory



Adaptive fusion of past and current info






How is The Performance of MemoryVLA?

3 Experimental Setup

3 Robots, 10 Suites,
150+ Tasks, 500+ Variations

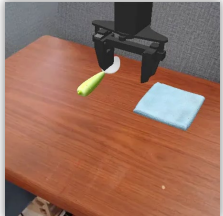


SimplerEnv-Bridge




WidowX Robot


VM Suite




Spoon on Towel



Carrot on Plate




Stack Cube




Eggplant in Basket

SimplerEnv-Fractal




Google Robot


VM Suite
VA Suite




Coke Can



Move Near




Open/Close Drawer




Place in Drawer


LIBERO



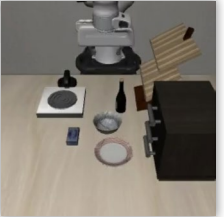
Franka Robot




Spatial Suite




Object Suite



Goal Suite




Long Suite




LIBERO-90 Suite


Real-world Robots




Insert Circle




Egg in Pan




Egg in Oven




Stack Cups




Stack Blocks



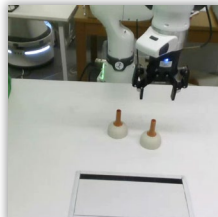
Pick Diverse Fruits



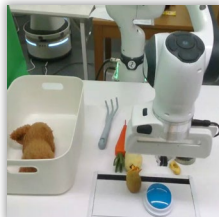
Seq Push Buttons



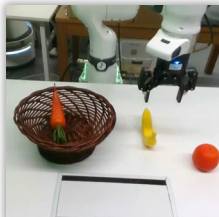
Change Food




Guess Where



Clean Table Count




Pick Place Order




Clean Rest. Table

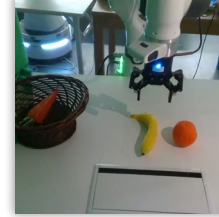
Robustness & Generalization




Background




Distractors




Lighting



Object



Container



Occlusion

3 Real Robots Setup



Franka Robot



WidowX Robot



4 Simulation: SimplerEnv-Bridge

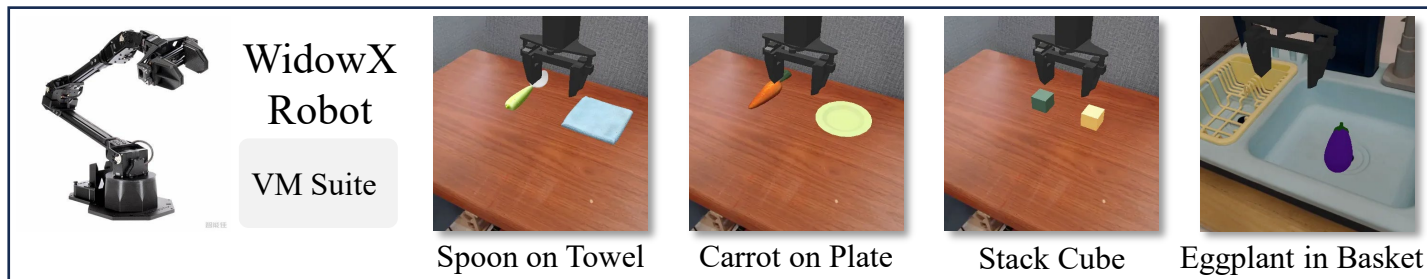
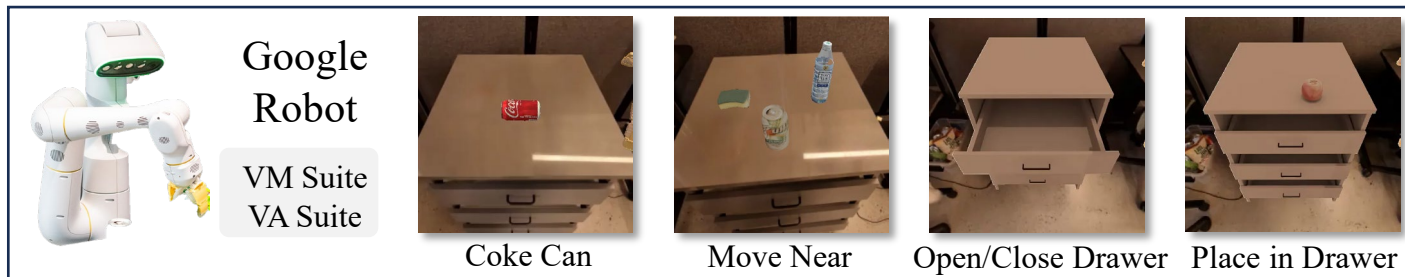


Table 1: Performance comparison on SimplerEnv-Bridge (Li et al., 2024b) with WidowX robot. CogACT-Large is our re-evaluated baseline using official weight, and MemoryVLA achieves a +14.6 gain in average success. Entries marked with * are reproduced from open-pi-zero, which leverage additional proprioceptive state inputs; they also adopt Uniform/Beta timestep sampling.

Method	Spoon on Towel	Carrot on Plate	Stack Cube	Eggplant in Basket	Avg. Success
RT-1-X (O’Neill et al., 2024)	0.0	4.2	0.0	0.0	1.1
OpenVLA (Kim et al., 2024)	4.2	0.0	0.0	12.5	4.2
Octo-Base (Team et al., 2024)	15.8	12.5	0.0	41.7	17.5
TraceVLA (Zheng et al., 2024b)	12.5	16.6	16.6	65.0	27.7
RoboVLMs (Liu et al., 2025a)	45.8	20.8	4.2	79.2	37.5
SpatialVLA (Qu et al., 2025)	16.7	25.0	29.2	100.0	42.7
Magma (Yang et al., 2025)	37.5	29.2	20.8	91.7	44.8
CogACT-Base (Li et al., 2024a)	71.7	50.8	15.0	67.5	51.3
π_0 -Uniform* (Black et al., 2024)	63.3	58.8	21.3	79.2	55.7
CogACT-Large (Li et al., 2024a)	58.3	45.8	29.2	95.8	57.3
π_0 -Beta* (Black et al., 2024)	84.6	55.8	47.9	85.4	68.4
MemoryVLA (Ours)	75.0	75.0	37.5	100.0	71.9 (+14.6)

4 Simulation: SimplerEnv-Fractal



- VM: Real2Sim Benchmark
- VA: Evaluate robustness, generalization

Table 2: Performance comparison on SimplerEnv-Fractal (Li et al., 2024b) with Google robot. Success rates (%) are reported for Visual Matching (VM) and Visual Aggregation (VA) suites. MemoryVLA achieves an overall +4.6 gain over CogACT. O./C. denotes Open/Close, and * follow Tab. 1.

Method	Visual Matching (VM)					Visual Aggregation (VA)					Overall
	Coke Can	Move Near	O. / C. Drawer	Put in Drawer	Avg.	Coke Can	Move Near	O. / C. Drawer	Put in Drawer	Avg.	
Octo-Base (Team et al., 2024)	17.0	4.2	22.7	0.0	11.0	0.6	3.1	1.1	0.0	1.2	6.1
RT-1-X (O'Neill et al., 2024)	56.7	31.7	59.7	21.3	42.4	49.0	32.3	29.4	10.1	30.2	36.3
OpenVLA (Kim et al., 2024)	18.0	56.3	63.0	0.0	34.3	60.8	67.7	28.8	0.0	39.3	36.8
RoboVLMs (Liu et al., 2025a)	76.3	79.0	44.9	27.8	57.0	50.7	62.5	10.3	0.0	30.9	44.0
TraceVLA (Zheng et al., 2024b)	45.0	63.8	63.1	11.1	45.8	64.3	60.6	61.6	12.5	49.8	47.8
RT-2-X (O'Neill et al., 2024)	78.7	77.9	25.0	3.7	46.3	82.3	79.2	35.5	20.6	54.4	50.4
Magma (Yang et al., 2025)	75.0	53.0	58.9	8.3	48.8	68.6	78.5	59.0	24.0	57.5	53.2
SpatialVLA (Qu et al., 2025)	79.3	90.0	54.6	0.0	56.0	78.7	83.0	39.2	6.3	51.8	53.9
π_0 -Uniform* (Black et al., 2024)	88.0	80.3	56.0	52.2	69.1	–	–	–	–	–	–
π_0 -Beta* (Black et al., 2024)	97.9	78.7	62.3	46.6	71.4	–	–	–	–	–	–
CogACT (Li et al., 2024a)	91.3	85.0	71.8	50.9	74.8	89.6	80.8	28.3	46.6	61.3	68.1
MemoryVLA (Ours)	90.7	88.0	84.7	47.2	77.7	80.5	78.8	53.2	58.3	67.7	72.7 (+4.6)

4 Simulation: LIBERO

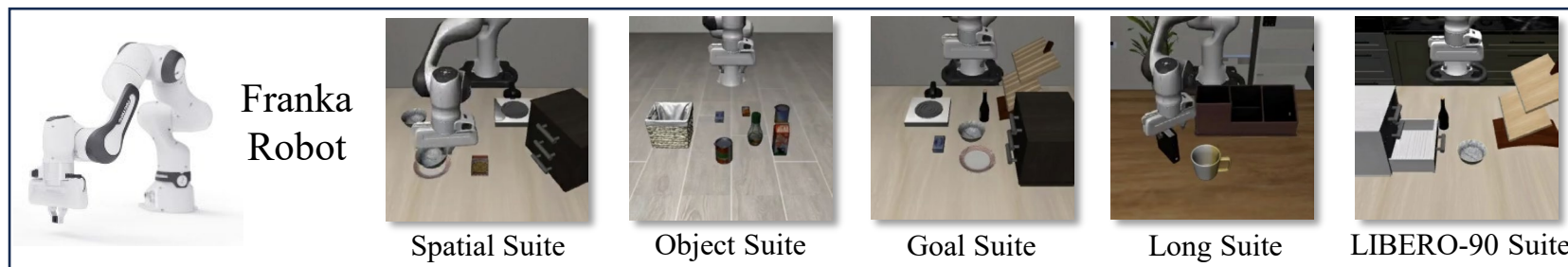


Table 3: **Performance comparison on LIBERO (Liu et al., 2023a) with Franka robot.** Success rates (%) are reported across five suites. * indicates methods using additional proprioceptive and wrist-camera inputs. CogACT results are reproduced by us. For methods without LIBERO-90 results, we report the average over the first four suites.

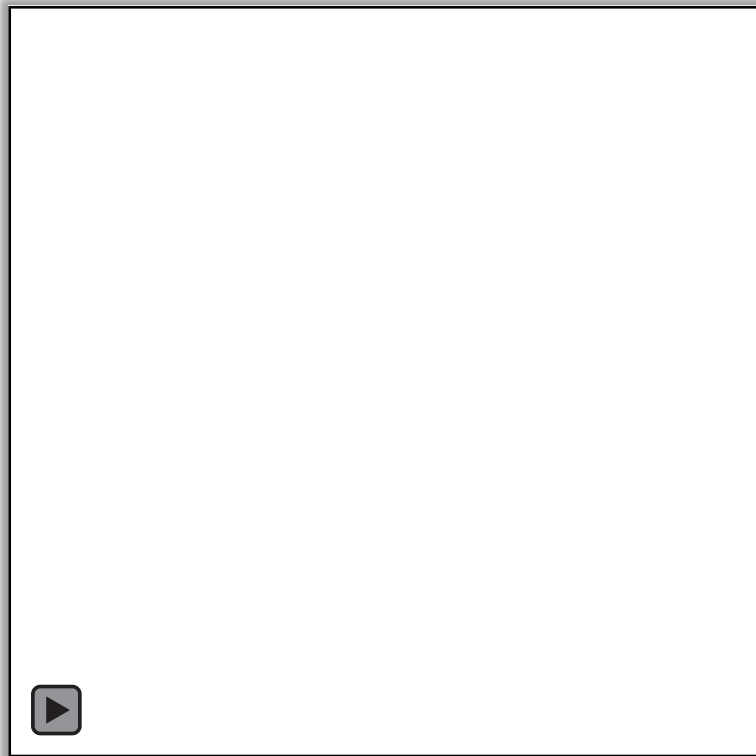
Method	Spatial	Object	Goal	Long	LIBERO-90	Avg. Success
Diffusion Policy (Chi et al., 2023)	78.3	92.5	68.3	50.5	–	72.4
Octo (Team et al., 2024)	78.9	85.7	84.6	51.1	–	75.1
MDT (Reuss et al., 2024)	78.5	87.5	73.5	64.8	–	76.1
UniACT (Zheng et al., 2025b)	77.0	87.0	77.0	70.0	73.0	76.8
MaIL (Jia et al., 2024)	74.3	90.1	81.8	78.6	–	83.5
SpatialVLA (Qu et al., 2025)	88.2	89.9	78.6	55.5	46.2	71.7
TraceVLA (Zheng et al., 2024b)	84.6	85.2	75.1	54.1	–	74.8
OpenVLA (Kim et al., 2024)	84.7	88.4	79.2	53.7	73.5	75.9
CoT-VLA (Zhao et al., 2025)	87.5	91.6	87.6	69.0	–	81.1
π_0 -FAST* (Pertsch et al., 2025)	96.4	96.8	88.6	60.2	83.1	85.0
TriVLA (Liu et al., 2025c)	91.2	93.8	89.8	73.2	–	87.0
4D-VLA (Zhang et al., 2025a)	88.9	95.2	90.9	79.1	–	88.6
CogACT (Li et al., 2024a)	97.2	98.0	90.2	88.8	92.1	<u>93.2</u>
π_0 * (Black et al., 2024)	96.8	98.8	95.8	85.2	–	94.2
MemoryVLA (Ours)	98.4	98.4	96.4	93.4	95.6	96.5 (+3.3)

5 Real-world Demos

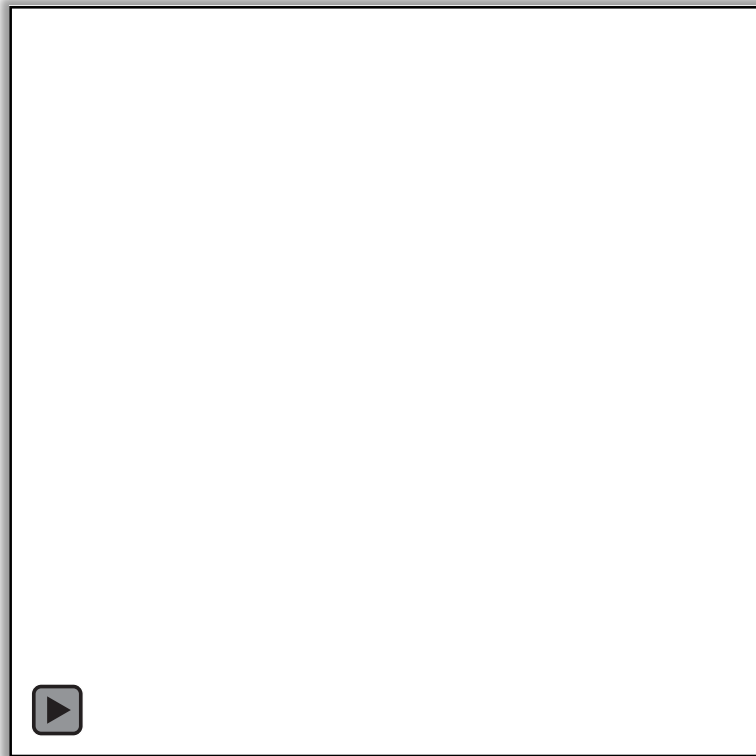


Long-horizon / Temporal Tasks

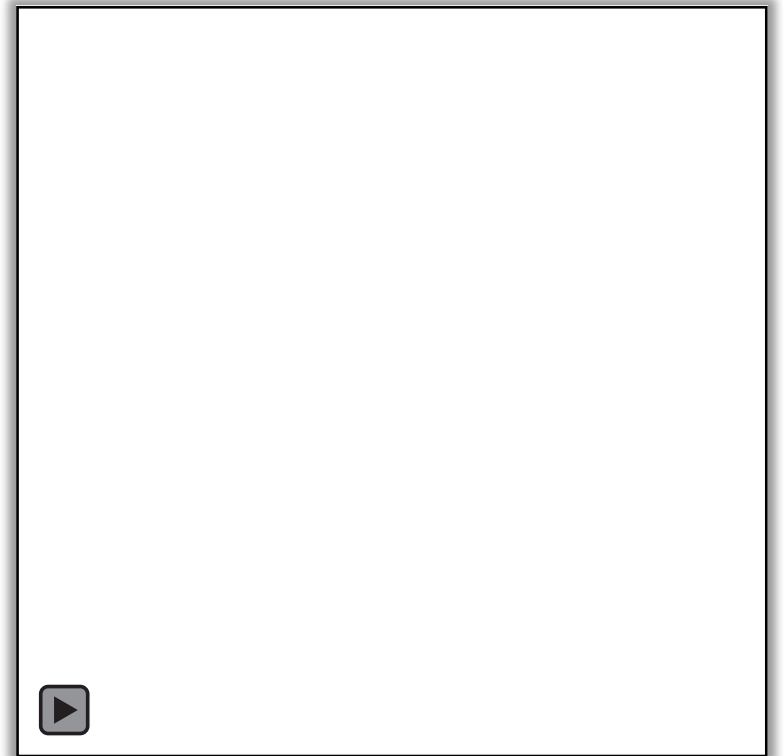
Clean Table & Count



Change Food



Pick Place Order

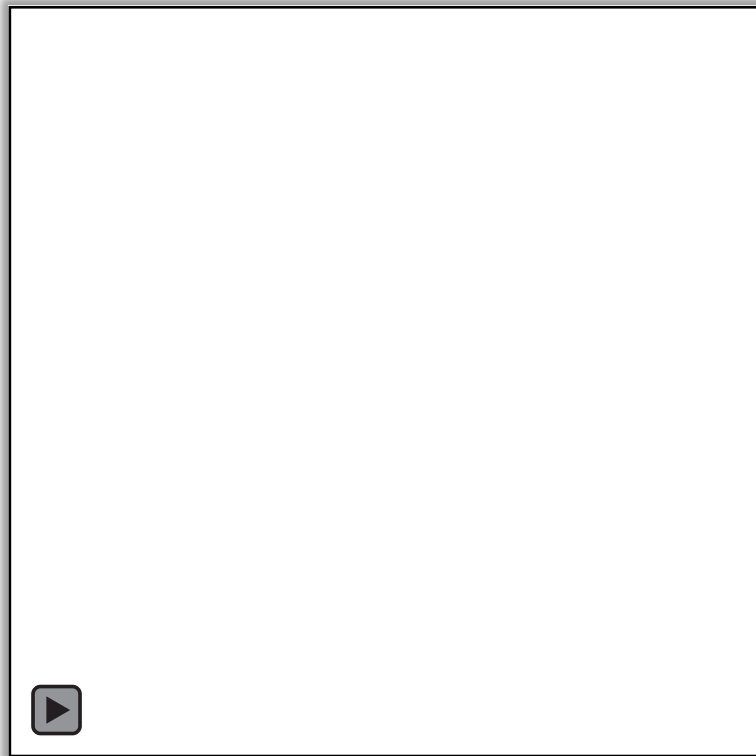


5 Real-world Demos

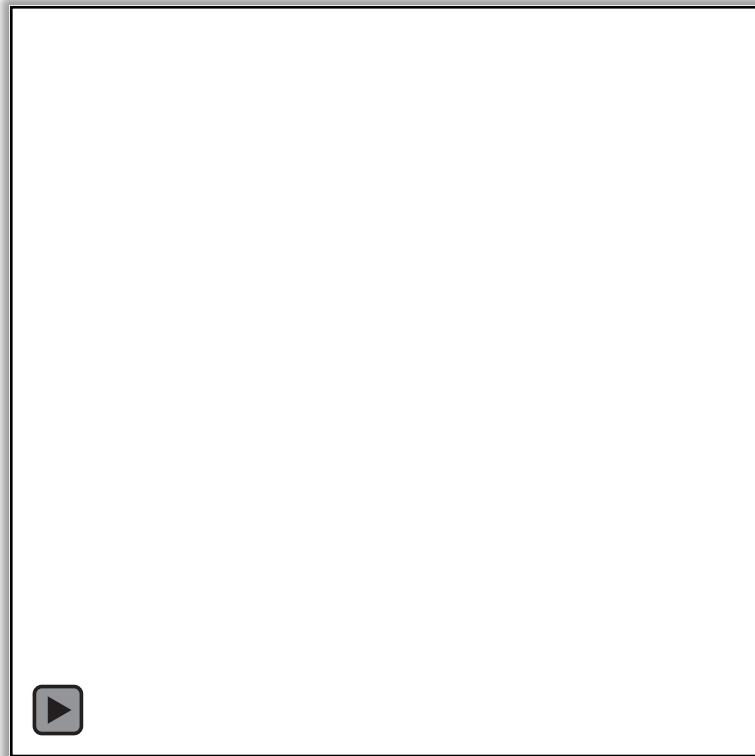


Long-horizon / Temporal Tasks

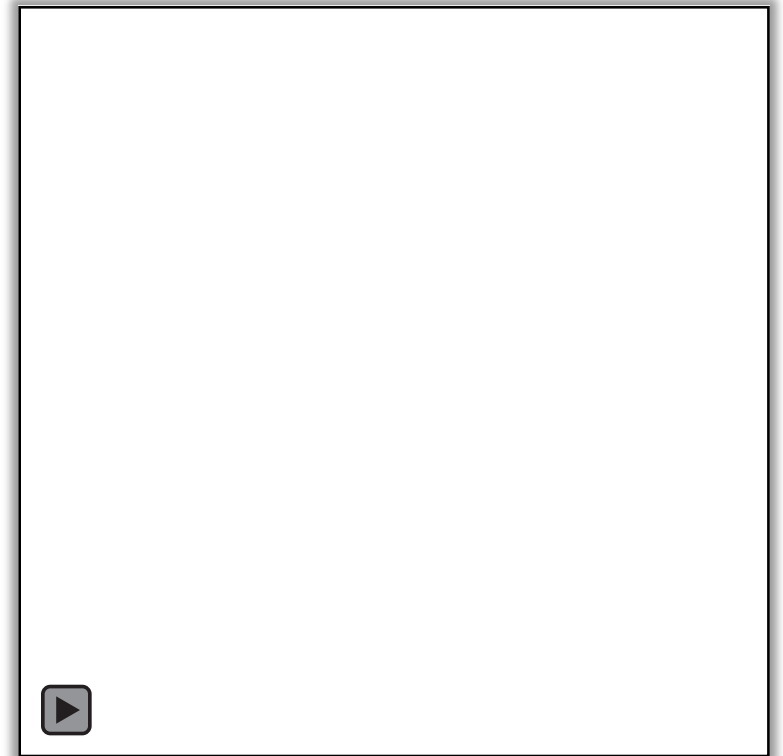
Clean Restaurant Table



Seq Push Buttons



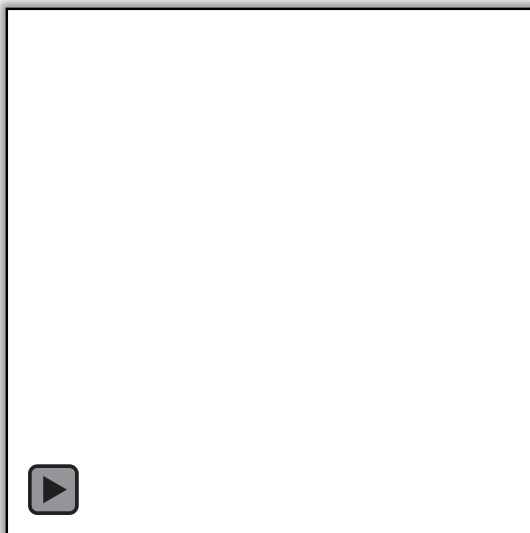
Guess Where



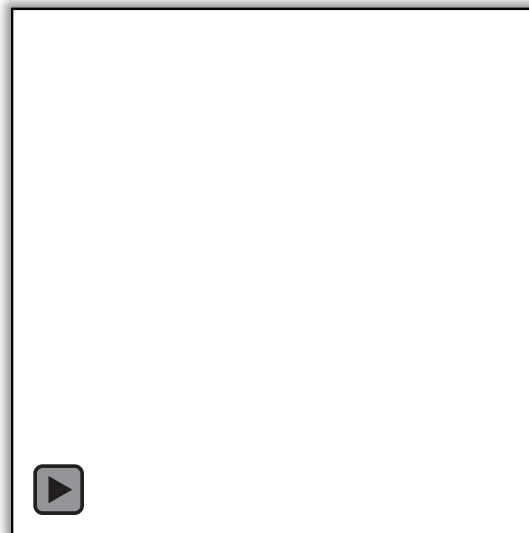
5 Real-world Demos



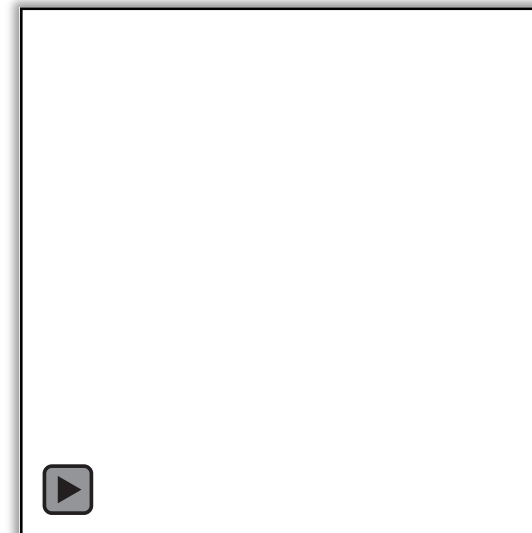
Pick Diverse Fruits



Egg In Oven

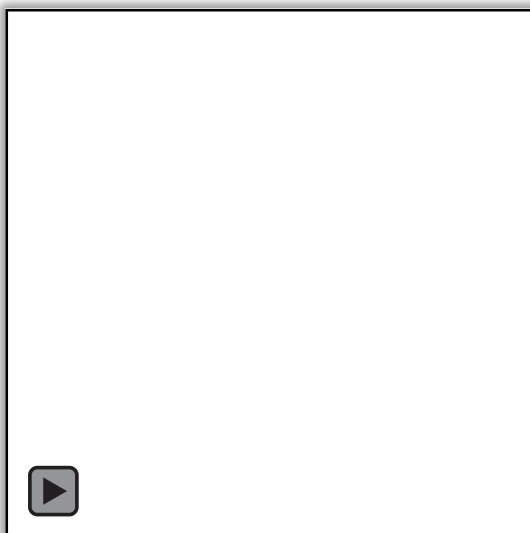


Stack Block

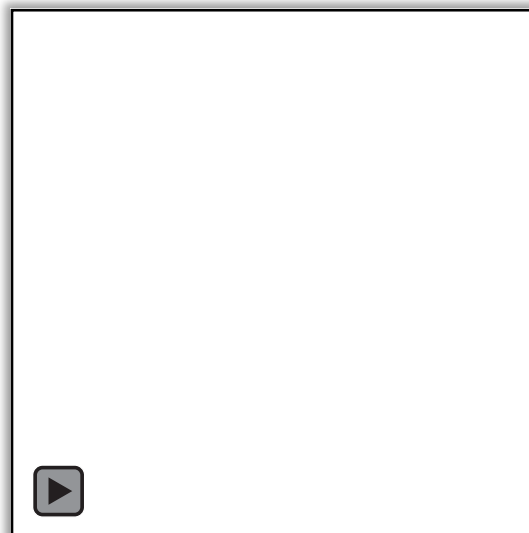


General Manipulation Tasks

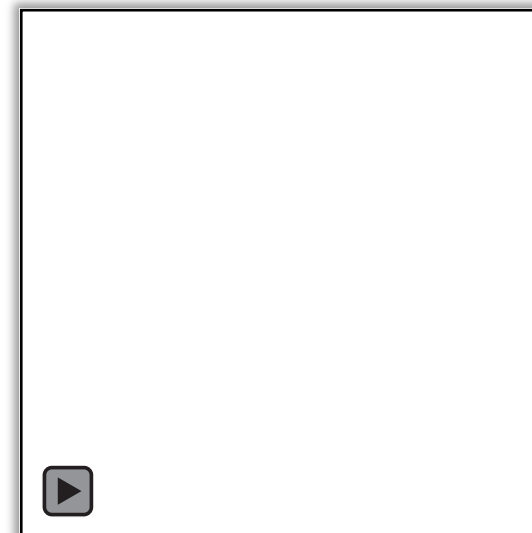
Stack Cup



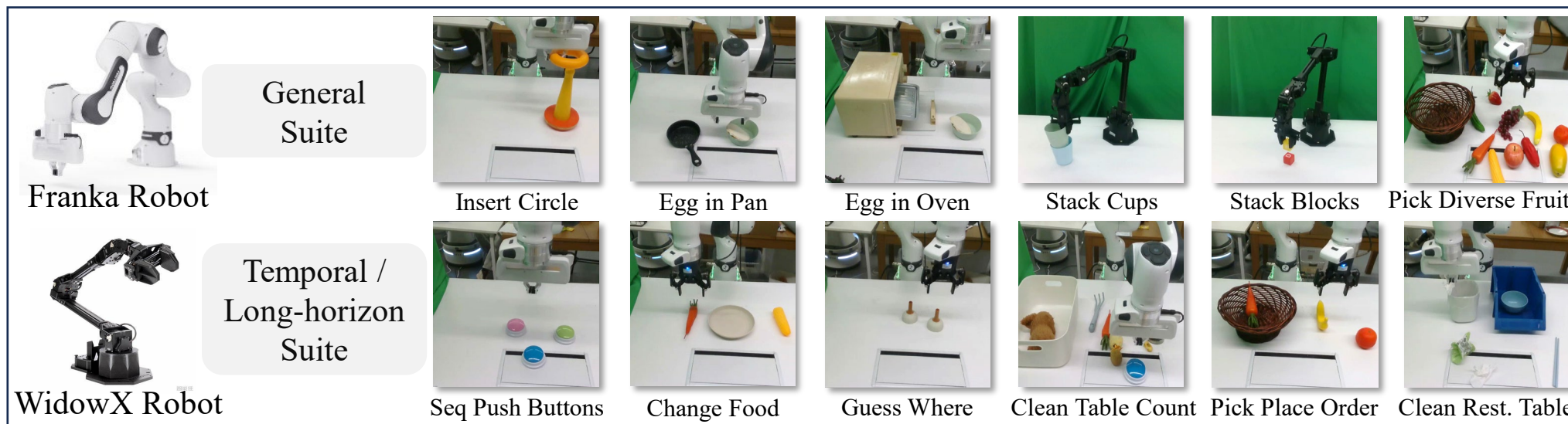
Insert Circle



Egg In Pan



5 Real-world Performance



Method	General Tasks						
	Insert Circle	Egg in Pan	Egg in Oven	Stack Cups	Stack Blocks	Pick Diverse Fruits	Avg. Success
OpenVLA (Kim et al., 2024)	47	27	53	40	13	4	31
π_0 (Black et al., 2024)	67	73	73	87	53	80	72
CogACT (Li et al., 2024a)	80	67	60	93	80	76	<u>76</u>
MemoryVLA (Ours)	87	80	80	93	87	84	85 (+9)

Method	Long-horizon Temporal Tasks						
	Seq. Push Buttons	Change Food	Guess Where	Clean Table & Count	Pick Place Order	Clean Rest. Table	Avg. Success
OpenVLA (Kim et al., 2024)	6	3	0	15	27	0	9
π_0 (Black et al., 2024)	25	42	24	61	82	80	52
CogACT (Li et al., 2024a)	15	47	40	67	90	84	<u>57</u>
MemoryVLA (Ours)	58	85	72	84	100	96	83 (+26)

- ❑ All methods use a single third-person view.
- ❑ Evaluation is step-wise for Long-horizon Temporal Tasks.

6 Robustness & Generalization

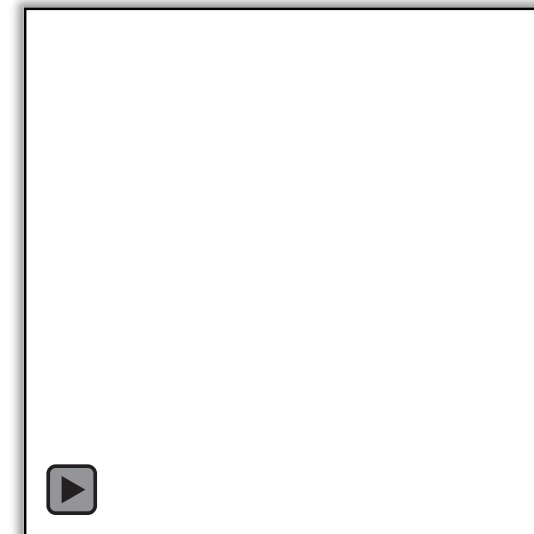
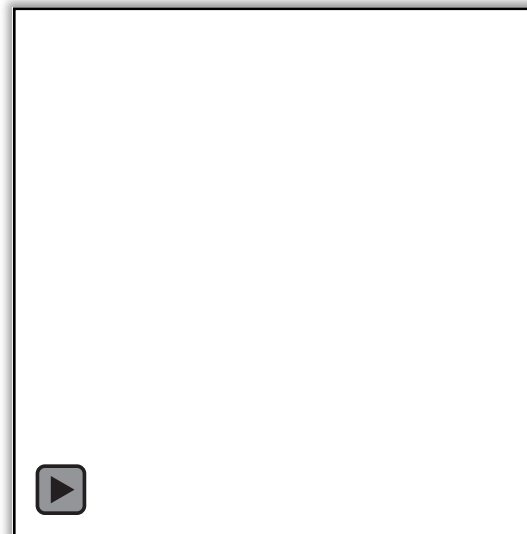
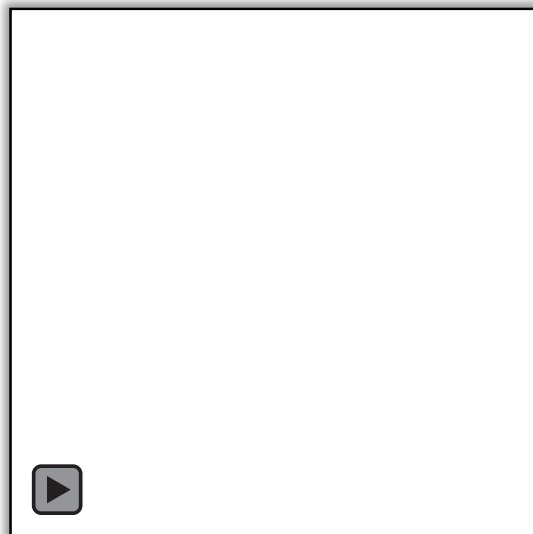
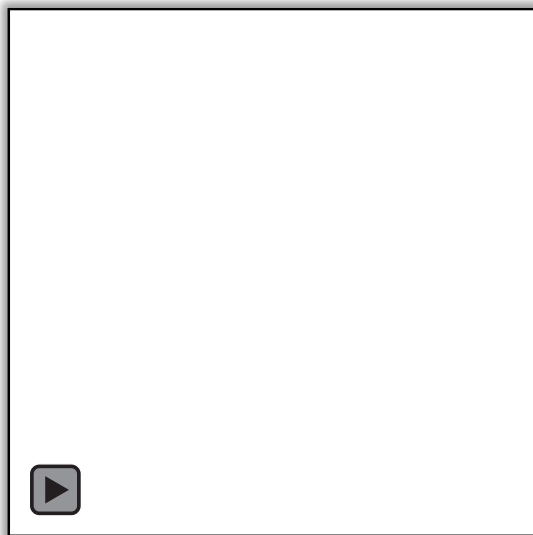


Unseen Background

Unseen Distractors

Unseen Lighting

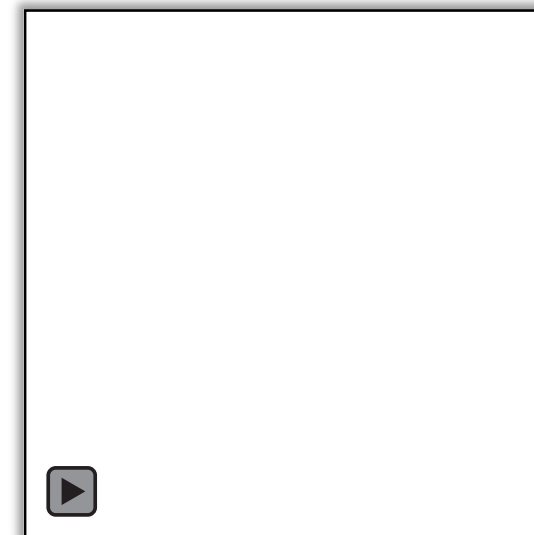
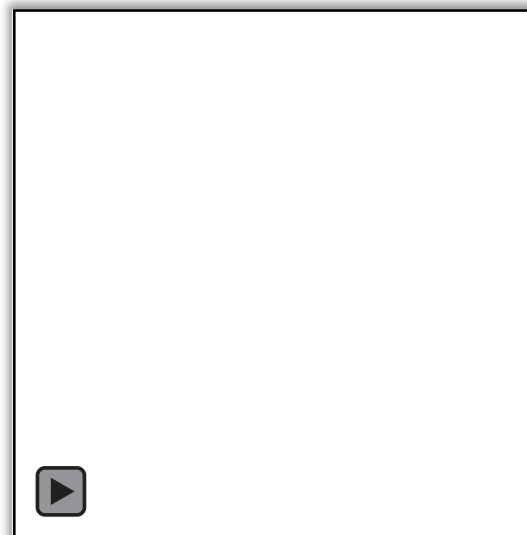
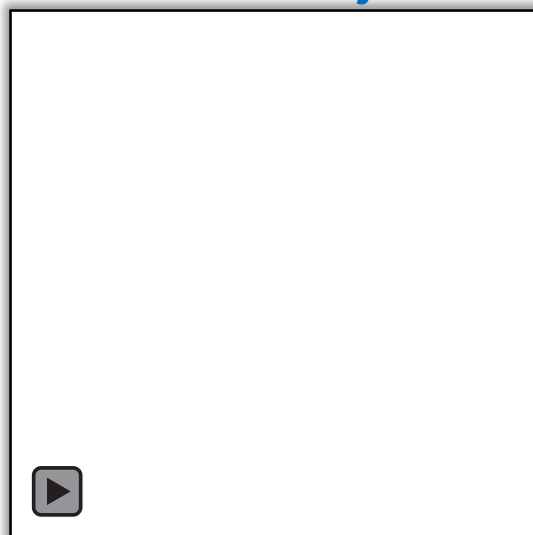
Base Setup



Unseen Object

Unseen Container

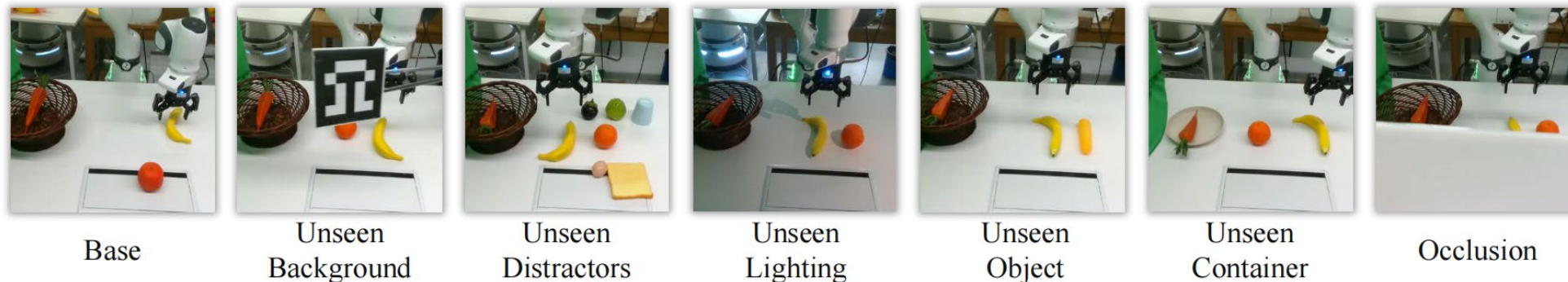
Occlusion



6 Robustness & Generalization



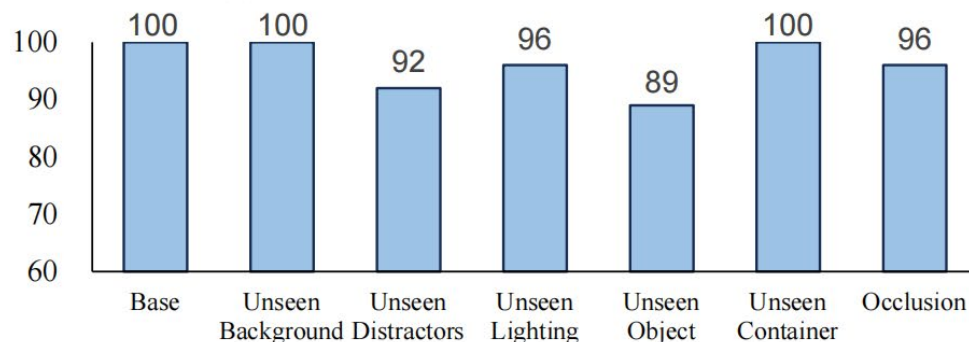
(a) Various OOD Variants of Pick Place Order



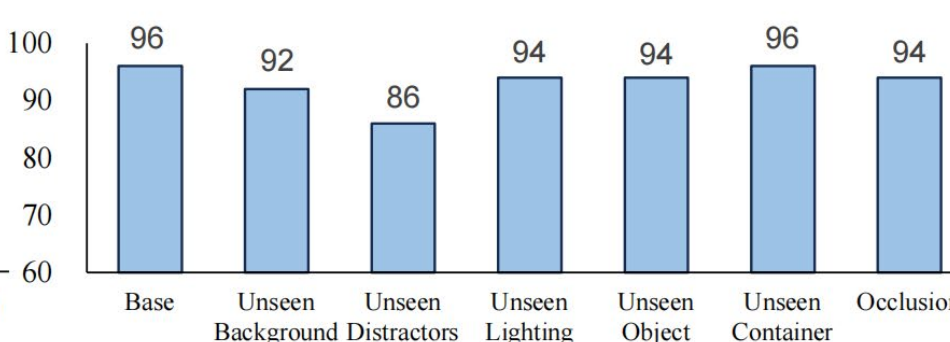
(b) Various OOD Variants of Clean Restaurant Table



(c) Generalization of Pick Place Order



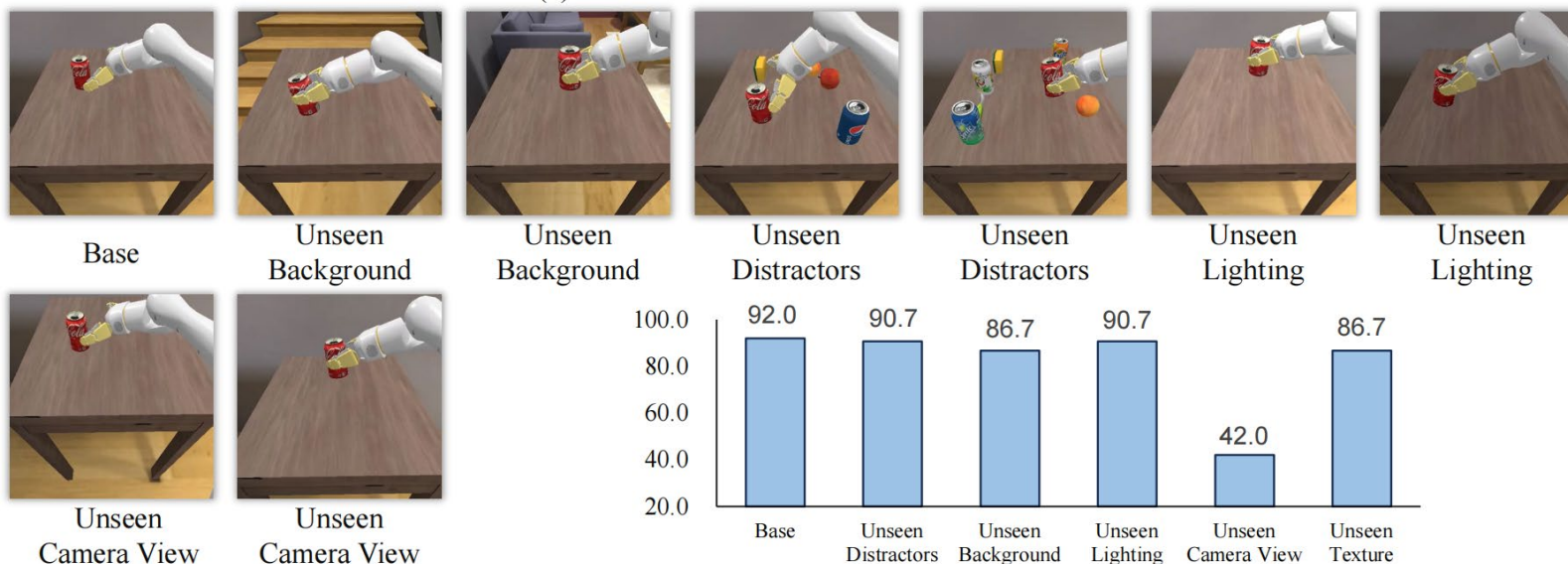
(d) Generalization of Clean Restaurant Table



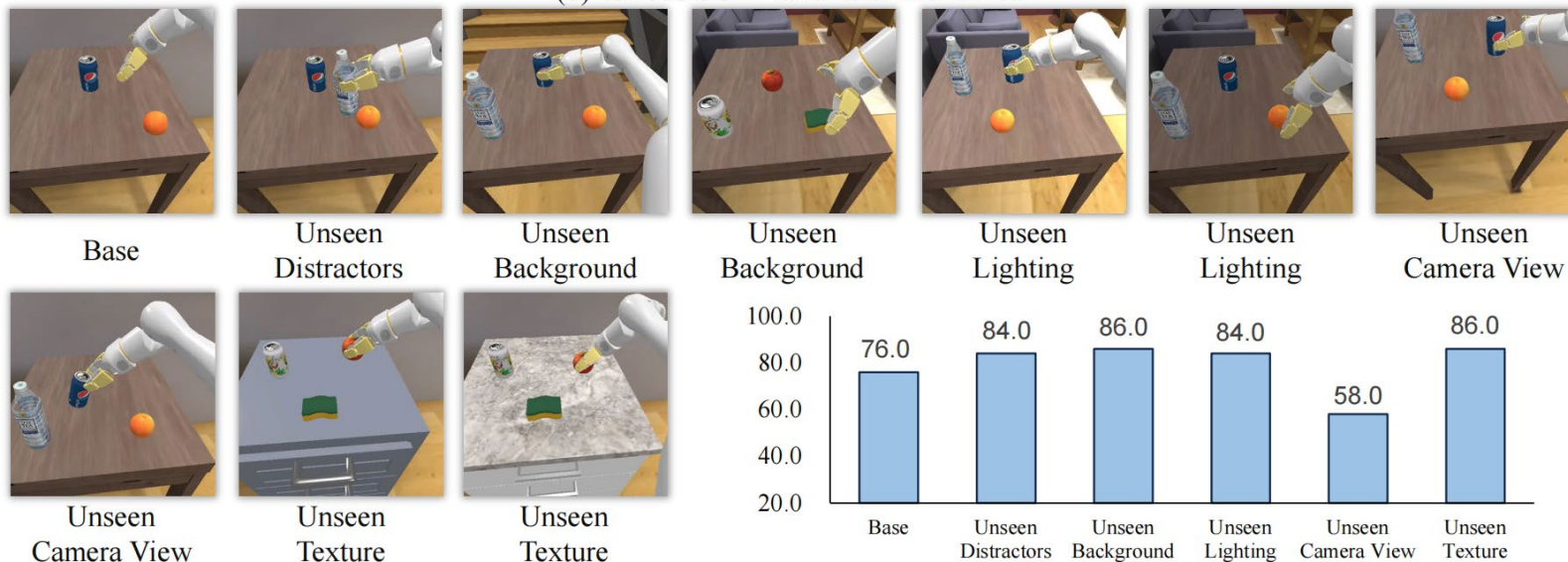
6 Robustness & Generalization



(a) Various OOD Variants of Pick Coke Can



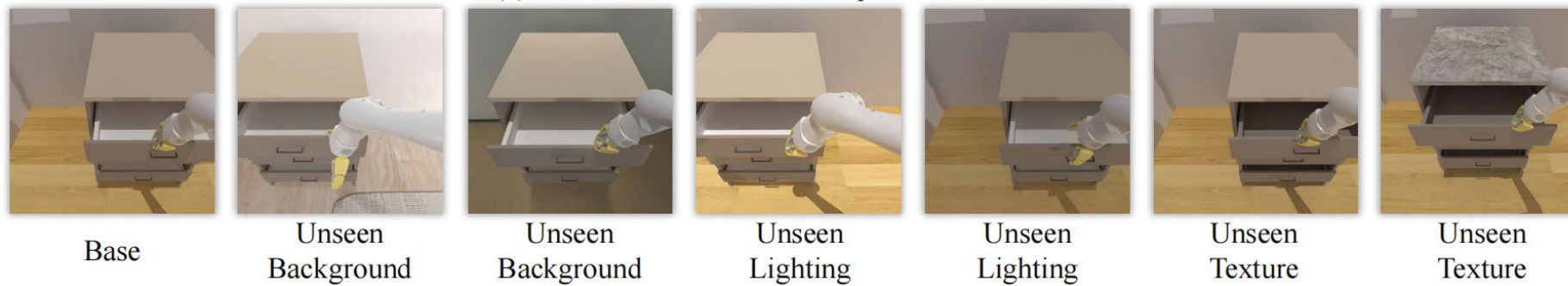
(b) Various OOD Variants of Move Near



6 Robustness & Generalization



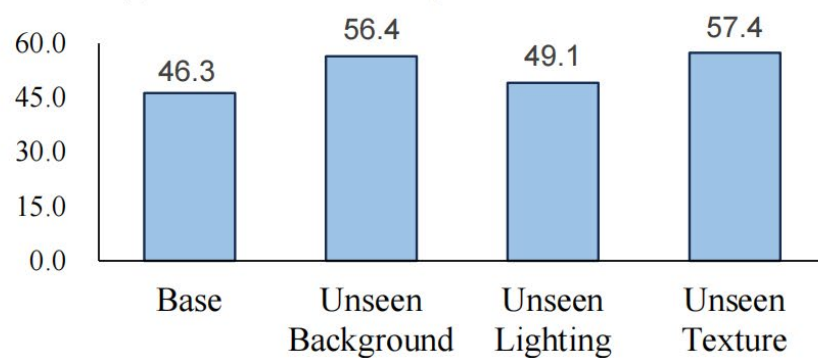
(a) Various OOD Variants of Open / Closed Drawer



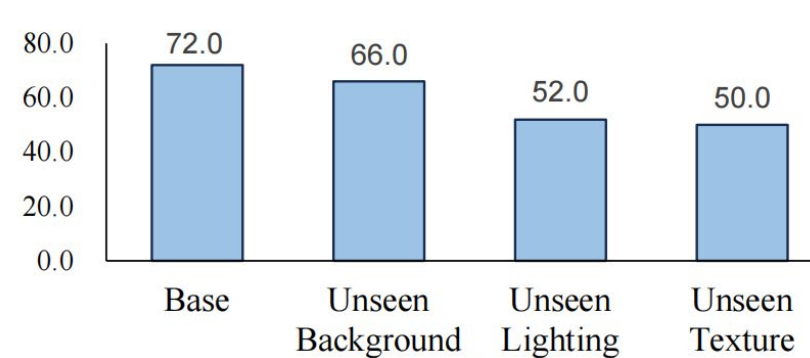
(b) Various OOD Variants of Place Apple Into Drawer



(c) Generalization of Open / Closed Drawer



(d) Generalization of Place Apple Into Drawer





Deep Exploration: Why Memory Important

1 Robotics vs CV

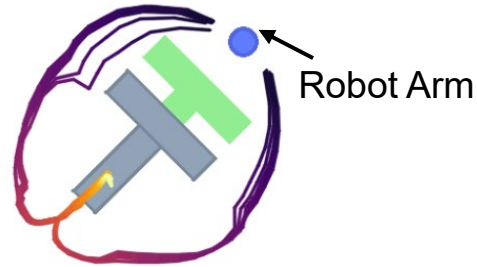


CV



Deterministic prediction

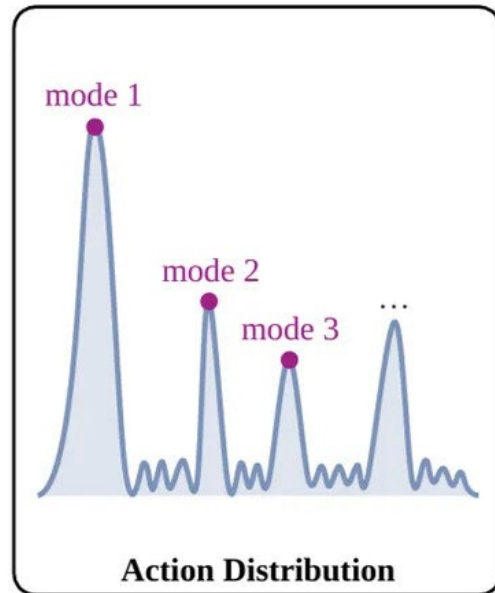
Robotics



Non-deterministic prediction

One observation corresponds to **multiple action label**.

Supervised labels **fight** each other!!



Multimodal Behavior

Multimodal behavior limits the scaling law of robotic models.

More data, more fights

Longer horizons, tougher fights

DP can solve it!

Really?



Didn't DP Solve Multimodal Behavior?

2 DP Really Solve Multimodal Behavior?

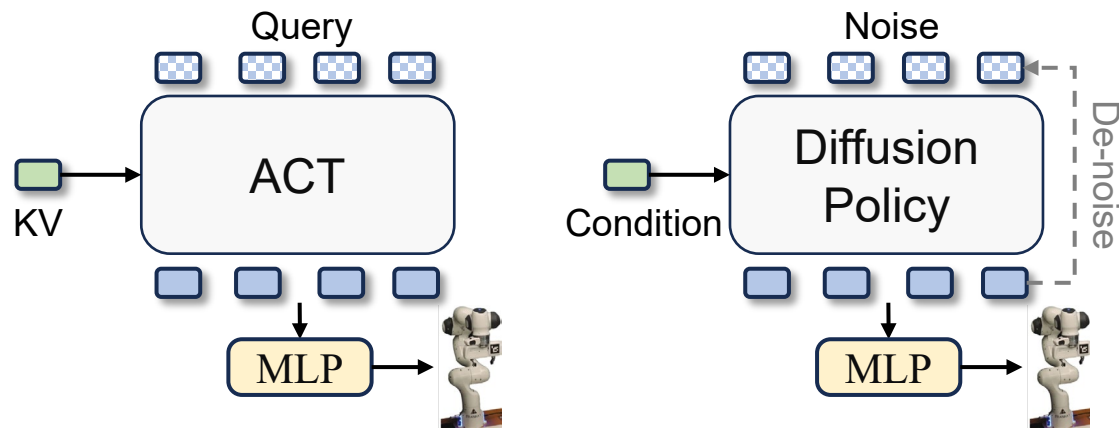


Multimodal Behavior

Multimodal behavior limits the scaling law of robotic models.

Supervised labels fight each other!!

Why Diffusion Policy Work?



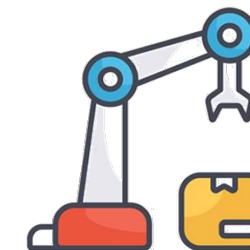
- ❑ Query is with prior, Noise is **random**.
- ❑ Denoise: one step generation → **multi-step** generation

Does It Really Work?

Robotics vs Generation



Generation



Robotics

- | | |
|-------------------|--------------------|
| ➤ Large data | ➤ Limited data |
| ➤ Dense supervise | ➤ Sparse supervise |
| ➤ Short horizon | ➤ Long horizon |
| ➤ Digital world | ➤ Real world |

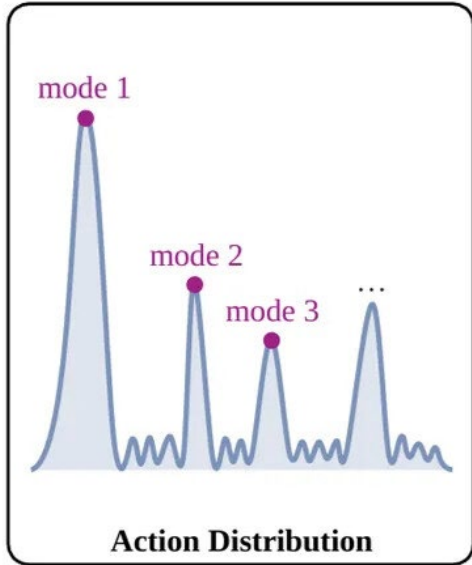
具身最难的问题，就是大家没认识到他到底有多难。——智源大会2025, 王鹤 @ PKU

Theory works, training fails.

3 One-to-Many Mapping



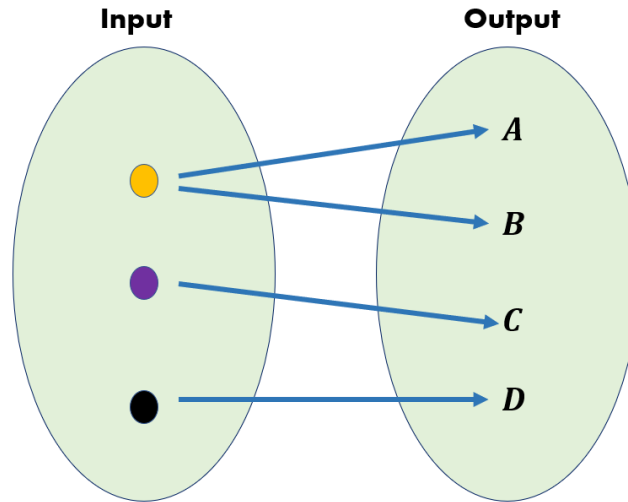
Multimodal Behavior



Multimodal behavior limits the scaling law of robotic models.

Supervised labels fight each other!!

One-to-Many Mapping



Imitation learning is difficult to model **non-functional mappings**.

How to solve?

- ❑ Additional Input
- ❑ Stronger Output Supervision

Additional Input

+ 3D	+ Haptics	+ View
+ Point	+ Mask	+

+ Memory

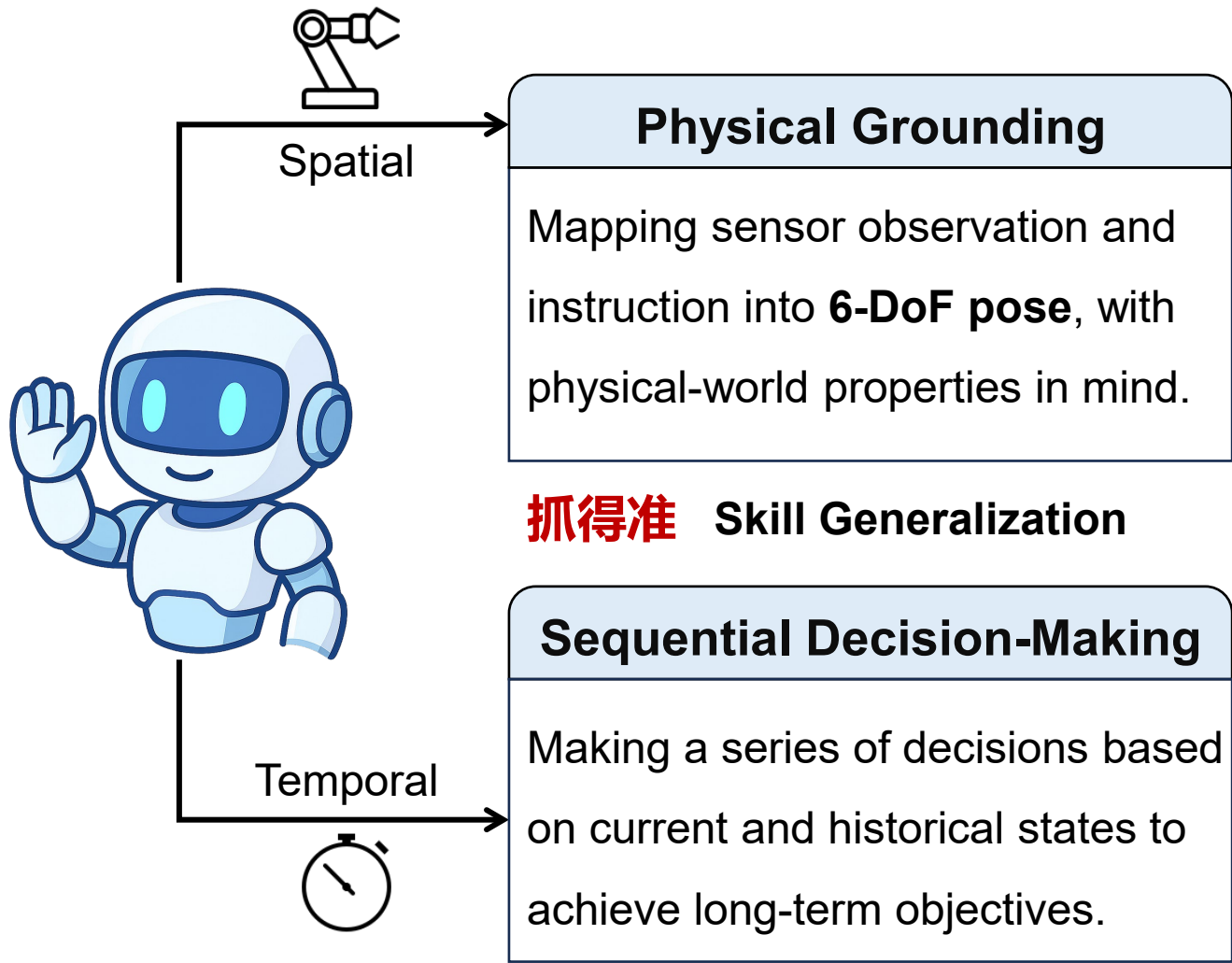
- ❑ Self-generated additional input.
- ❑ Low cost and strong effectiveness.
- ❑ Episode-adaptive dynamic prior.

Call for Attention to Memory in VLA

Strong Supervision

- ❑ Richer Distributions: Diffusion, GMM
- ❑ Denser Supervision Signals

4 What Matters in VLA



抓得准 Skill Generalization

抓得对 铭记历史 (Memory)
展望未来 (Reasoning)



(Convert language data into human reading time)

Robot data will never be sufficient in short term

Memory and reasoning as **data-efficient solutions**

- Memory is self-generated.
 - Reasoning transfers from vision–language data.
- 记忆推理不是骗投资人钱的大饼，可能是具身的必由之路**

- w/o memory & reasoning: 鼠目寸光 (49年入国军)
- w/o physical grounding: 眼高手低 (挥泪斩马谕)



Some Thoughts on Memory & Reasoning in VLA



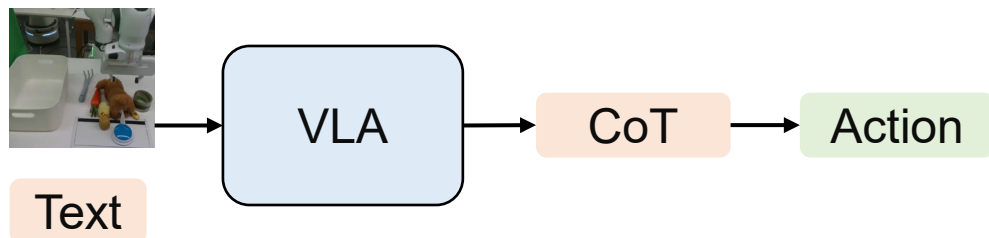
How VLA Performs Effective Reasoning?

1 Reasoning Paradigm of VLA



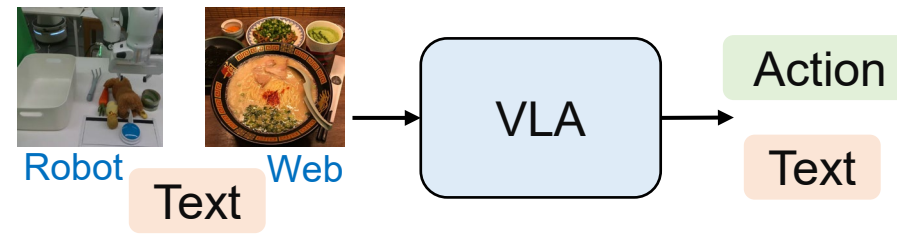
CoT

Huge gap between VL and A
ECoT, CoT-VLA, FSD, ...



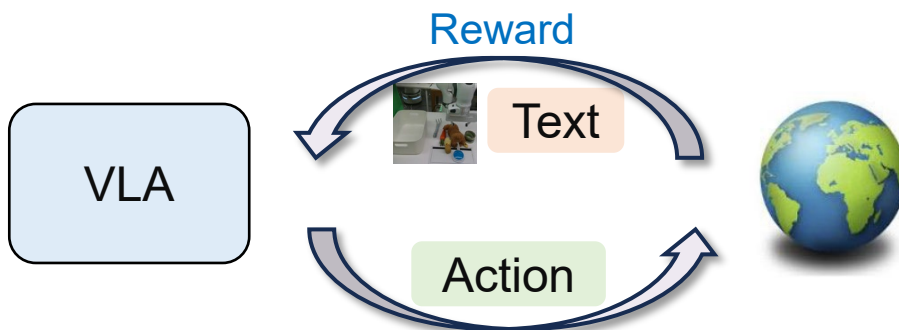
Co-train

VLA forgot the reasoning ability of VLM
PI-0.5, GR-3, ChatVLA, ...



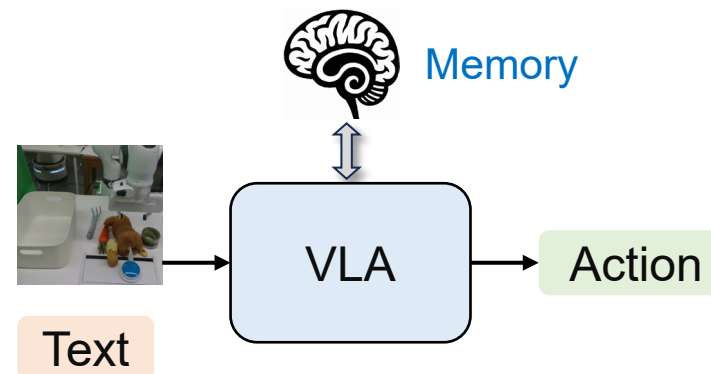
RL

Exploring Physical World
ConRFT, VLA-RL, Simple VLA-RL



Memory

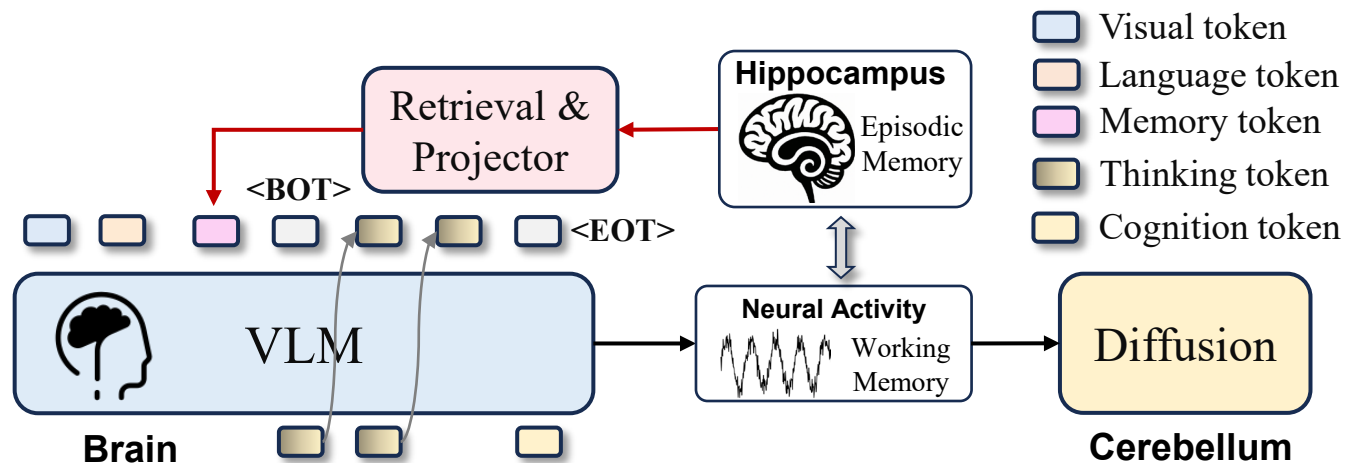
Current decision depend on history
MemoryVLA, Long-context DP, ...





What Are The Future Direction for Memory

2 Memory Reflection



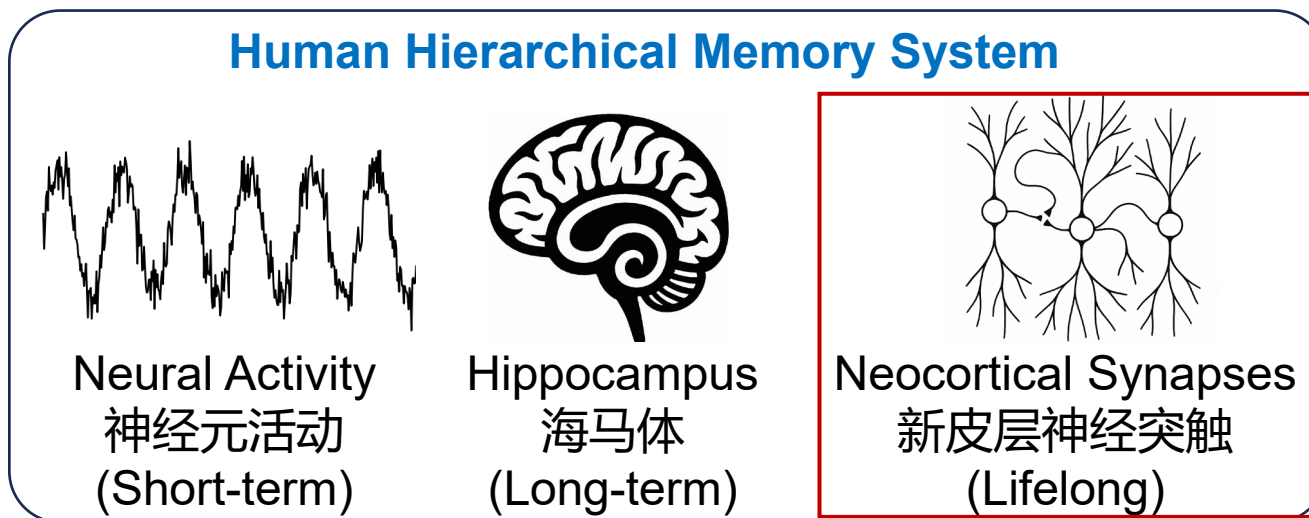
Memory Reflection

- ❑ Memory need **reflection**.
- ❑ Retrieve relevant memories and project them into the LLM input space for chain-of-thought reasoning.

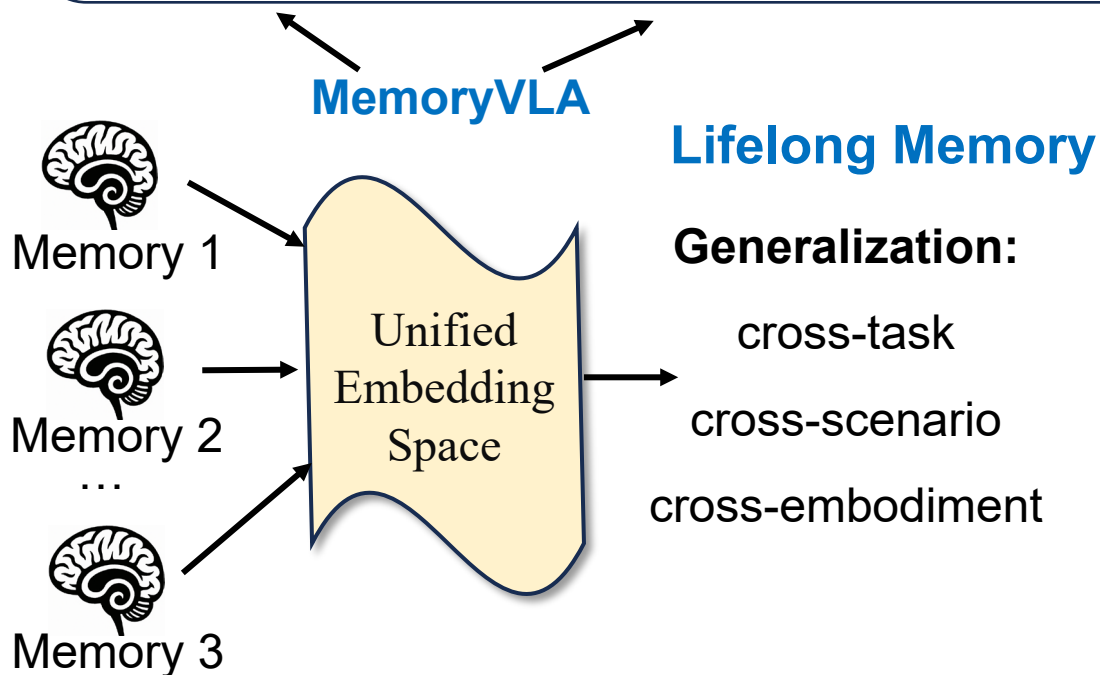
Embedding-space CoT

- ❑ Humans may not speak when thinking.
- ❑ N CoT steps $\approx N$ extra implicit layers, boosting model depth and capacity.

3 Lifelong Memory



Human: 吃一堑, 长一智



Spurious Recall

- ❑ Overabundant memory retrieval brings in unrelated noise, causing false recall.
- ❑ Stronger retrieval mechanism.

VLA + External Memory vs World Model?

4 Reasoning Paradigm for VLA



- ❑ **Memory: self-generated additional prior** for temporally dependencies.
 - Future Direction: Memory Reflection, Lifelong Memory
- ❑ **CoT: mid-representations bridging the gap** between Vision-Language and Action
 - It does **NOT** enhance reasoning, but provides **additional supervision**. (ECoT)
 - My similar discovery: SimplerEnv + CogACT setting

CoT-v2	AVG
L + bridge ft + cotv2 + 测试输出	58.3
L + bridge ft + cotv2 + 测试不输出	57.3

- VLA analysis from the perspective of mid-representations

A Survey on Vision-Language-Action Models: An Action Tokenization Perspective

Yifan Zhong^{1,2*}, Fengshuo Bai^{2*}, Shaofei Cai^{1,2}, Xuchuan Huang^{1,2}, Zhang Chen^{1,2}, Xiaowei Zhang^{1,2}, Yuanfei Wang^{2,3}, Shaoyang Guo^{1,2}, Tianrui Guan^{1,2}, Ka Nam Lui^{1,2}, Zhiquan Qi^{1,2}, Yitao Liang^{1,2}, Yuanpei Chen^{1,2†} and Yaodong Yang^{1,2†}

¹Institute for AI, Peking University, ²PKU-PsiBot Joint Lab, ³School of Computer Science, Peking University

- [1] <https://embodied-cot.github.io>
- [2] <https://ecot-lite.github.io>
- [3] <https://arxiv.org/abs/2507.01925>

4 Reasoning Paradigm for VLA



□ Memory

□ CoT

□ **Co-train:** Cross-domain Data Co-train

- High cost of real-robot data collection.
- VLA forgets the reasoning ability of VLM.
- **Which data is noisy? Which is useful?**

□ **RL:** Learn through exploring the world, beyond the training dataset.

- Pre-training: Adaptive utilization of suboptimal data.
- Post-training: Explore OOD cases beyond training set, and align anything.
- Test-time compute
- Bottleneck: Real-world RL. (Data efficiency, sim2real gap, ...)

Essential analysis of RLVR in VLM

Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?

Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning

[1] <https://arxiv.org/abs/2504.13837>

[2] <https://arxiv.org/abs/2506.01939>

5 Ultimate Goal



How to autonomously evolve intelligence in open-world exploration?

Closed-loop Feedback

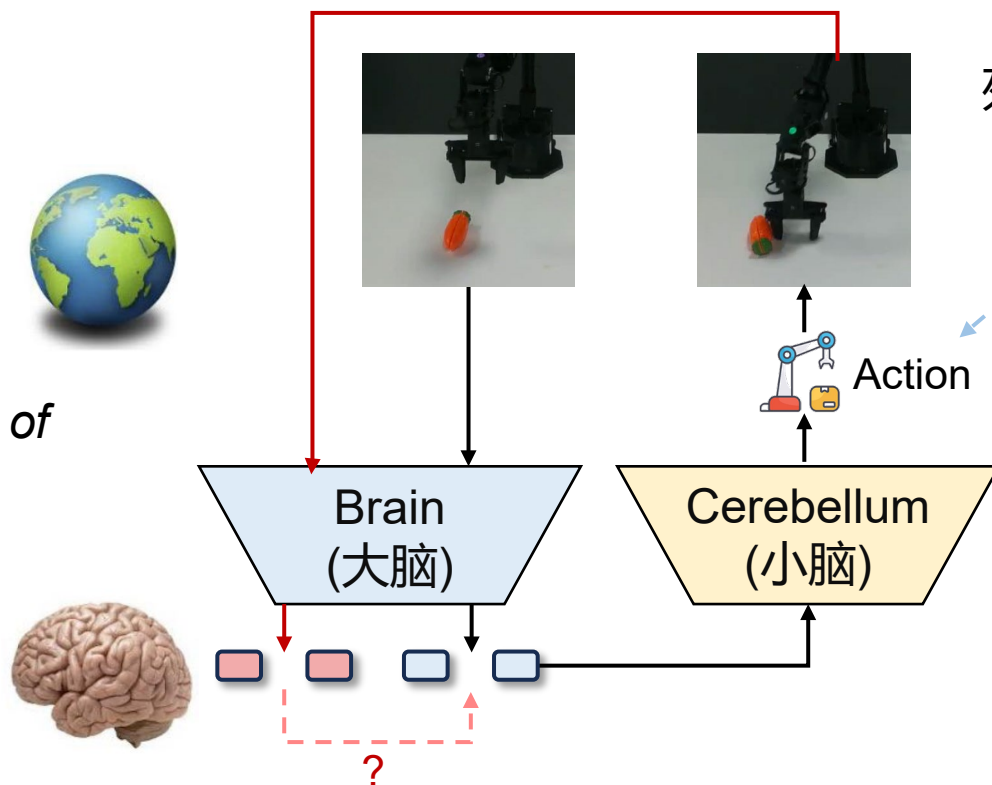


Yi Ma, HKU

"Pursuing the Nature of Intelligence"

<https://www.bilibili.com/video/BV1FtQzYUEyo>

$$\text{Knowledge} = \int_0^t \text{Intelligence},$$
$$\text{Intelligence} = \frac{d}{dt} \text{Knowledge}.$$



死记硬背训练集中的知识

Human is not like that
Human have no BP

Human: closed-loop feedback

Closed-loop Feedback is Promising for Autonomous Intelligence



哈哈哈

安徽 安庆



扫一扫上面的二维码图案，加我为朋友。

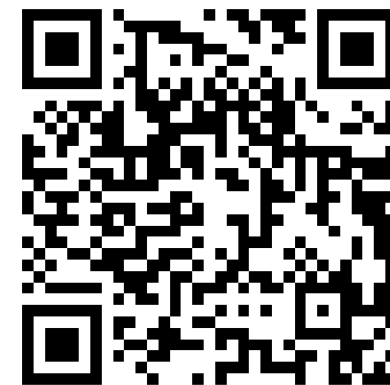
Thank You!

Hao Shi (石昊)

shi-h23@mails.tsinghua.edu.cn



Homepage



Paper