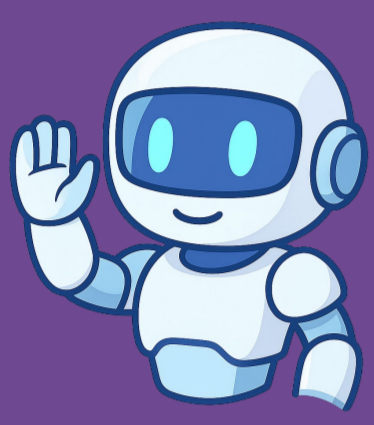


# Your Robots Need Memory!



# MemoryVLA

One of the earliest works to explicitly model memory in VLA

See More

Homepage (Video)

Official Code (200+ Star)

Dexbotic Code (900+ Star)

Contact

shihao1895@gmail.com

u1s11024

(石昊)

# 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, Gao Huang



## 1 Motivation

### ① Why Does Memory Matter in VLA?

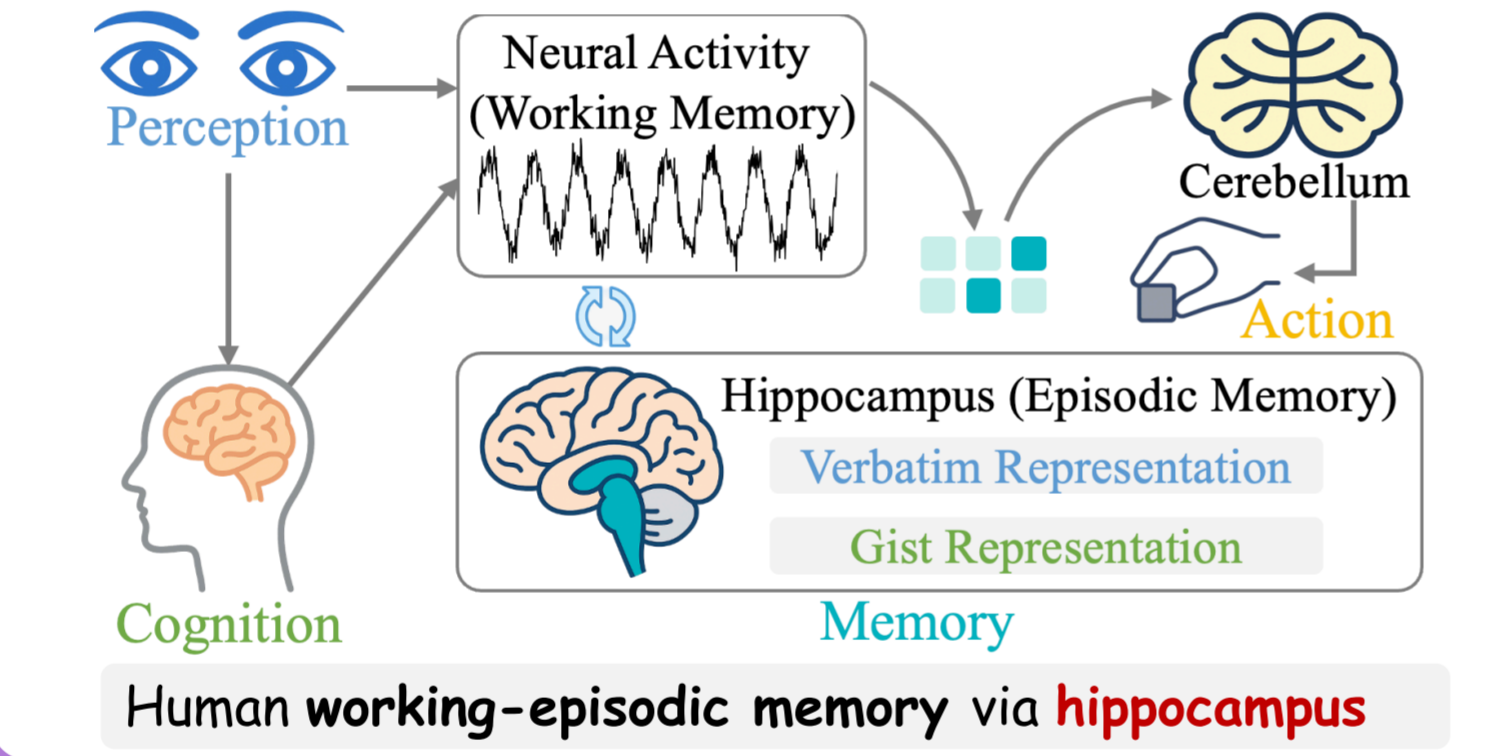
**Example of Temporal Confusion**

Clean an object    Ready to press    Have I already pressed button?

Miss!    Repeated!

Clean one object, then press the button once to count it.

### ② How Humans Do It?



### ④ Insight

Before press    After press

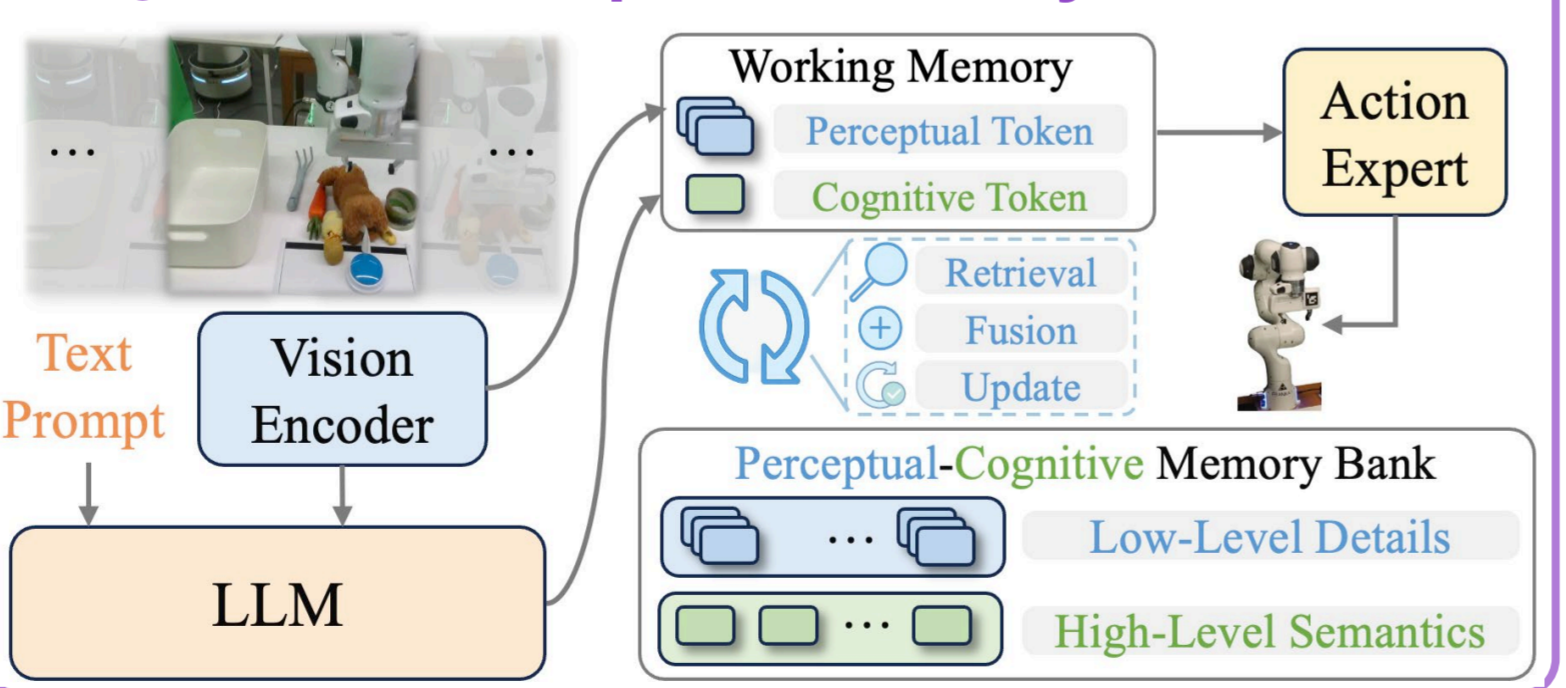
Same Observation    Different Actions

Typic VLA is **One-to-Many Mapping**

Observation    Action

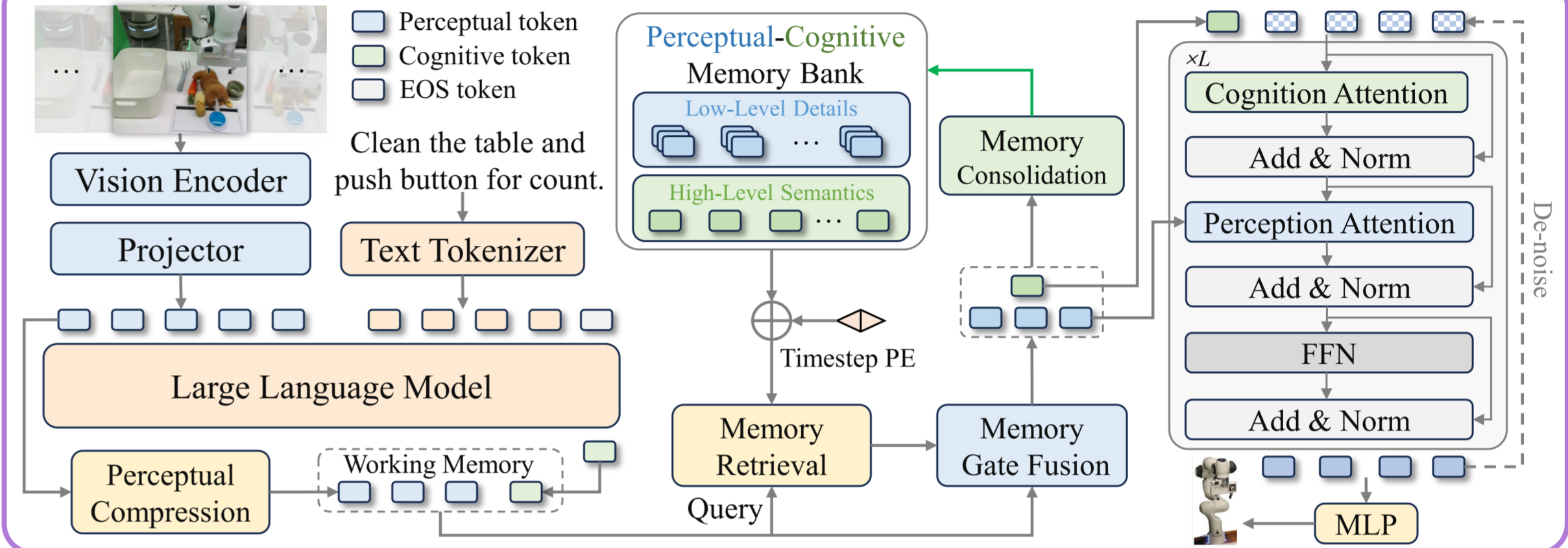
More robot data  
More label conflicts  
Harder scale up

### ③ Human-Inspired Memory for VLA



## 2 Method

### Overall framework



### Module Details

**(a) Memory Retrieval**  
Select past info relevant to current decision

**(b) Memory Gate Fusion**  
Adaptive fusion of past and current info

**(c) Memory Consolidation**  
Merge nearby & similar entries for compact memory

## 3 Experiments

Benchmark	Metric	SpatialVLA	OpenVLA	Ours	Improvement
SimplerEnv	SpatialVLA	42.7	57.3	71.9	+14.6
	CogACT	57.3	68.4	71.9	+3.5
	$\pi_0$	68.4	71.9	71.9	0
LIBERO	SpatialVLA	75.9	83.1	96.5	+3.3
	$\pi_0$ -FAST	83.1	93.2	96.5	+3.3
	CogACT	93.2	96.5	96.5	0
Mikasa-Robo	SpatialVLA	21.0	28.4	41.2	+11.8
	CogACT	28.4	29.4	41.2	+11.8
	$\pi_0$	29.4	41.2	41.2	0
Real-world-General	SpatialVLA	31	72	85	+9
	CogACT	72	76	85	+9
	$\pi_0$	76	85	85	0
Real-world-Temporal	SpatialVLA	9	52	83	+26
	CogACT	52	57	83	+26
	$\pi_0$	57	83	83	0

3 Robots, 6 Benchmarks, 150+ Tasks, 500+ Variations

**Future Direction**

- Hierarchical Memory for Efficient Retrieval
- Memory Reflection for Reasoning
- Lifelong Memory Across Episodes, Environments, and Embodiments
- Improved Memory Consolidation
- Memory with World Models