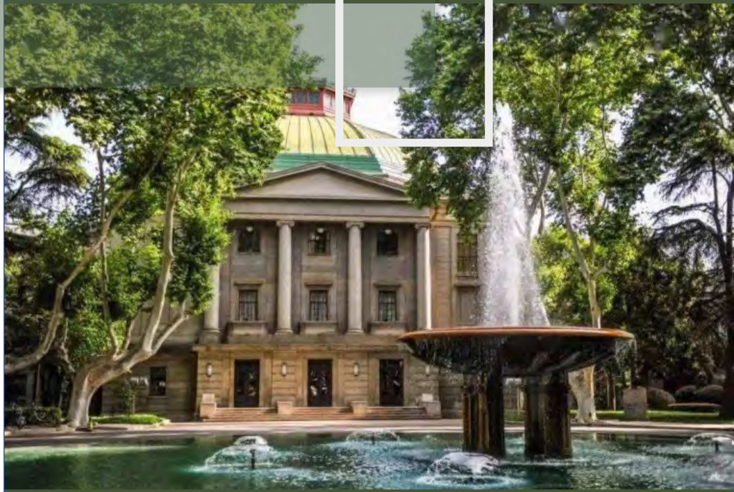
東南大學
SOUTHEAST UNIVERSITY

Learning to Remember: Exploring Multimodal Memory Mechanisms in Long Video Understanding

汇报人：李文卓

東南大學
SOUTHEAST UNIVERSITY

CONTENTS



SOUTHEAST UNIVERSITY

PART ONE

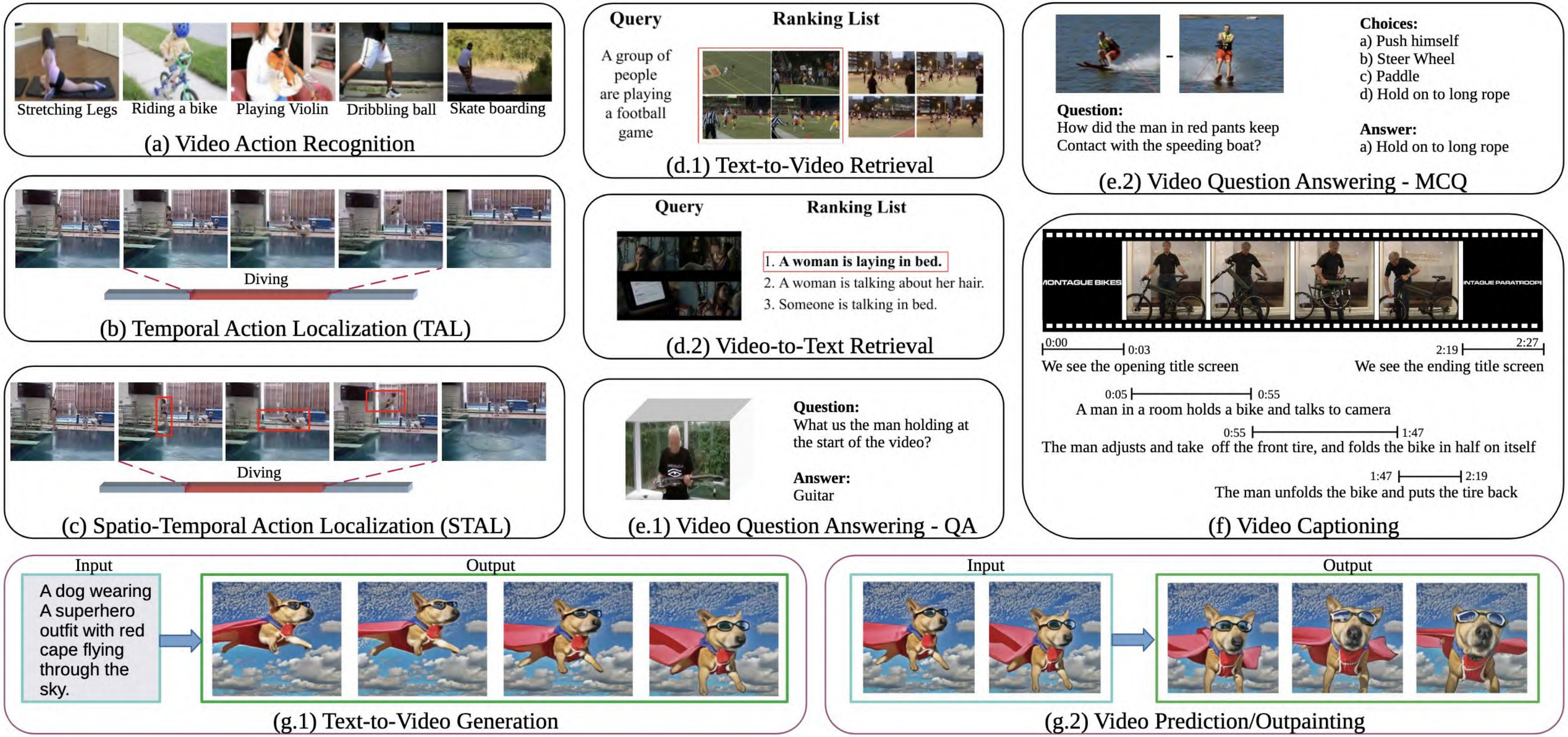
Introduction

— Why Long Video Understanding is Hard?

01

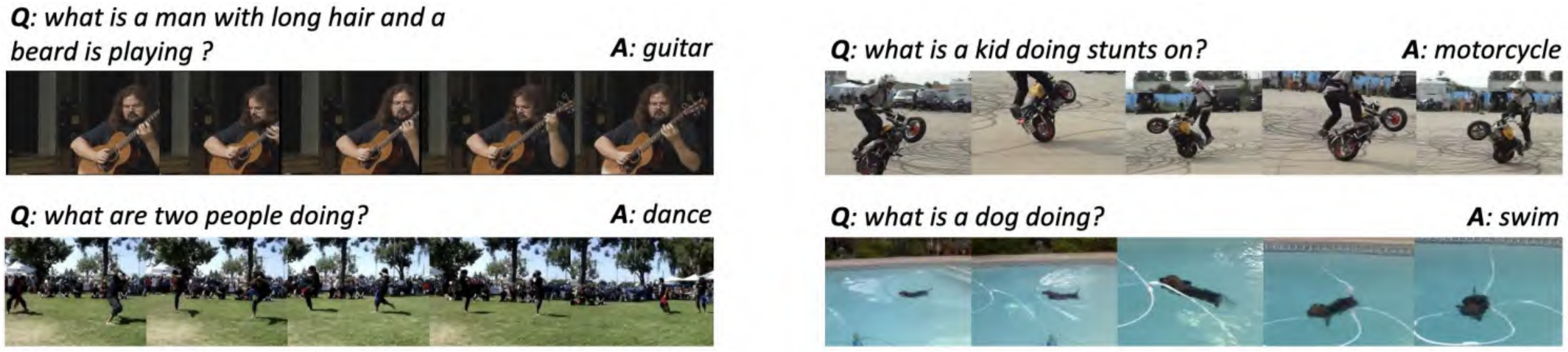
Introduction

Video Understanding: Action Recognition, Event Detection, **Video Question Answering**, Video Summarization etc.



Introduction

Short Video Understanding: eg. MSVD-QA

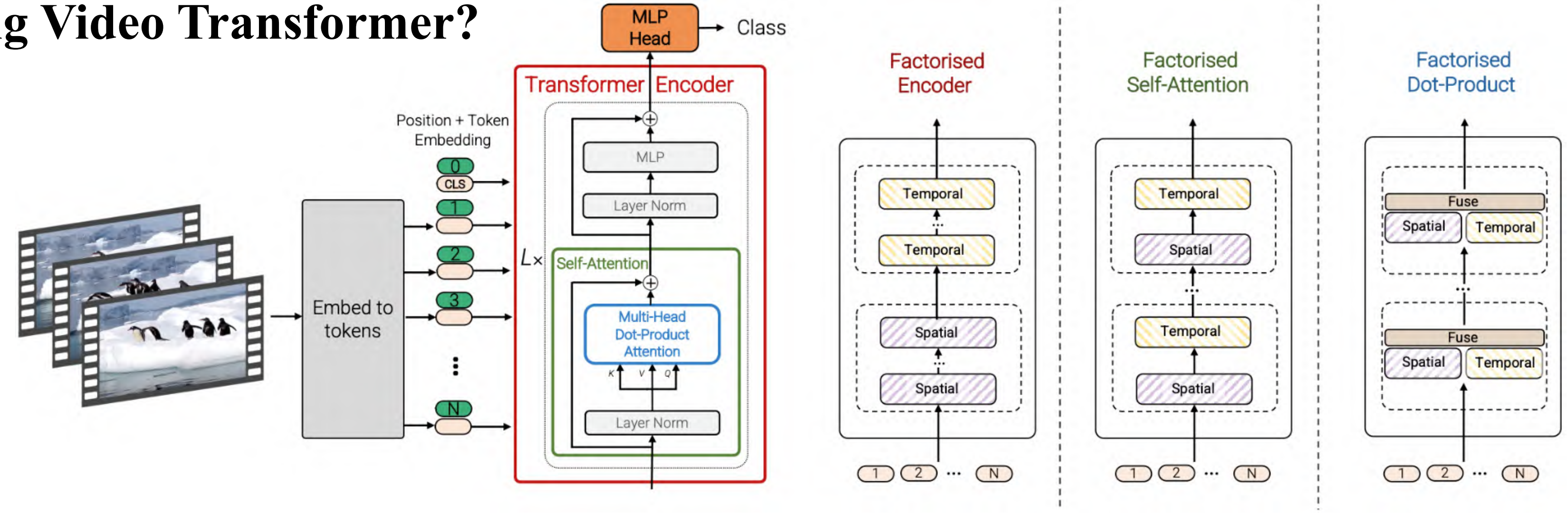


Long Video Understanding: eg. Video-MME

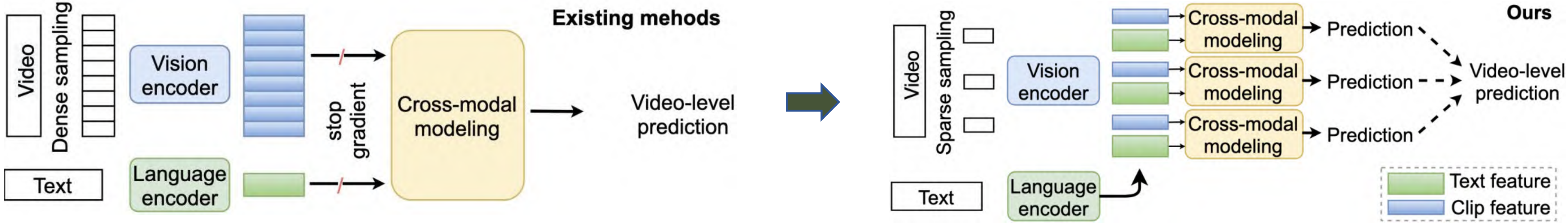
Introduction

Can we utilize the existing Video Transformer?

ViViT:



- **Simply use it: → Self-attention all tokens O(n²), computationally and memory Intolerant**
- **Sample: Extracting keyframes or key clip from long videos → Inevitable information loss**



- **Hierarchical structure: model locally, then aggregate globally. → Tradeoff: Information retention VS Complexity**

Introduction

The difficulties in understanding long videos

Modeling long temporal dependencies is difficult
- several minutes or even hours
- traditional Transformer or RNN are difficult to directly capture such long cross-segment dependencies.

Complex multimodal information fusion
- Video semantics span multiple modalities (vision scenes, character dialogue, audio cues)
- Required **cross-modal alignment and fusion capabilities.**

Information redundancy and semantic sparsity
- large number of frames lacking effective information and key events being sparsely distributed
- core challenge: how to filter and focus on important segments.

Event-level understanding and Computational efficiency
- Event-level action, causal, and intent reasoning required
- Global modeling over tens of thousands of frames is computationally expensive which limits the direct use of global attention methods.

Remember longer and more accurately

东南大学
SOUTHEAST UNIVERSITY

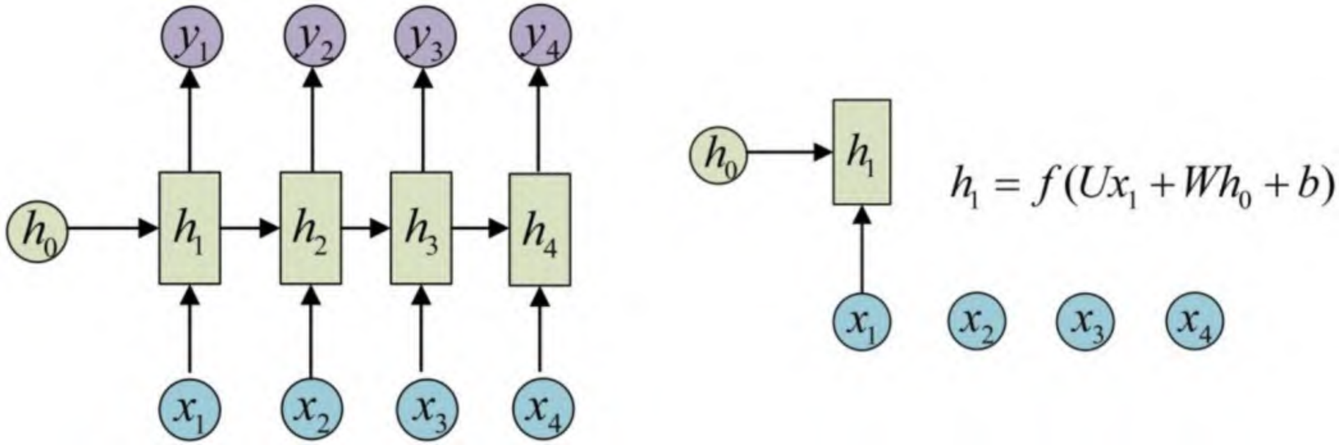
PART TWO

Problem Definition
— What Does “Learning to Remember” Mean?

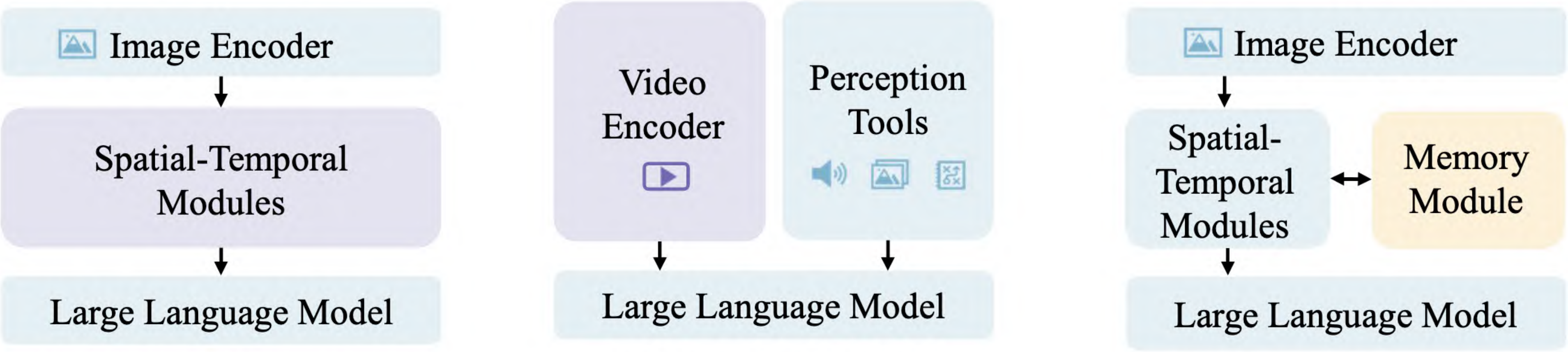
02

Problem Definition

Internal memory eg. RNN hidden states, Transformer cache: Implicitly maintain state



External memory eg. Memory Bank, Key-Value Memory: Explicitly store, update, and retrieve features



(a) Video MLLMs with Spatial-Temporal Modules

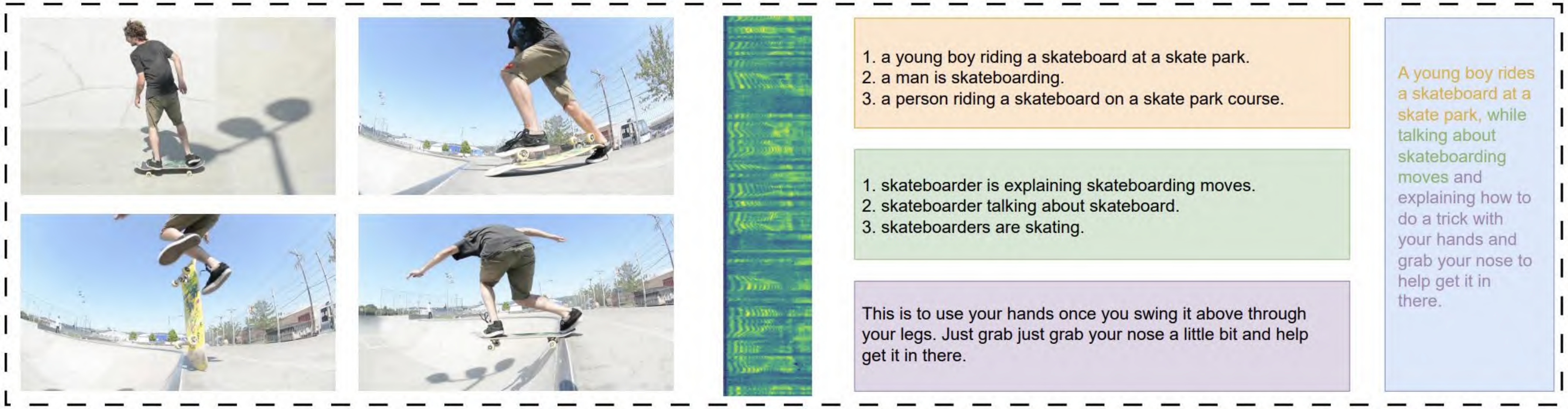
(b) Video MLLMs with Perception Tools

(c) Our Approach for Long Video Understanding

Problem Definition

Key points of memory mechanism

Inputs: Long video sequence + multimodal signal (frames, audio, subtitles)



Goal: Learn the most useful information summary within a limited context window

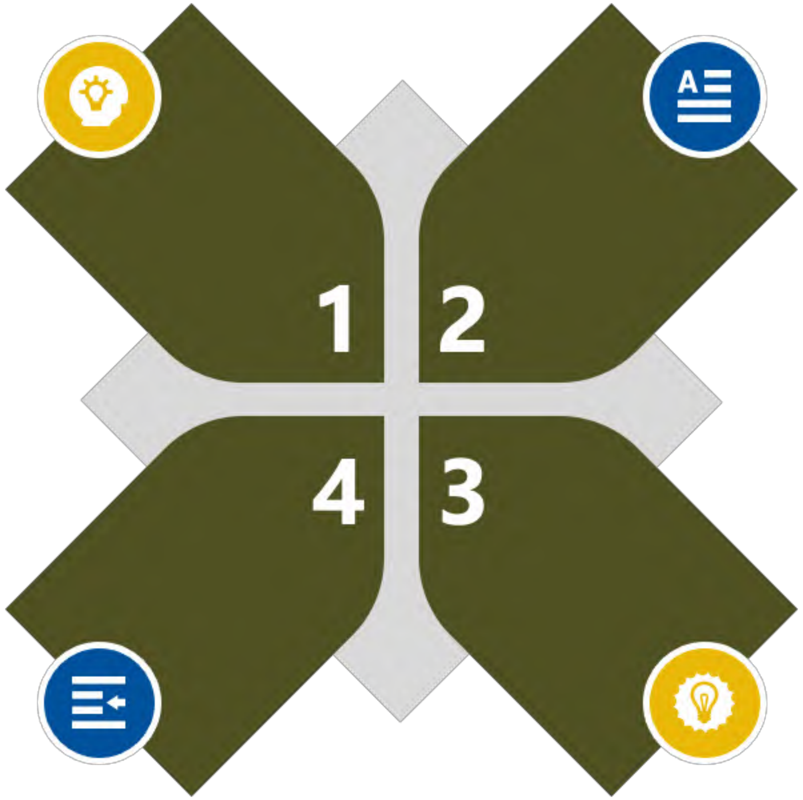
The model needs to have:

- Selective encoding（哪些片段值得记？）
- Efficient storage（如何组织记忆？）
- Dynamic retrieval（如何在问答/推理中召回？）

Problem Definition

Role of Memory Mechanisms in Long Videos

Short-term retention & context continuity:
- Preserves key information over seconds to minutes, maintaining semantic coherence.
- Avoid Catastrophic forgetting over time.



Long-term integration & global reasoning
- Integrates events across long time spans to capture global narrative and causal structure.
- Helping the model capture the global narrative structure and causal relationships

Redundancy filtering & key event selection
- Selectively stores key events
- Reducing computation on irrelevant segments.
- Improve efficiency and robustness

Cross-modal alignment & semantic fusion
- Serves as a shared space to align and fuse visual, textual, and audio cues at the event level.
- Facilitate the model's multimodal understanding at the event level.

SOUTHEAST UNIVERSITY

PART THREE

Architectures

— Representative Multimodal Memory Designs

03

東南大學
SOUTHEAST UNIVERSITY

PART THREE

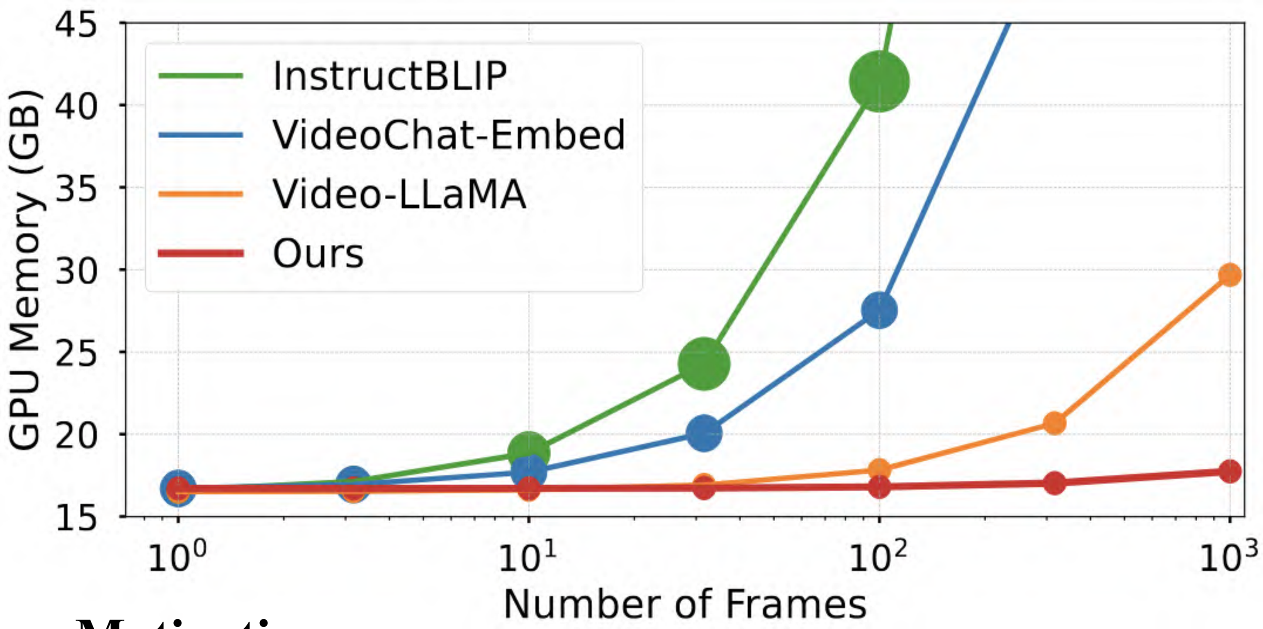
Architectures
—Feature Space Memory

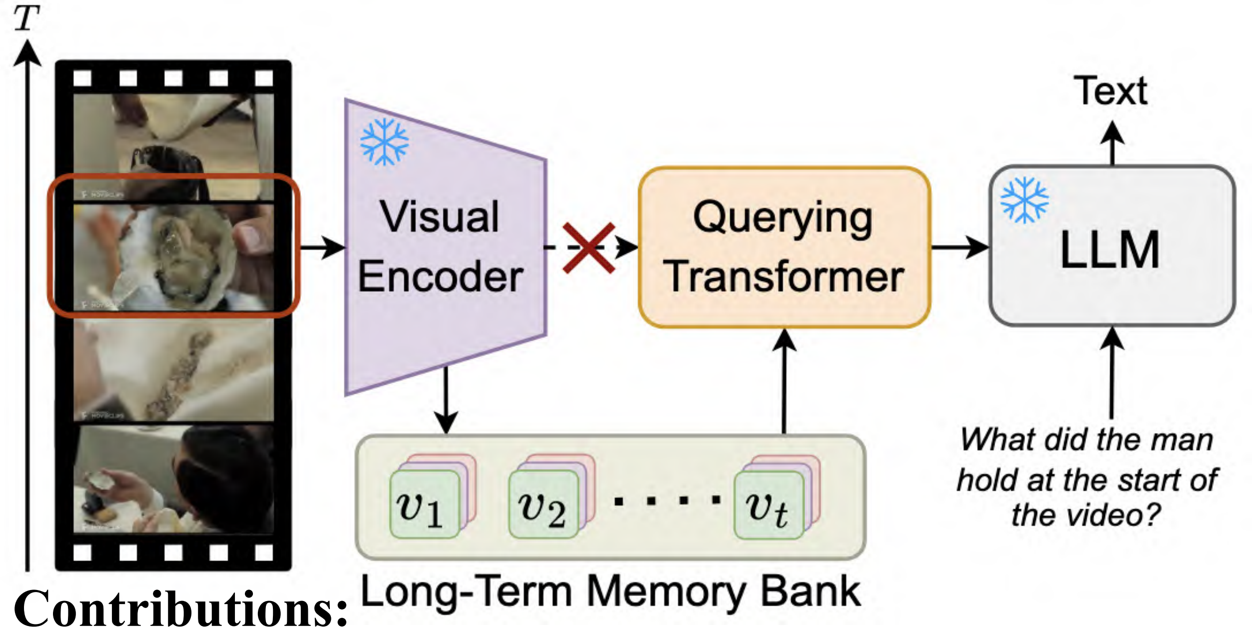
3.1

Architectures

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

Link: https://arxiv.org/pdf/2404.05726 (**CVPR2024**)





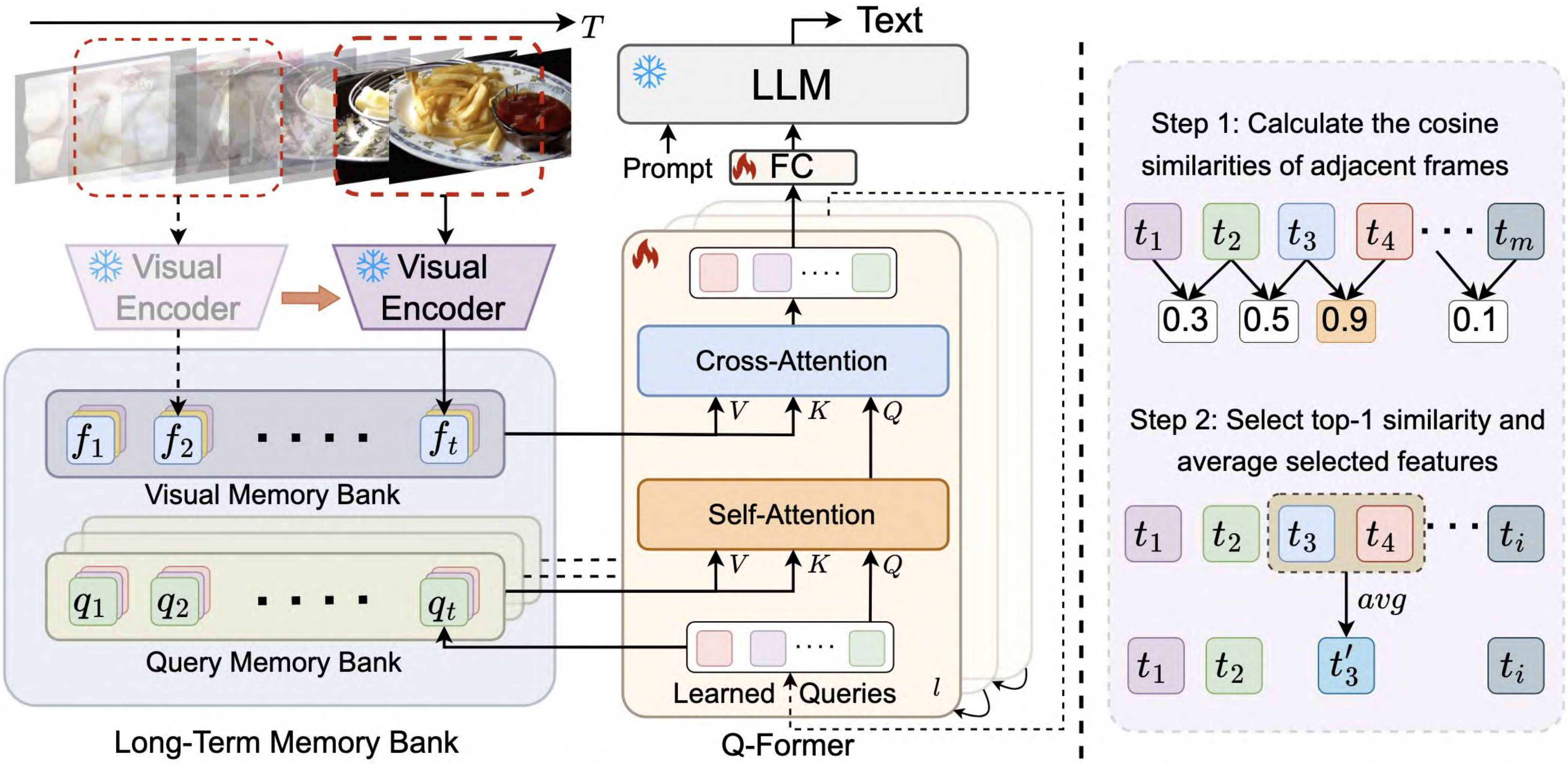
Motivation:

- Existing LLM-based multimodal models (e.g., Video-LLaMA, VideoChat) Limited by context length and GPU memory
- Only process a small number of frames
- Primarily suited for short video understanding

Contributions:

- Proposes a novel long-term memory bank that can be seamlessly integrated into existing large multimodal models, enabling long-video modeling.
- Processes video streams in an online manner, significantly reducing GPU memory usage and effectively alleviating LLM context-length limitations.

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding



(a) Framework Overview

(b) Memory Bank Compression

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

Comparison with state-of-the-art methods on the **LVU dataset: top-1** and <u>top-2</u>

Model	Content			Metadata				Avg
	Relation	Speak	Scene	Director	Genre	Writer	Year	
Obj_T4mer [29]	54.8	33.2	52.9	47.7	52.7	36.3	37.8	45.0
Performer [39]	50.0	38.8	60.5	58.9	49.5	48.2	41.3	49.6
Orthoformer [40]	50.0	38.3	66.3	55.1	55.8	47.0	43.4	50.8
VideoBERT [41]	52.8	37.9	54.9	47.3	51.9	38.5	36.1	45.6
LST [32]	52.5	37.3	62.8	56.1	52.7	42.3	39.2	49.0
VIS4mer [32]	57.1	40.8	67.4	62.6	54.7	48.8	44.8	53.7
S5 [33]	**67.1**	<u>42.1</u>	<u>73.5</u>	<u>67.3</u>	**65.4**	<u>51.3</u>	<u>48.0</u>	<u>59.2</u>
Ours	<u>58.2</u>	**44.8**	**80.3**	**74.6**	<u>61.0</u>	**70.4**	**51.9**	**63.0**

Comparison on the **Breakfast and COIN** datasets: The top-1 accuracy

Model	Breakfast	COIN
TSN [44]	-	73.4
VideoGraph [45]	69.5	-
Timeception [28]	71.3	-
GHRM [46]	75.5	-
D-Sprv. [47]	89.9	90.0
ViS4mer [32]	88.2	88.4
S5 [33]	<u>90.7</u>	<u>90.8</u>
Ours	**93.0**	**93.2**

Comparison with state-of-the-art methods on the **video question answering task**: Top-1 accuracy

Model	MSRVTT	MSVD	ActivityNet
JustAsk [60]	41.8	47.5	38.9
FrozenBiLM [61]	47.0	54.8	43.2
SINGULARITY [62]	43.5	–	44.1
VIOLETv2 [63]	44.5	54.7	–
GiT [64]	43.2	56.8	–
mPLUG-2 [65]	<u>48.0</u>	58.1	–
UMT-L [66]	47.1	55.2	47.9
VideoCoCa [67]	46.3	56.9	**56.1**
Video-LLaMA [12]	46.5	<u>58.3</u>	45.5
Ours	**48.5**	**60.6**	<u>49.8</u>

Comparison with state-of-the-art methods on the **video captioning task: METEOR (M) and CIDEr (C)**

Model	MSRVTT		MSVD		YouCook2	
	M	C	M	C	M	C
UniVL [68]	28.2	49.9	29.3	52.8	–	127.0
SwinBERT [69]	29.9	53.8	41.3	120.6	15.6	109.0
GIT [64]	32.9	73.9	**51.1**	**180.2**	<u>17.3</u>	<u>129.8</u>
mPLUG-2 [65]	**34.9**	**80.3**	48.4	165.8	–	–
VideoCoca [67]	–	73.2	–	–	–	128.0
Video-LLaMA	32.9	71.6	49.8	175.3	16.5	123.7
Ours	<u>33.4</u>	<u>74.6</u>	<u>51.0</u>	<u>179.1</u>	**17.6**	**131.2**

東南大學
SOUTHEAST UNIVERSITY

PART THREE

Architectures
— Video Caption Memory

3.2

Architectures

VideoLucy: Deep Memory Backtracking for Long Video Understanding

Link: https://arxiv.org/abs/2510.12422 (**NeurIPS 2025**)

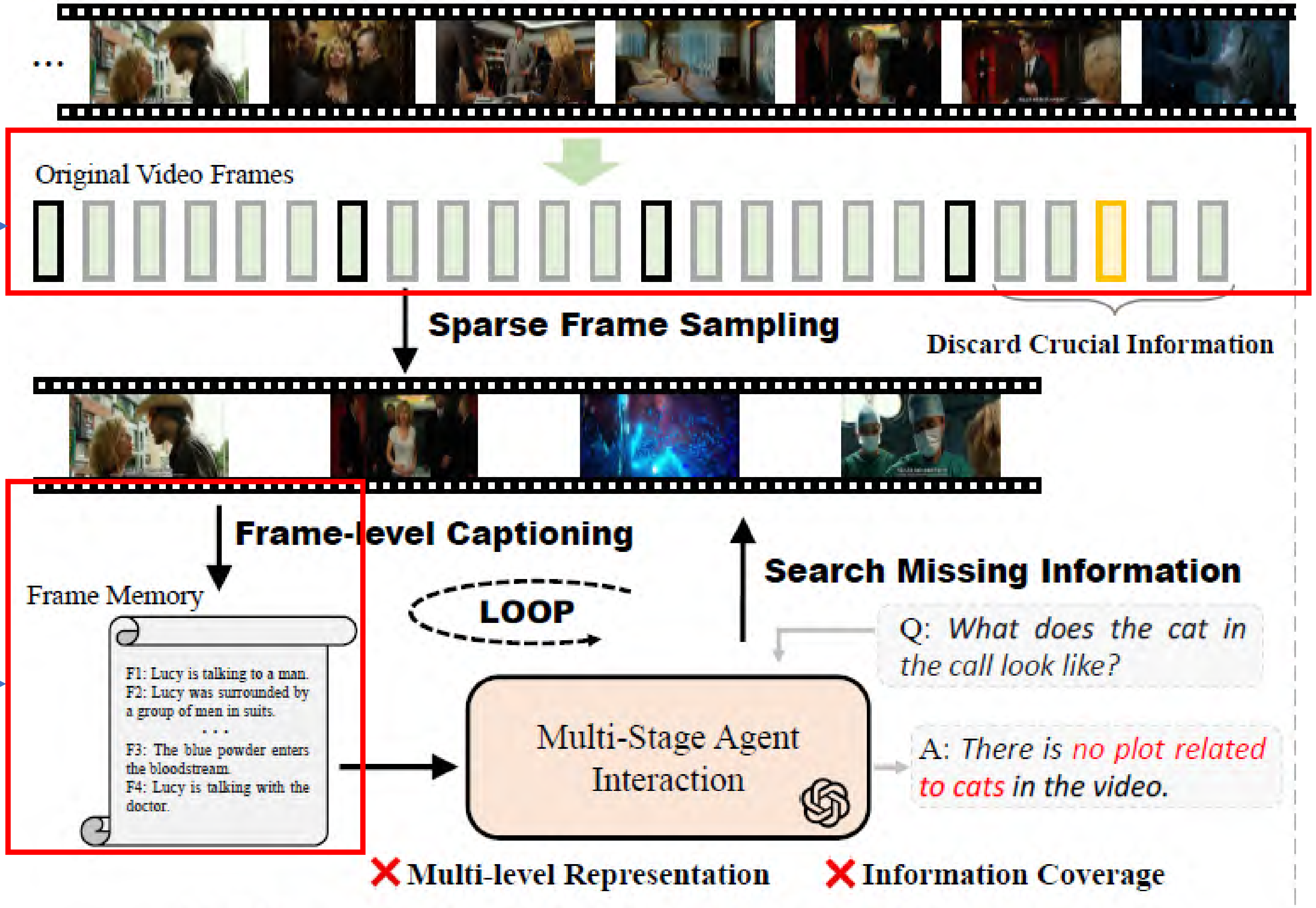
Prior agent-based video understanding methods suffer from two key limitations

- **Modeling and inference based on a single frame**
  - Difficult to capture the temporal contextual information of **consecutive frames**.
  - Essentially, this approach utilizes a **pre-trained captioning model** to generate **text descriptions** for each specified frame in the video
  - Using a **large language model** as the core, an iterative information search loop is constructed to **obtain** keyframes related to the problem and their supplementary descriptions.
- **Sparse frame sampling strategy,** To reduce the cost of generating dense frame-level subtitles, but obviously carries the risk of losing critical information.

VideoLucy: Deep Memory Backtracking for Long Video Understanding

Prior agent-based video understanding methods suffer from two key limitations

- **Sparse frame sampling strategy**
- **Modeling and inference based on a single frame**



(a) A Representative of Existing Video Agent-based Systems

VideoLucy: Deep Memory Backtracking for Long Video Understanding

Innovation point

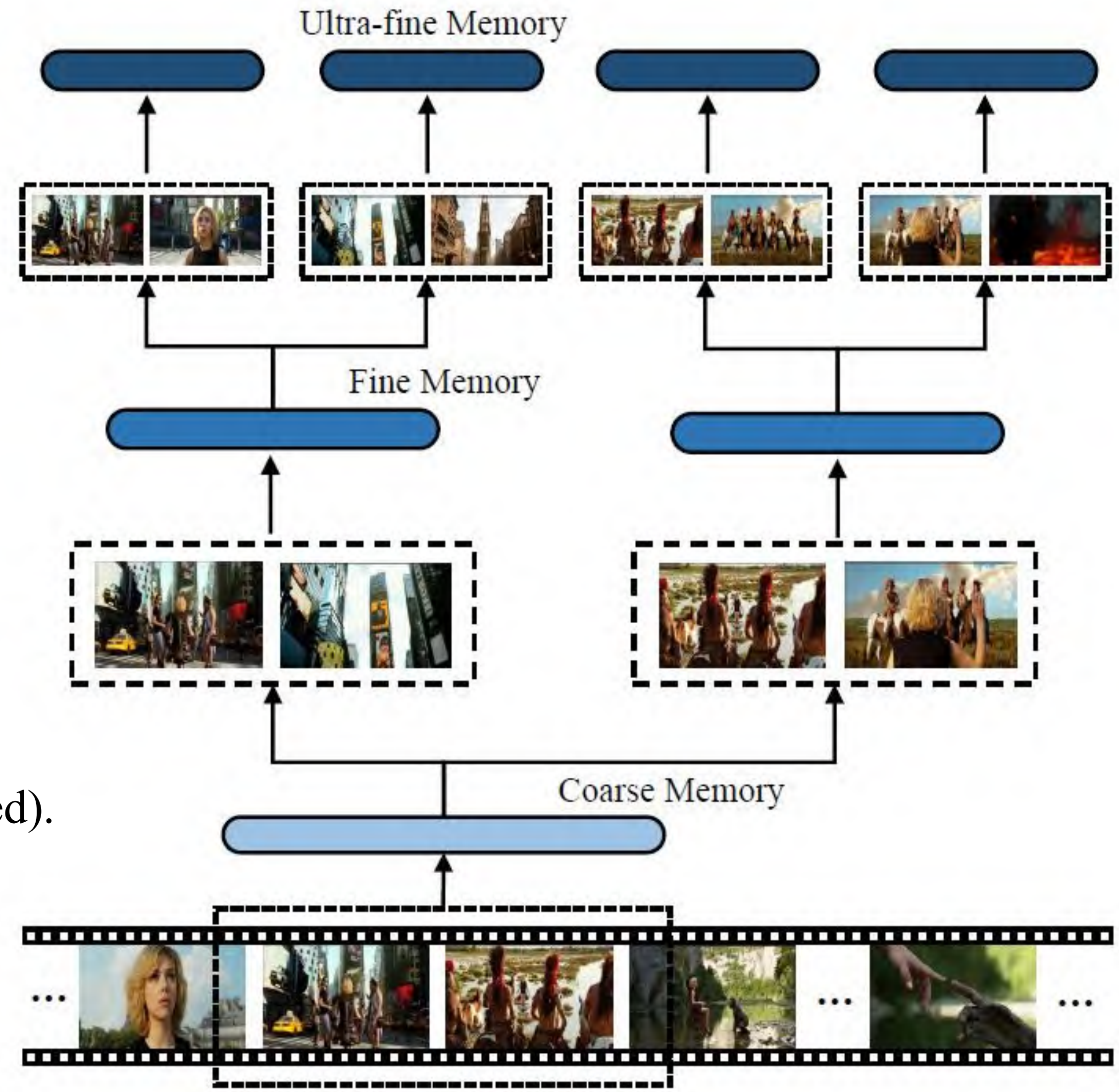
Hierarchical Memory Structure

For a video with **N frames**, it can be divided into **k non-overlapping sub-segments**, each containing N/k frames.

Then, based on the value of k, we can further divide the sub-segments into **three segments of different granularities**. MLLM can then be used to interpret these segments separately.

$$m_k = VidCap(v_k, p_k)$$

- k=1, understands the entire video (coarse-grained).
- k=m (1<m<N), understands each segment of the video (fine-grained).
- k=N, understands every frame of the video (ultra-fine-grained).

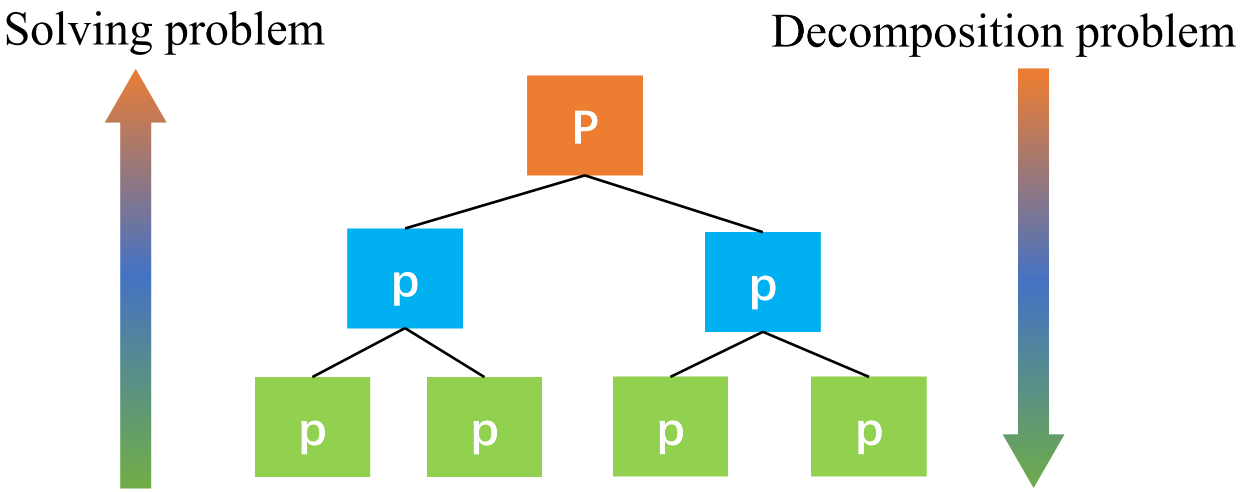


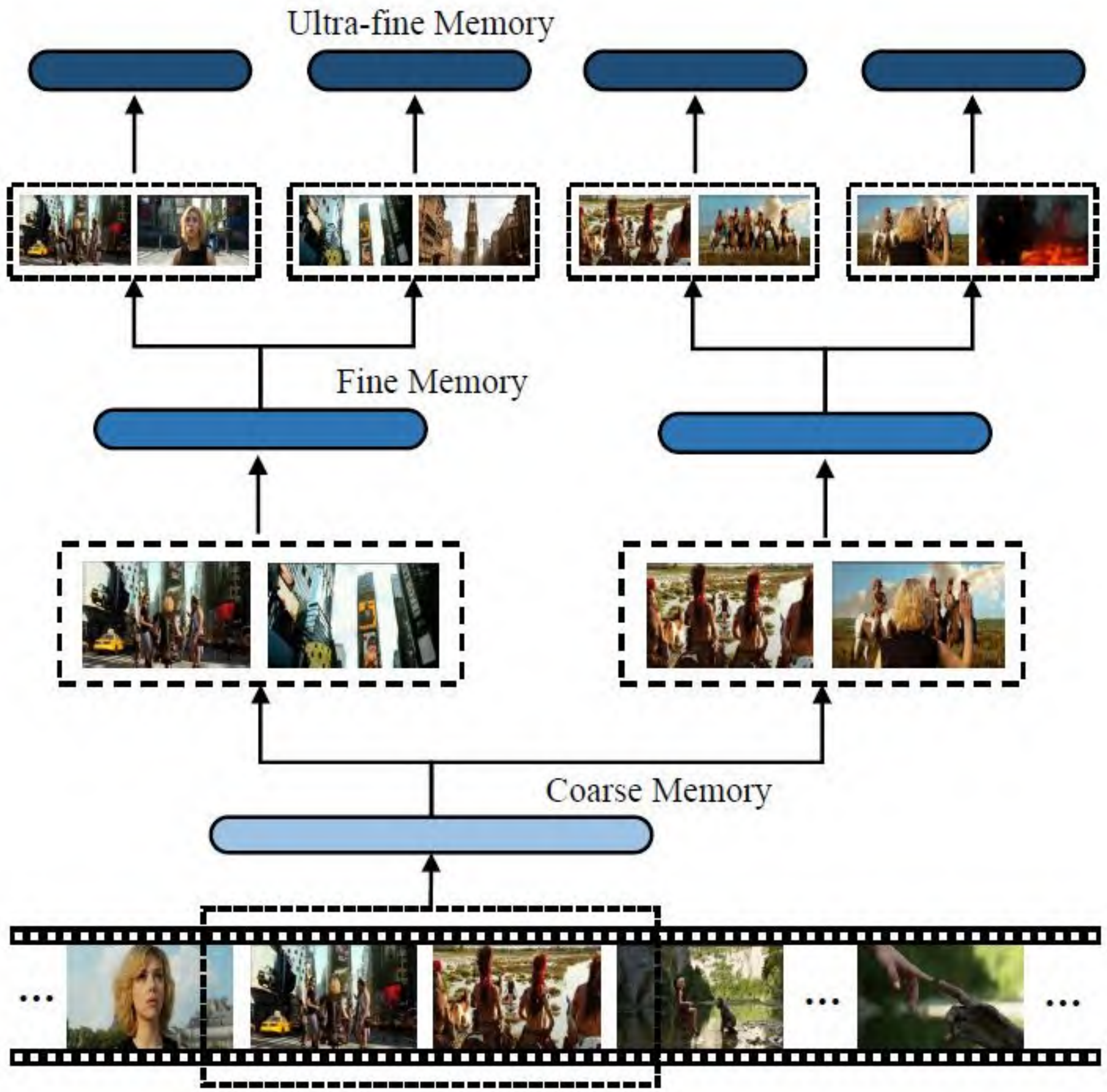
VideoLucy: Deep Memory Backtracking for Long Video Understanding

Innovation point

Hierarchical Memory Structure

Essentially, it's the **divide-and-conquer strategy**.
For a problem P, we first break it down into multiple smaller subproblems p from top to bottom, then solve each subproblem one by one from bottom to top, and finally combine them to solve the overall problem P.





VideoLucy: Deep Memory Backtracking for Long Video Understanding

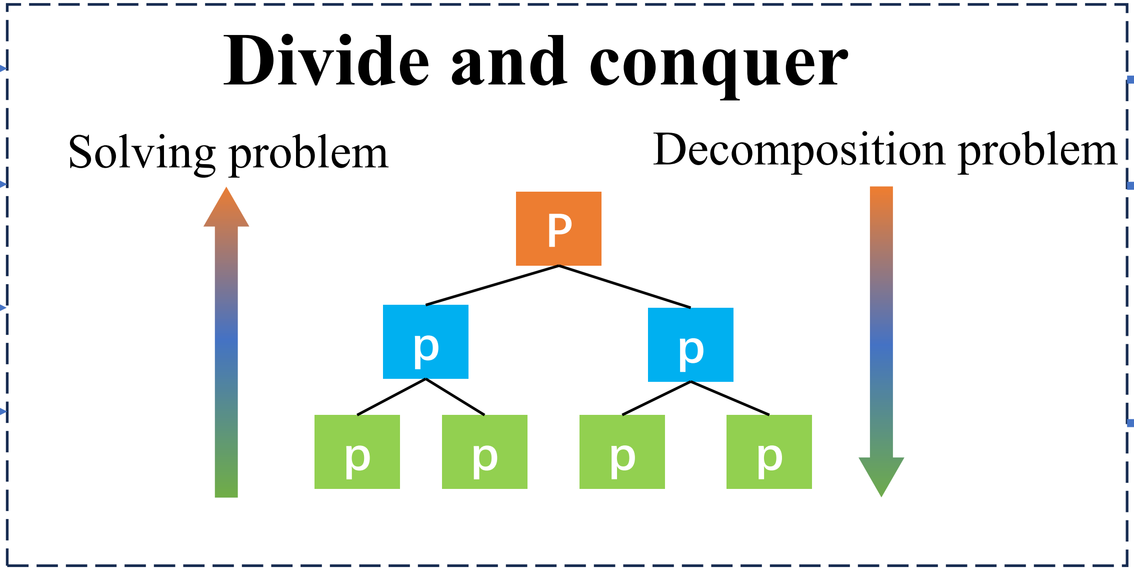
Array sorting

Attention

CNN

Video Memory

....



Merge Sort

Swin Transformer

Feature Pyramid Network

VideoLucy

VideoLucy: Deep Memory Backtracking for Long Video Understanding

Innovation point

Multi-agent system design

Localization Agent
Locating video clips related to the question

Captioning Agent
Based on the input video clip and prompt, provide the caption.

Instruction Agent
Design a Prompt for Captioning Agent

Answering Agent
Answer the question based on your current exploration and memory.

Algorithm 1 The Iterative Backtracking Mechanism

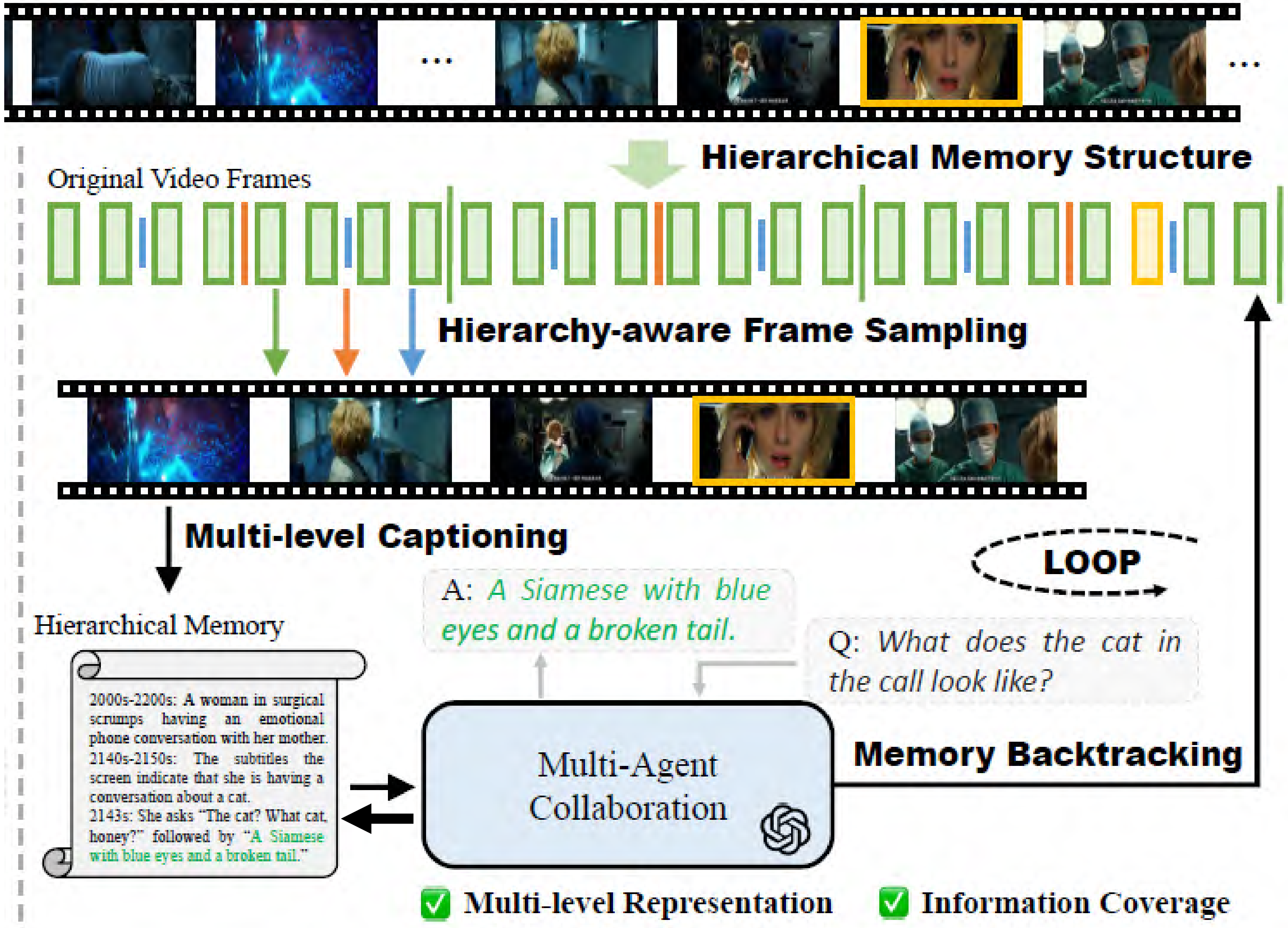
Input: The video V, question Q, captioning agent $CapAGT$, localization agent $LocAGT$, instruction agent $InsAGT$, answering agent $AnsAGT$ and the specified temporal scopes T_c, T_f, T_{uf} corresponding to coarse, fine, and ultra-fine memory.

```
Implement sparse coarse memory initialization to obtain an initial current memory list CM.
Initialize a relevant set of time periods S_rt = {}.
Obtain the response based on the current memory R = AnsAGT(CM, Q).
while R is not confident do
    Locate the single most question-relevant time period t = LocAGT(CM \ S_rt, Q) not in S_rt.
    Add this time period t to the relevant set S_rt, i.e., S_rt ← S_rt ∪ {t}.
    Analyze missing question-key info and provide instruction prompt p = InsAGT(CM, Q, t).
    Obtain the video clip V_t corresponding to this period t from the video V.
    Divide V_t into short clips{(t^i, V_t^i)}_{i=1}^L by T_d, |t| = T_c ⇒ T_d = T_f, |t| = T_f ⇒ T_d = T_uf.
    Obtain the updated current-depth memory of this time period m_c = CapAGT(V_t, p).
    Obtain the deeper memories of this time period {m_d^i}_{i=1}^L = {CapAGT(V_t^i, p)}_{i=1}^L.
    Update CM: CM ← CM ∪ {(t, m_c)} ∪ {(t^i, m_d^i) | i = 1, ···, L}.
    Obtain the response based on the updated current memory R = AnsAGT(CM, Q).
end while
```

Output: The final response R with a confident answer to the question Q.

VideoLucy: Deep Memory Backtracking for Long Video Understanding

Method Overview



(b) VideoLucy (Ours)

Limitations

1. **Hyperparameter sensitivity**

- the parameter K for segmenting video clips needs to be manually specified.
- While the experiments in the paper achieved SOTA results, different hyperparameters are required for different datasets to achieve the corresponding SOTA performance.

2. **The reasoning expense is too high.**

東南大學
SOUTHEAST UNIVERSITY

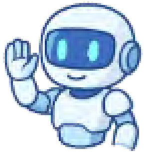
PART FOUR

Frontier Field Integration

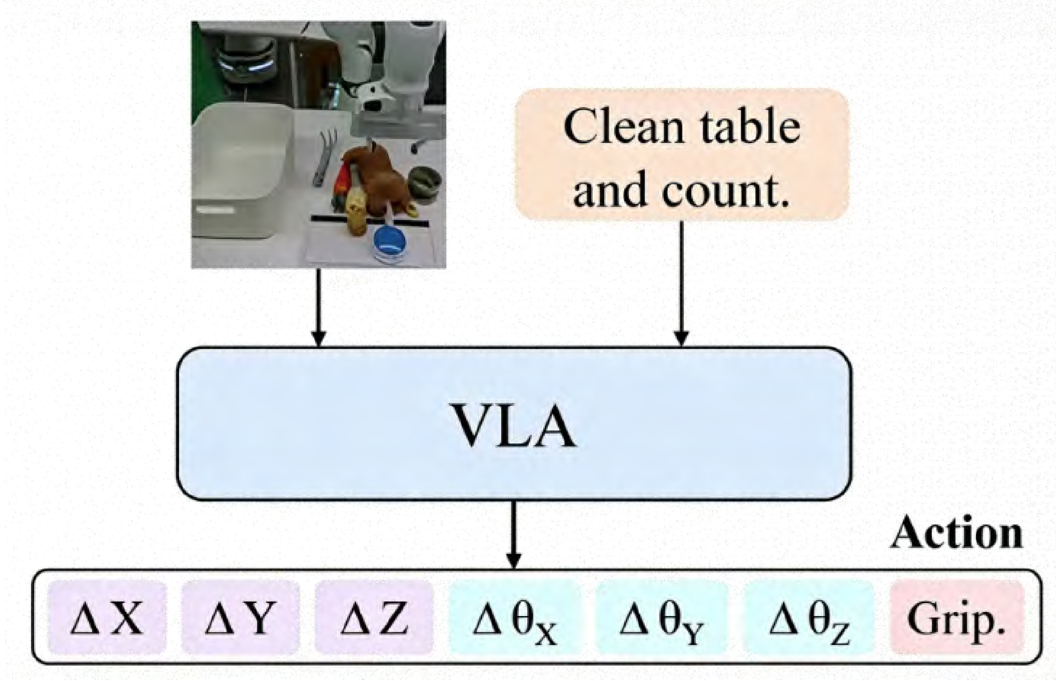
— Application In VLA

04

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

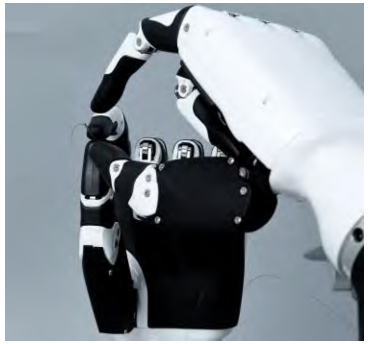


Link: https://arxiv.org/abs/2508.19236 (**OpenReview ICLR 2026**)



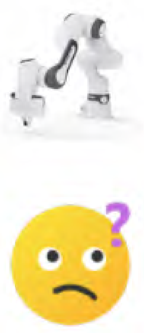
Physical Grounding

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



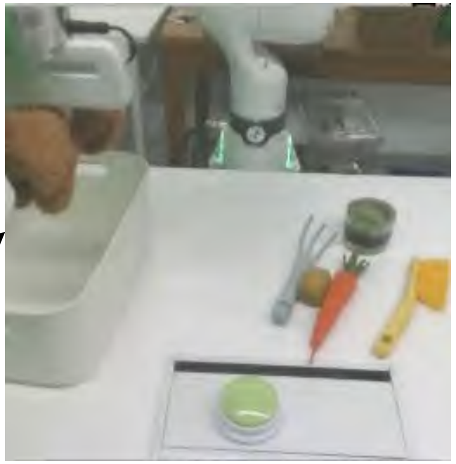
Is Good Physical Grounding Enough?

Clean Table & Count

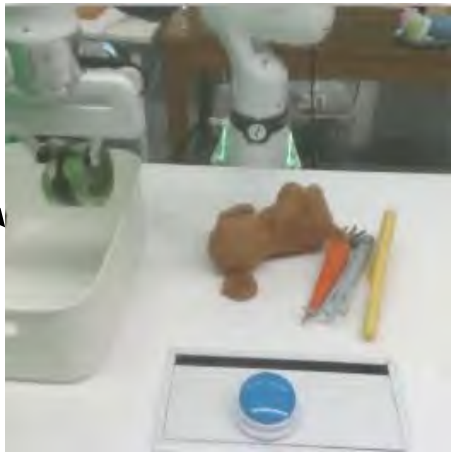


Have I pressed it before?

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



Miss

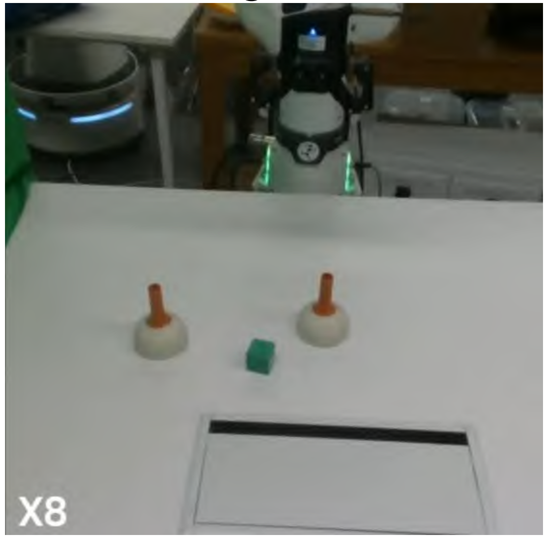


Repeat

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

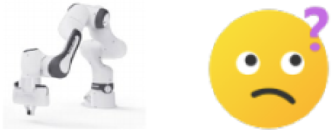
Change Food

Change Food

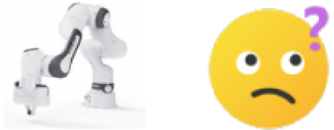


Guess Where

Guess Where



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



Which cup is the block really under?

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

Clean Table & Count

Change Food

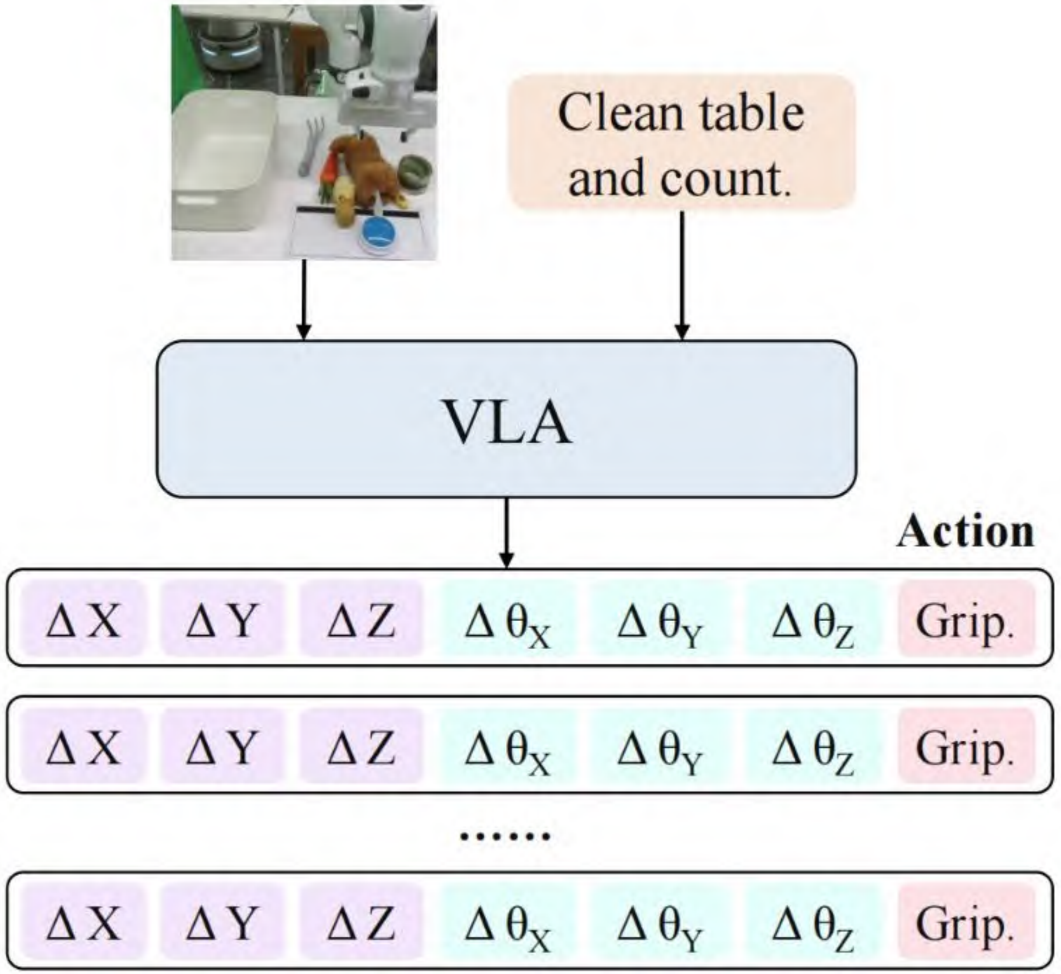
Guess Where

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.

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



Spatial

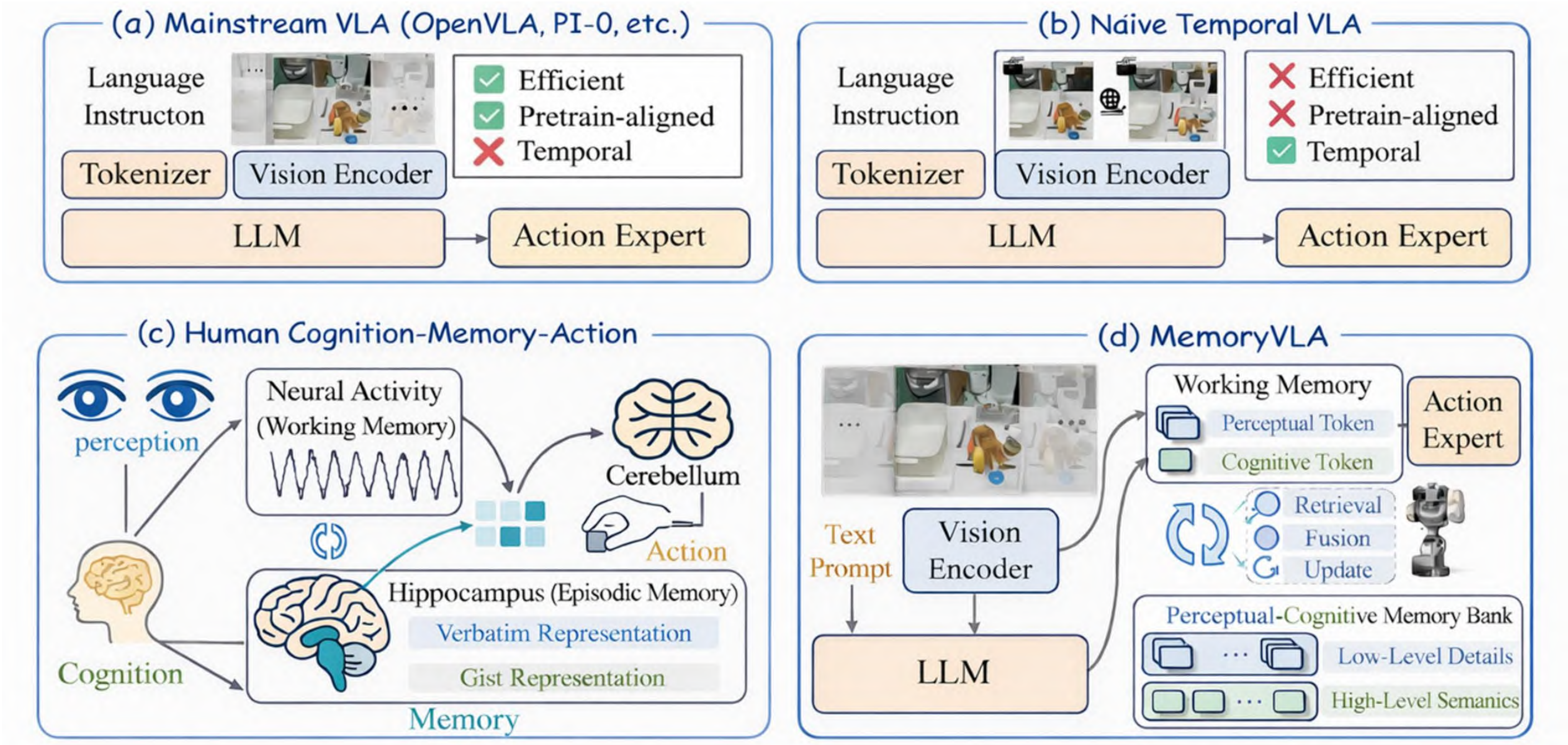
Physical Grounding
Mapping sensor observation and instruction into **6-DoF pose**, with physical-world properties in mind.

Temporal

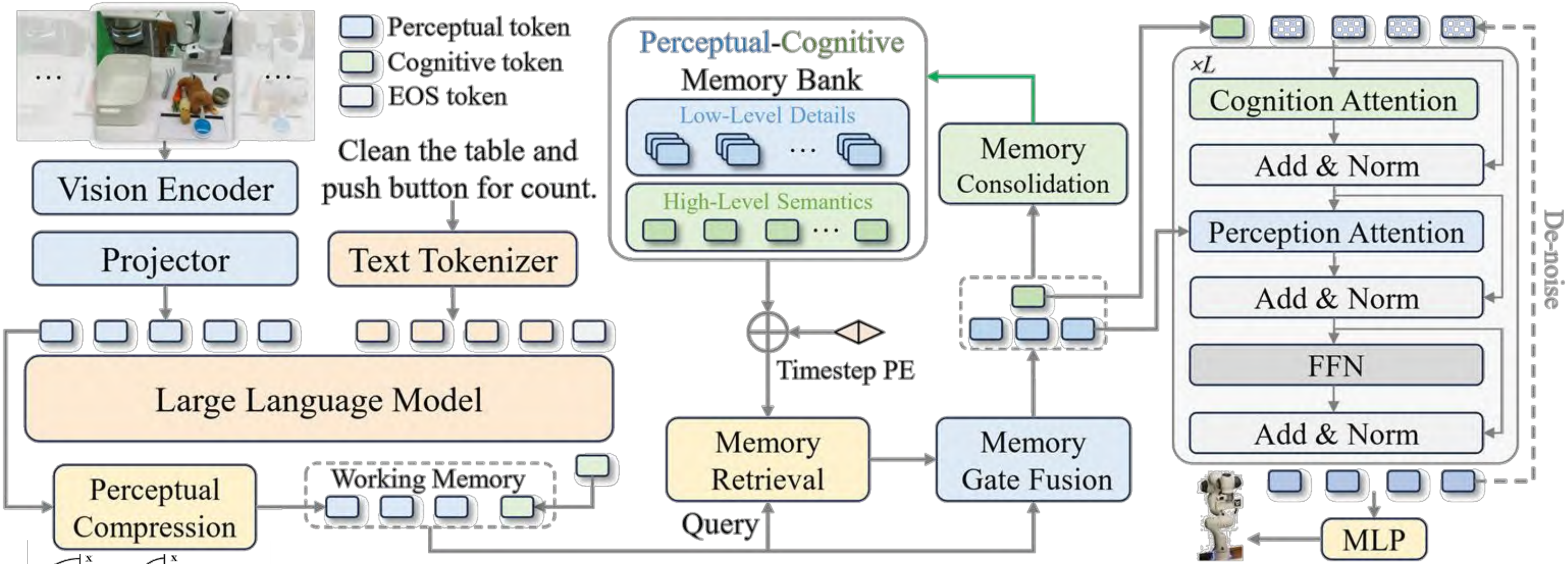
Sequential Decision-Making
Making a series of decisions based on current and historical states to achieve long-term objectives

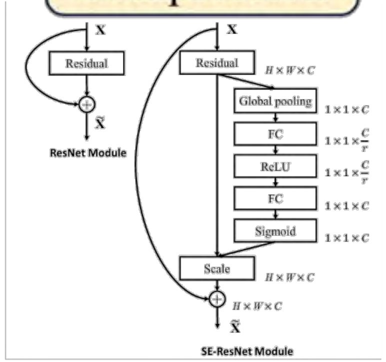
MemoryVLA: Perceptual-Cognitive Memory in Vision-Language-Action Models for Robotic Manipulation

How to Capture Temporal Dependencies?



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

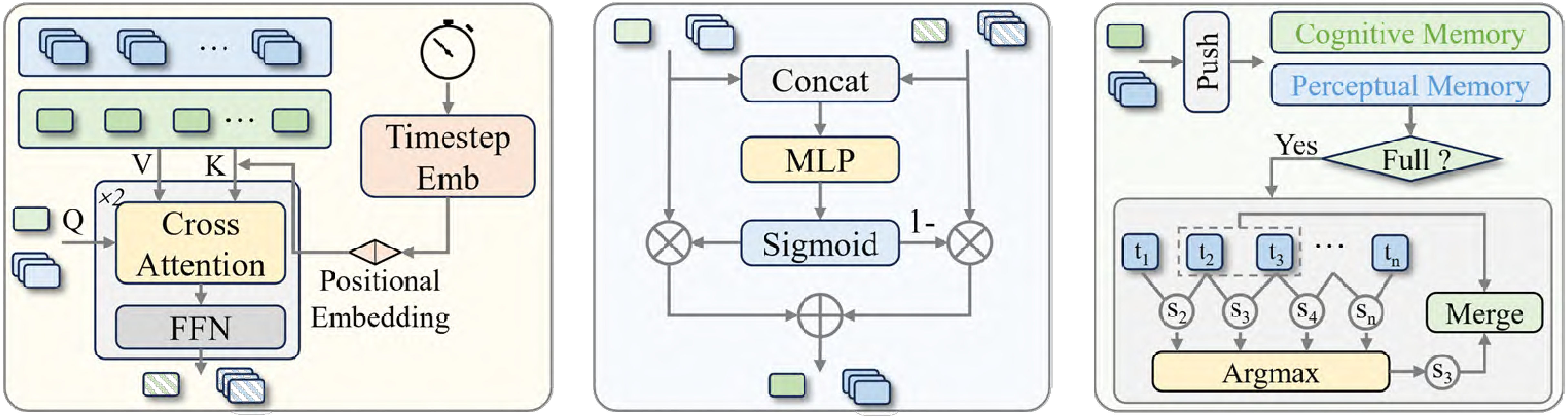




Backbone: Llama-7B/Qwen 2.5-7B, OXE pretrained

Acion Expert: DiT-L

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



(a) Memory Retrieval **(b) Memory Gate Fusion** **(c) Memory Consolidation**

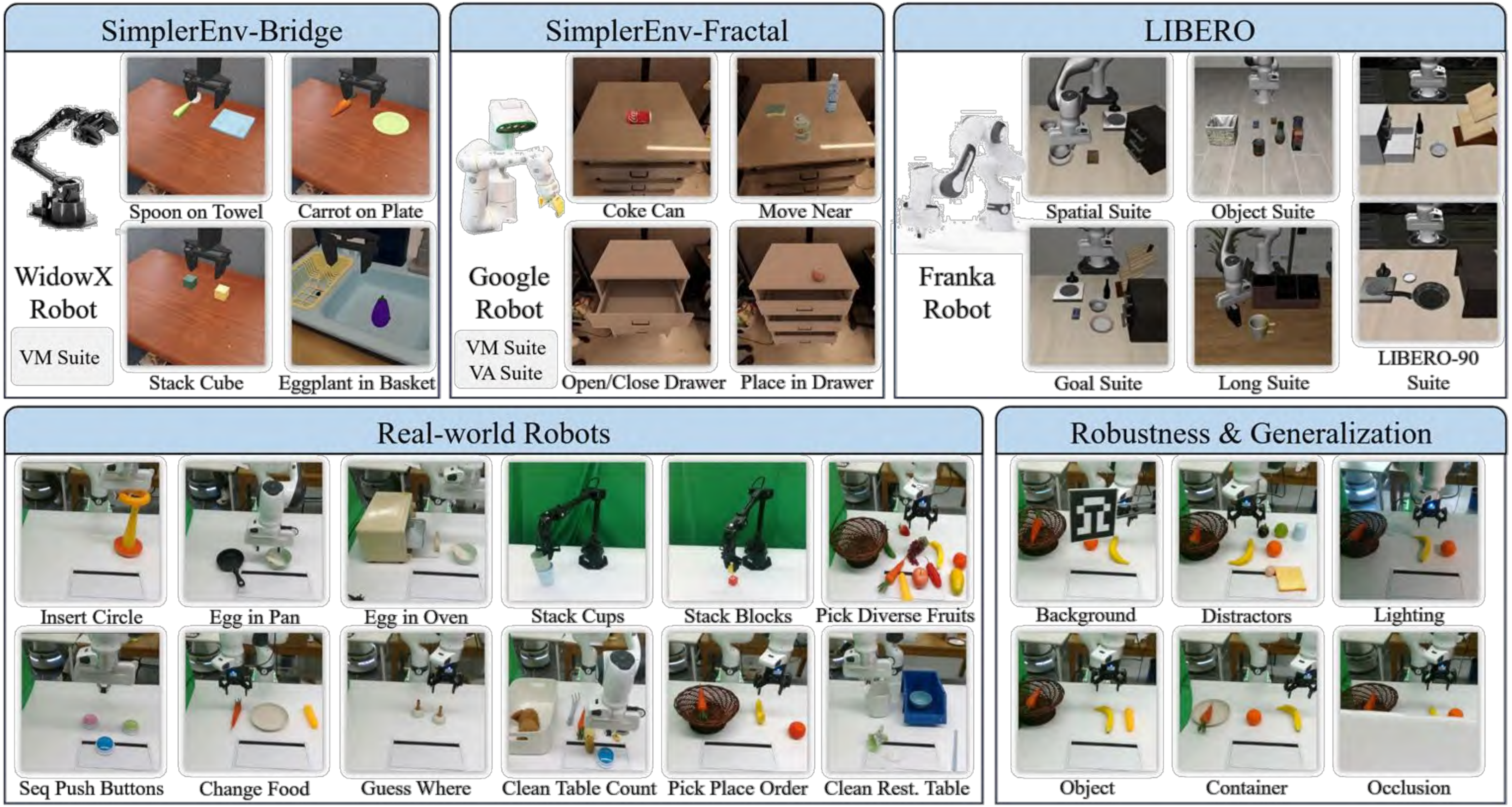
Select past info relevant to current decision

Adaptive fusion of past and current info

Merge nearby & similar entries for compact memory

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

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



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

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)

SimplerEnv-Bridge

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	93.2
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)

LIBERO

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)

SimplerEnv-Fractal

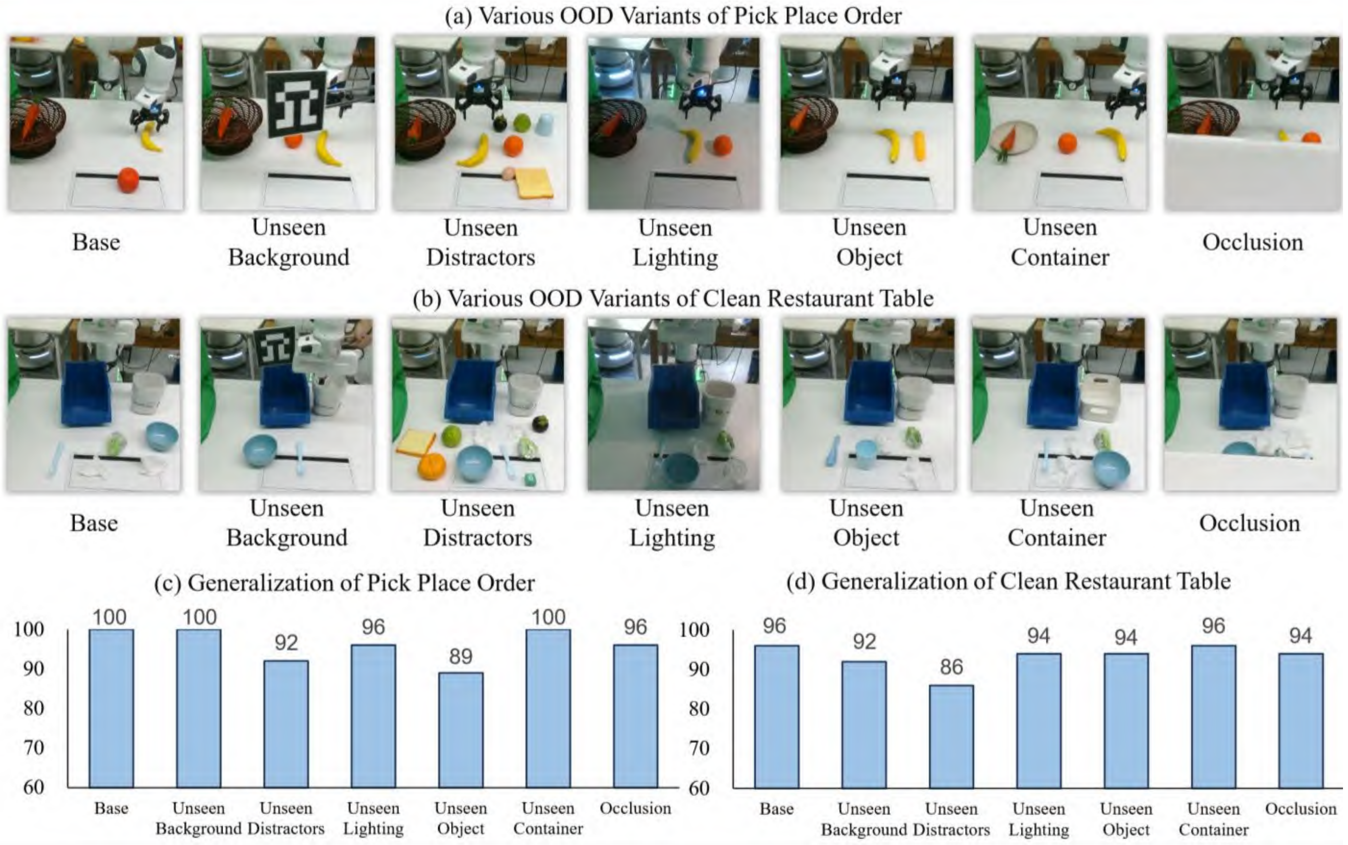
Method	General Tasks						
	Insert Circle	Egg in Pan	Egg in Oven	Stack Cups	Stack Blocks	Pick Diverse Fruits	Avg. Success
OpenVLA (Kim et al., 2024)	47	27	53	40	13	4	31
π_0 (Black et al., 2024)	67	73	73	87	53	80	72
CogACT (Li et al., 2024a)	80	67	60	93	80	76	76
MemoryVLA (Ours)	87	80	80	93	87	84	**85** (+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	57
MemoryVLA (Ours)	58	85	72	84	100	96	**83** (+26)

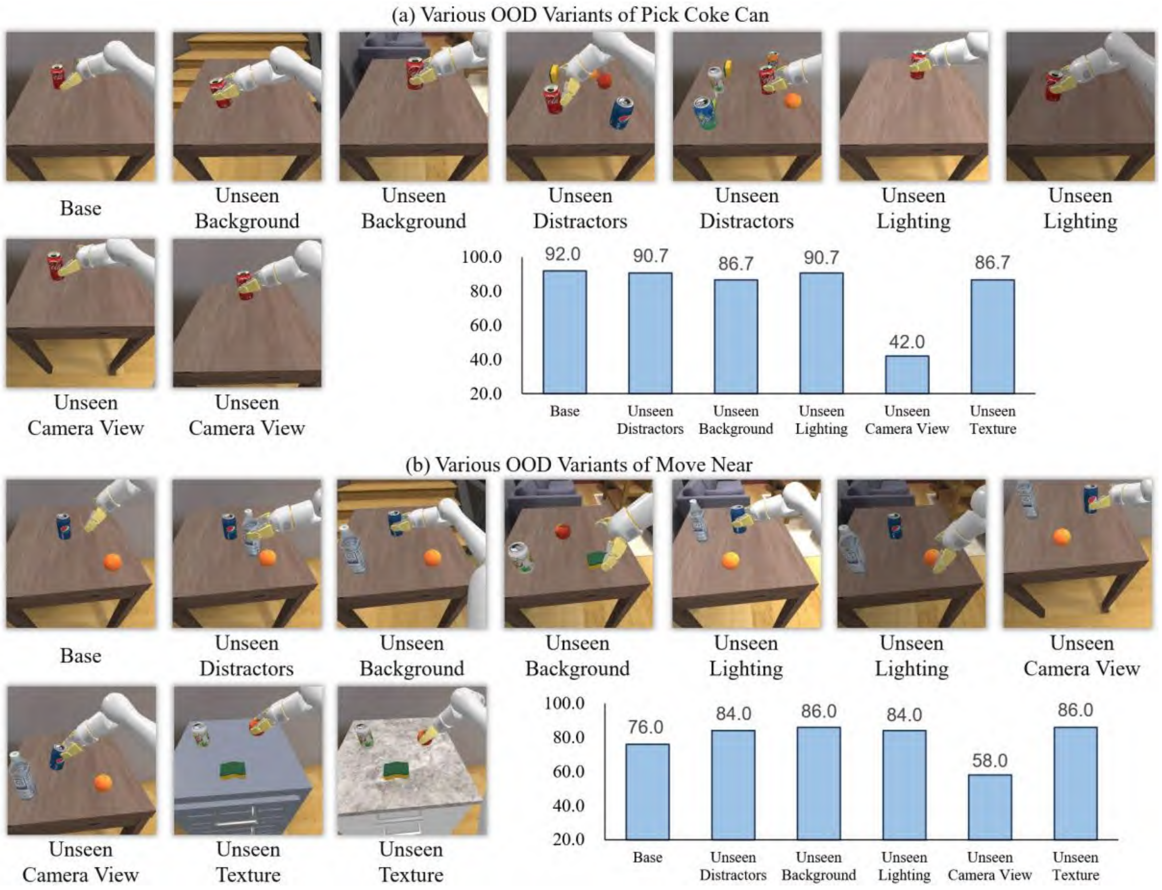
Real-World

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

Robustness and generalization



Real-World



Simulation

东南大学
SOUTHEAST UNIVERSITY

Thanks for listening!

汇报人：李文卓