ASPLOS

ReCA: Integrated Acceleration for Real-Time and Efficient Cooperative Embodied Autonomous Agents

Zishen Wan[1], Yuhang Du[2], Mohamed Ibrahim[1], Jiayi Qian[1], Jason Jabbour[3], Yang (Katie) Zhao[2], Tushar Krishna[1], Arijit Raychowdhury[1], Vijay Janapa Reddi[3]

Georgia Tech University of Minnesota Harvard University

ACM Inter Conf on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2025

Autonomous Machine Era

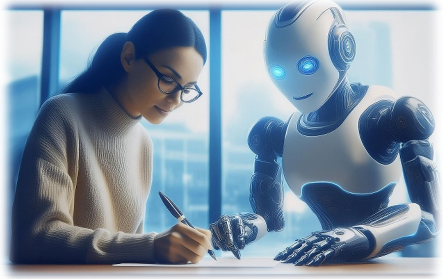
- Autonomous Machines on the Rise

Self-Driving Cars

Drones

Legged Robot

AR/VR



Embodied AI Robot

- Wide Application Potential

Package Delivery

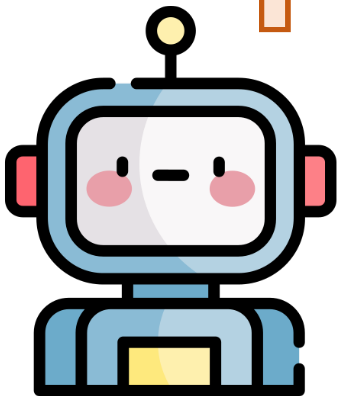
Search & Rescue

Agriculture

Manufacture

Space

Embodied Agentic Systems



Cognition Capability (System 2)

Human-like reasoning
Trustworthy decision making
Human-agent interaction

Autonomy Capability (System 1)

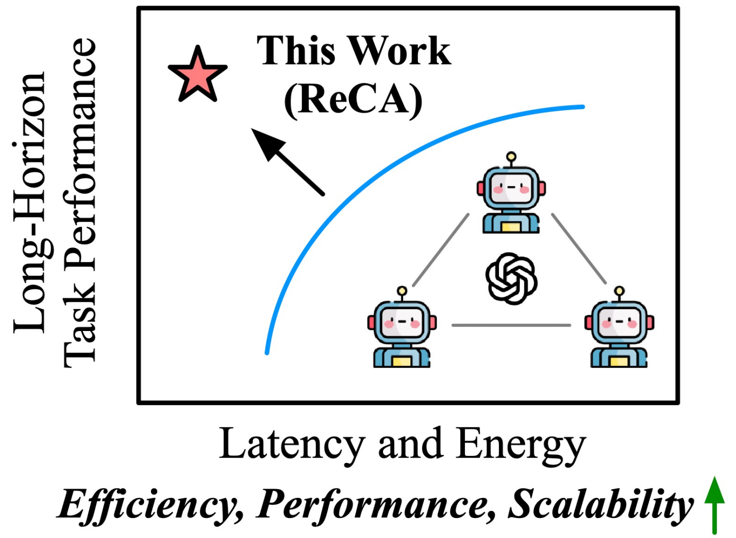
Perception, localization and mapping, path planning, control and action

Requirements

- Real-Time ***Performance***
- Energy ***Efficiency***
- Design ***Scalability***

Goal of this Work (Executive Summary)

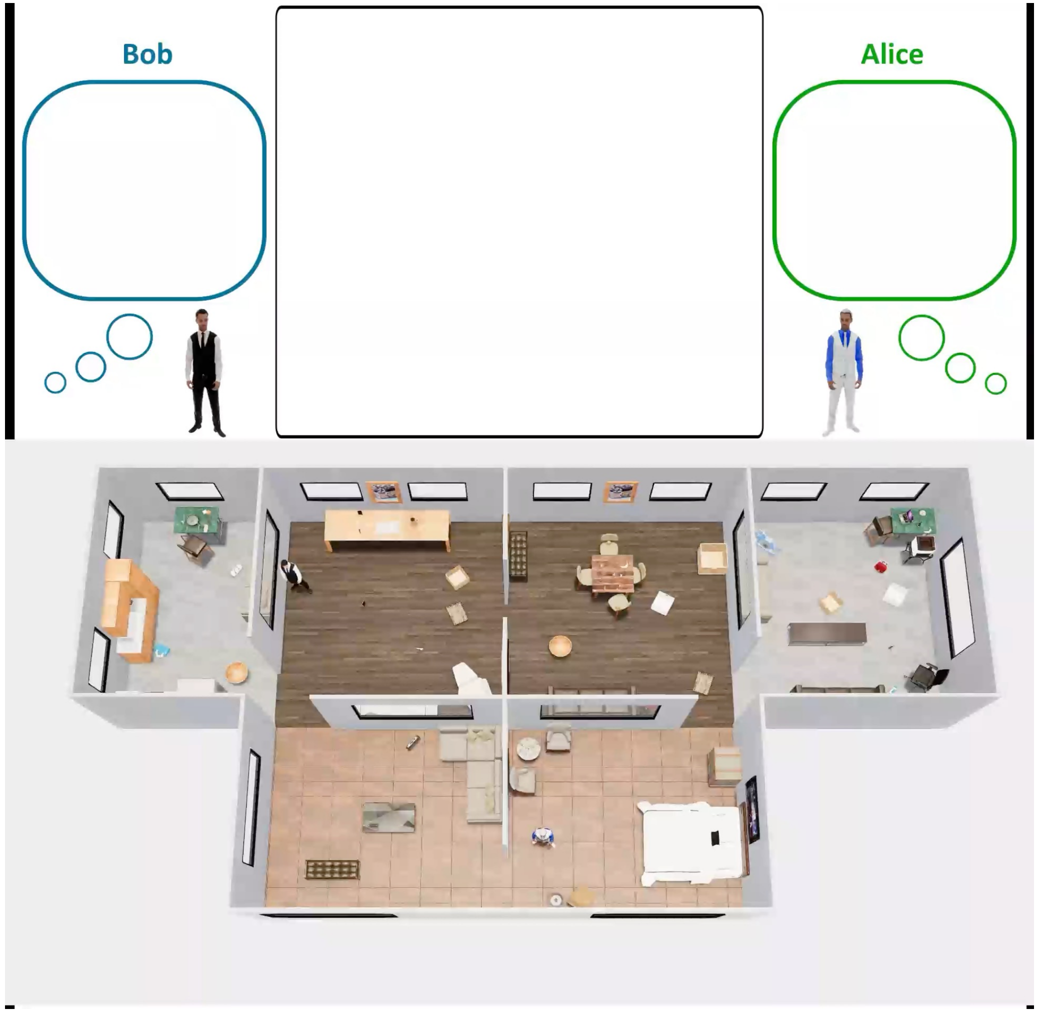
- *Understand* fundamental **building blocks** and **characteristics** of embodied systems.
- *Identify* **optimization opportunities** for embodied systems.
- *Demonstrate* scalability and efficiency improvement of embodied systems via **co-design** intelligence.



Embodied Autonomous Agent System

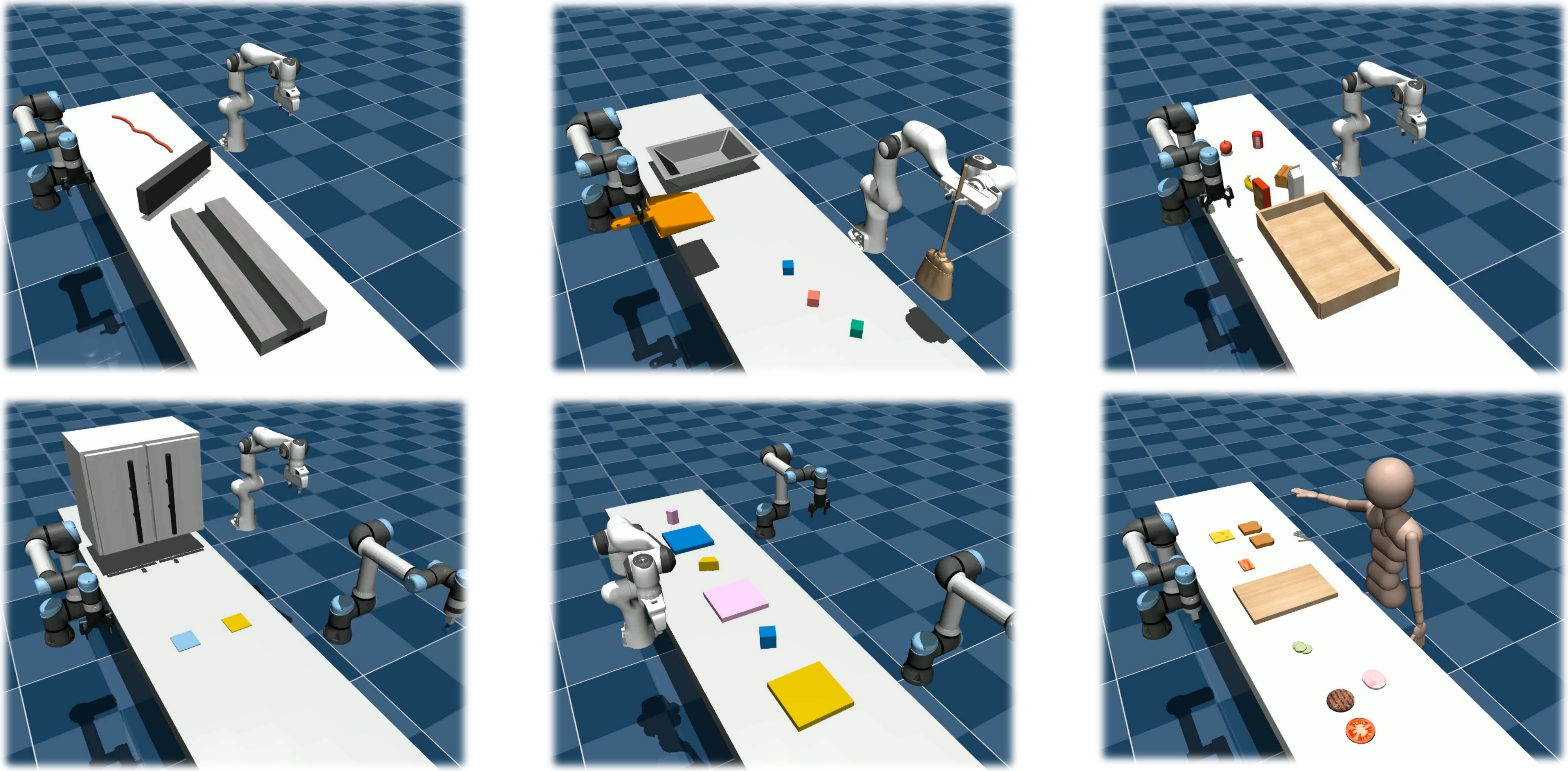
- **Task**: long-horizon multi-objective task and motion planning
  - Examples: household tasks, transport objects, make meal, set up table, cook…

Demo: Long-Horizon Multi-Objective Planning



Zhang et al, "CoELA: Building Cooperative Embodied Agents Modularly with Large Language Models", in ICLR 2024

Demo: Long-Horizon Multi-Objective Planning



Zhao et al, “RoCo: Dialectic Multi-Robot Collaboration with Large Language Models”, in arXiv 2023

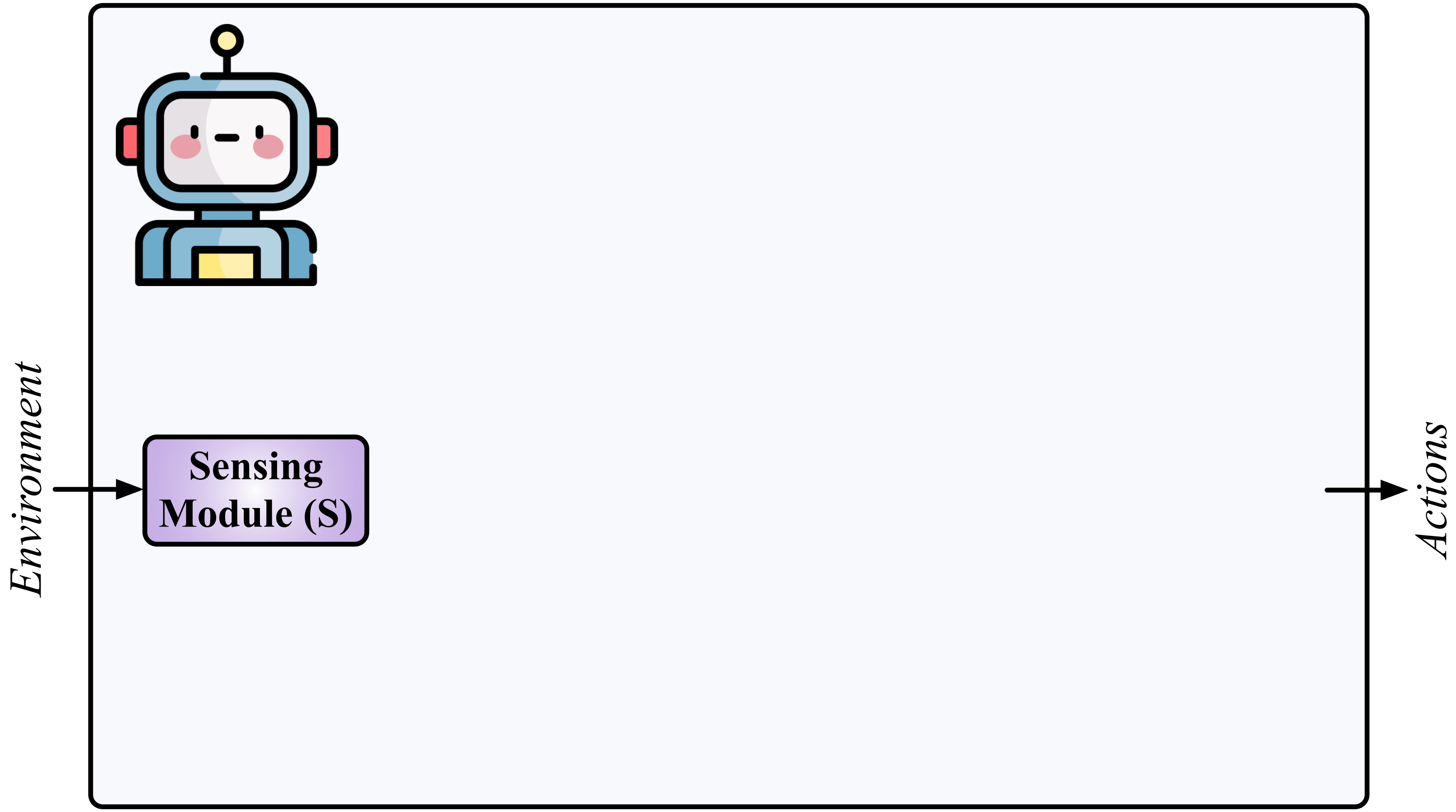
Research Question:

What are the fundamental **building blocks** and **paradigms** of embodied systems?

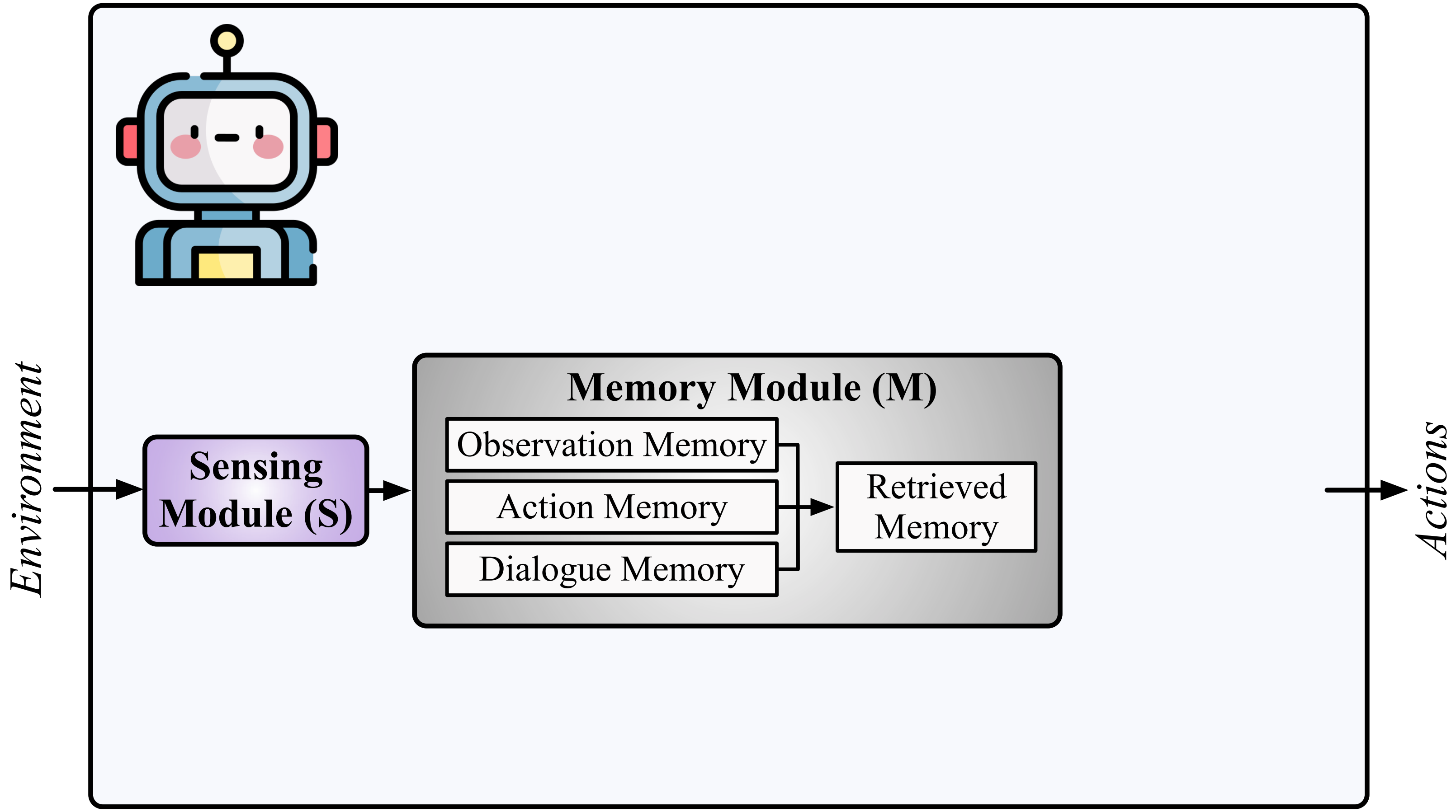
What are the **system characteristics** and **sources of inefficiencies** in these embodied systems?

Embodied Agent System Paradigm

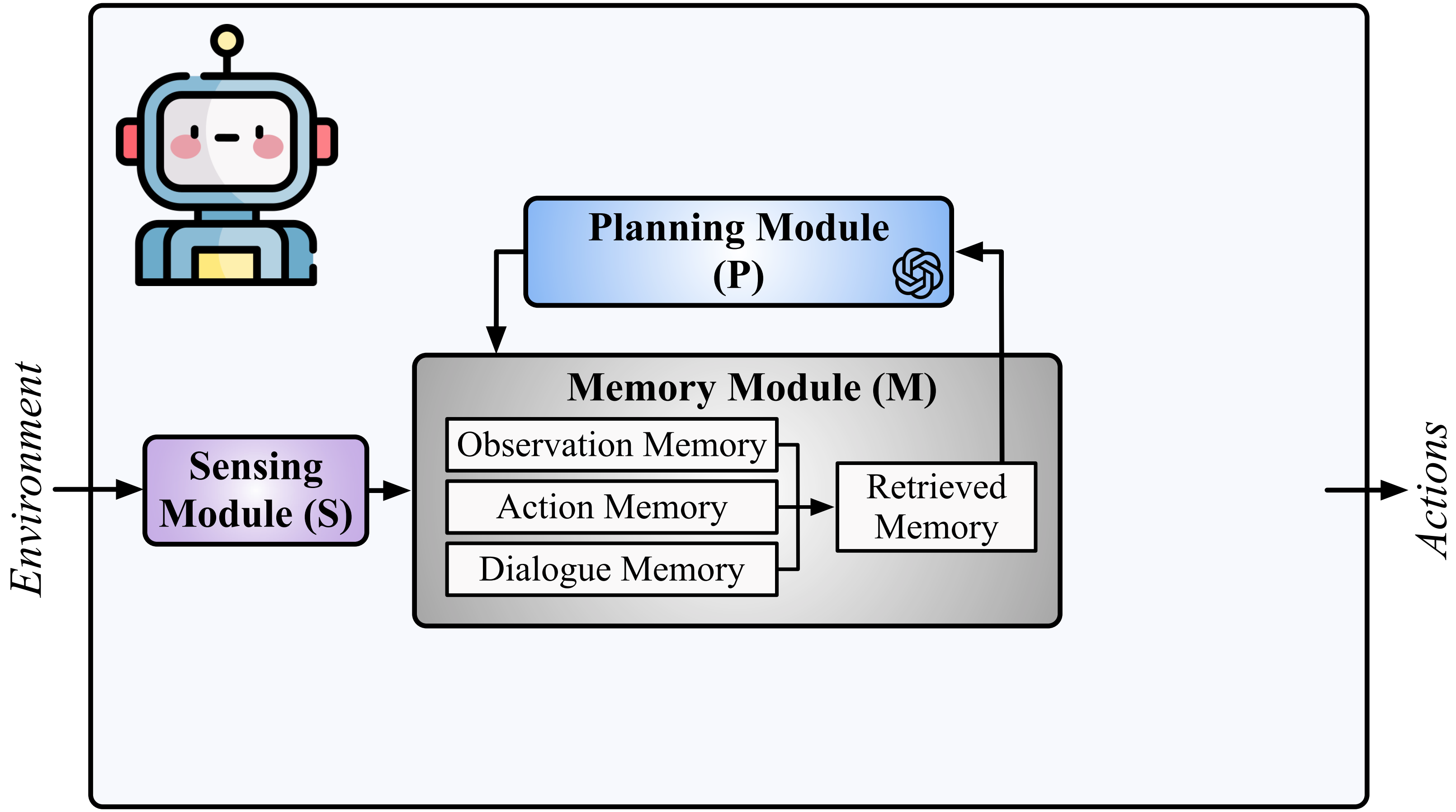
Embodied Agent System Paradigm



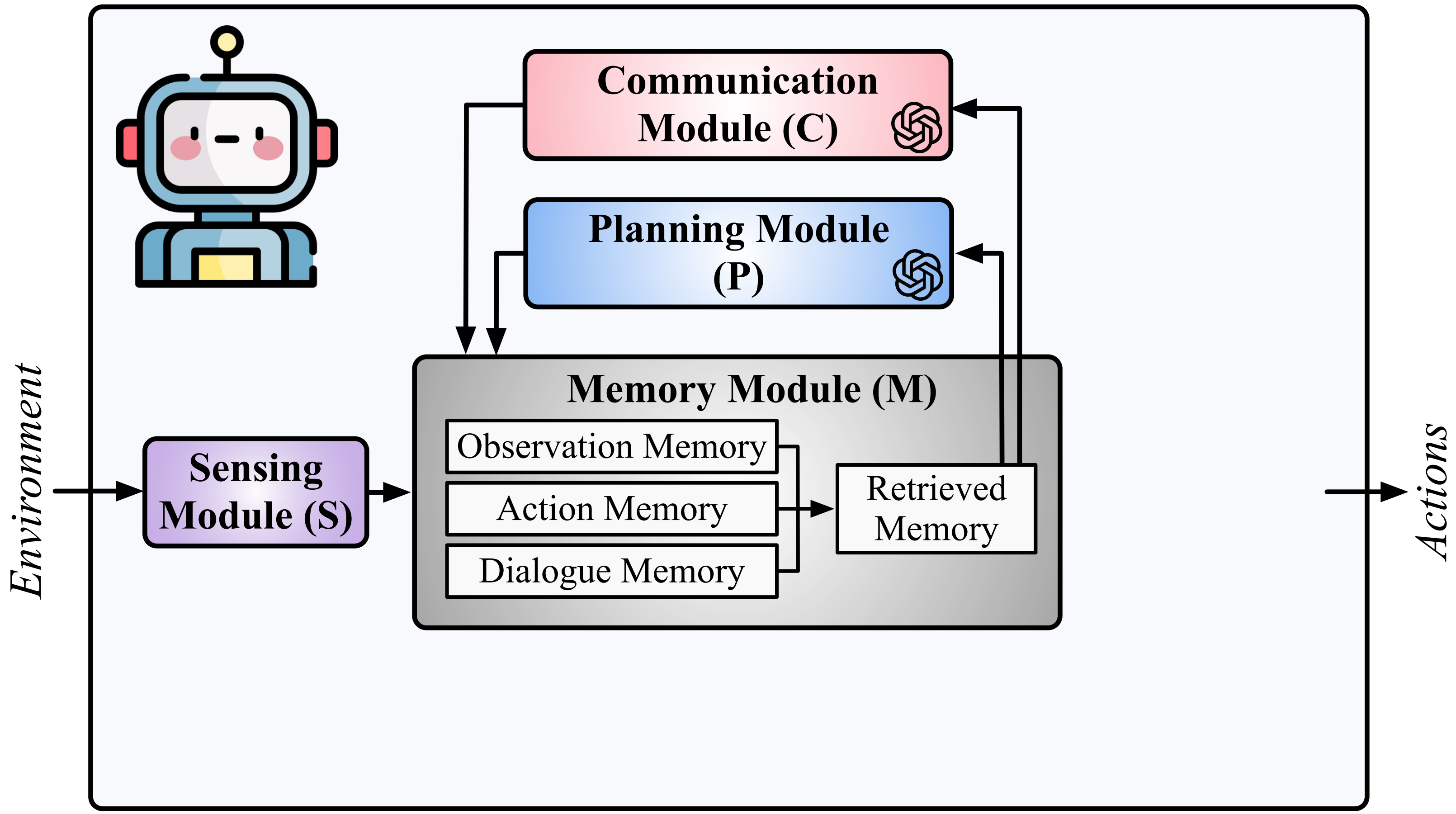
Embodied Agent System Paradigm



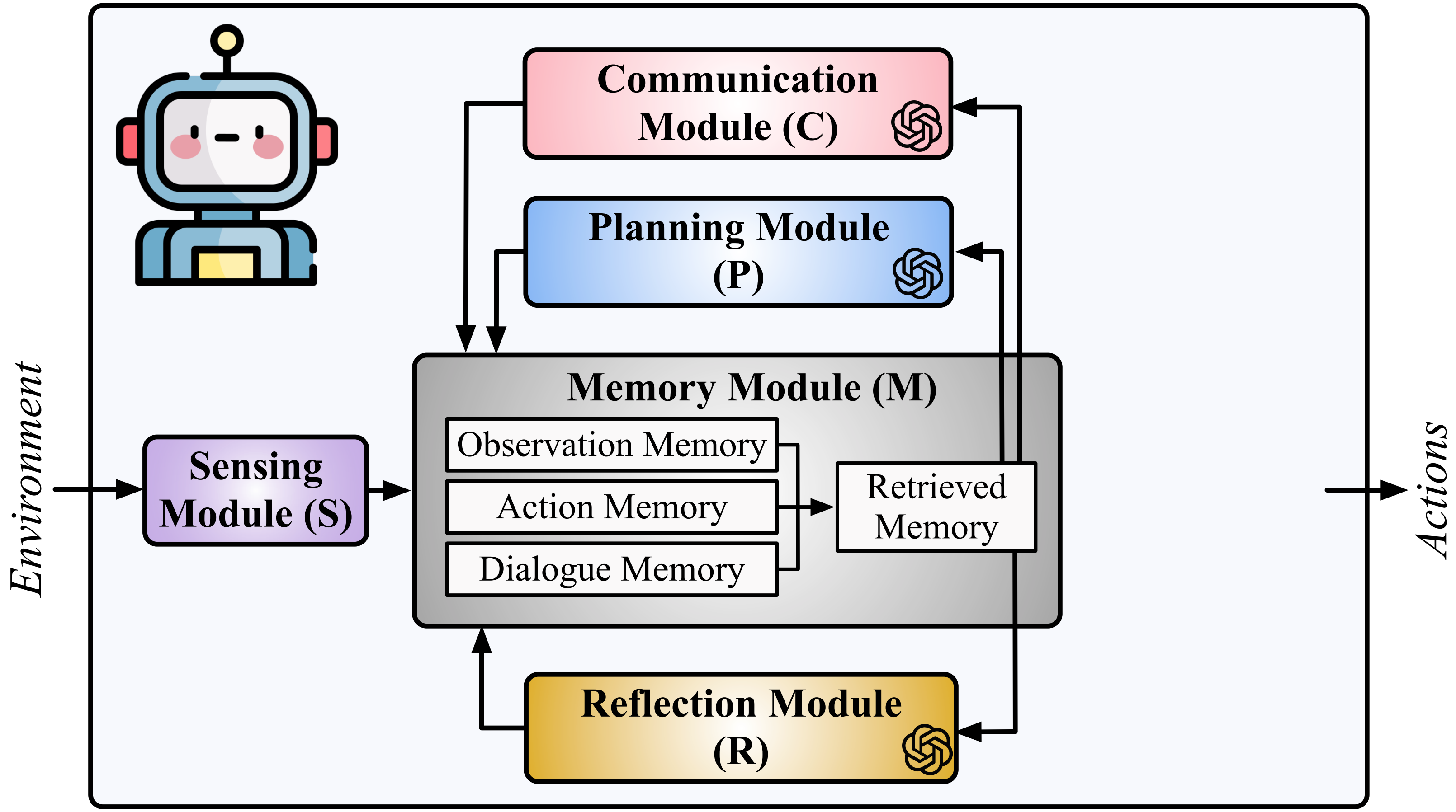
Embodied Agent System Paradigm



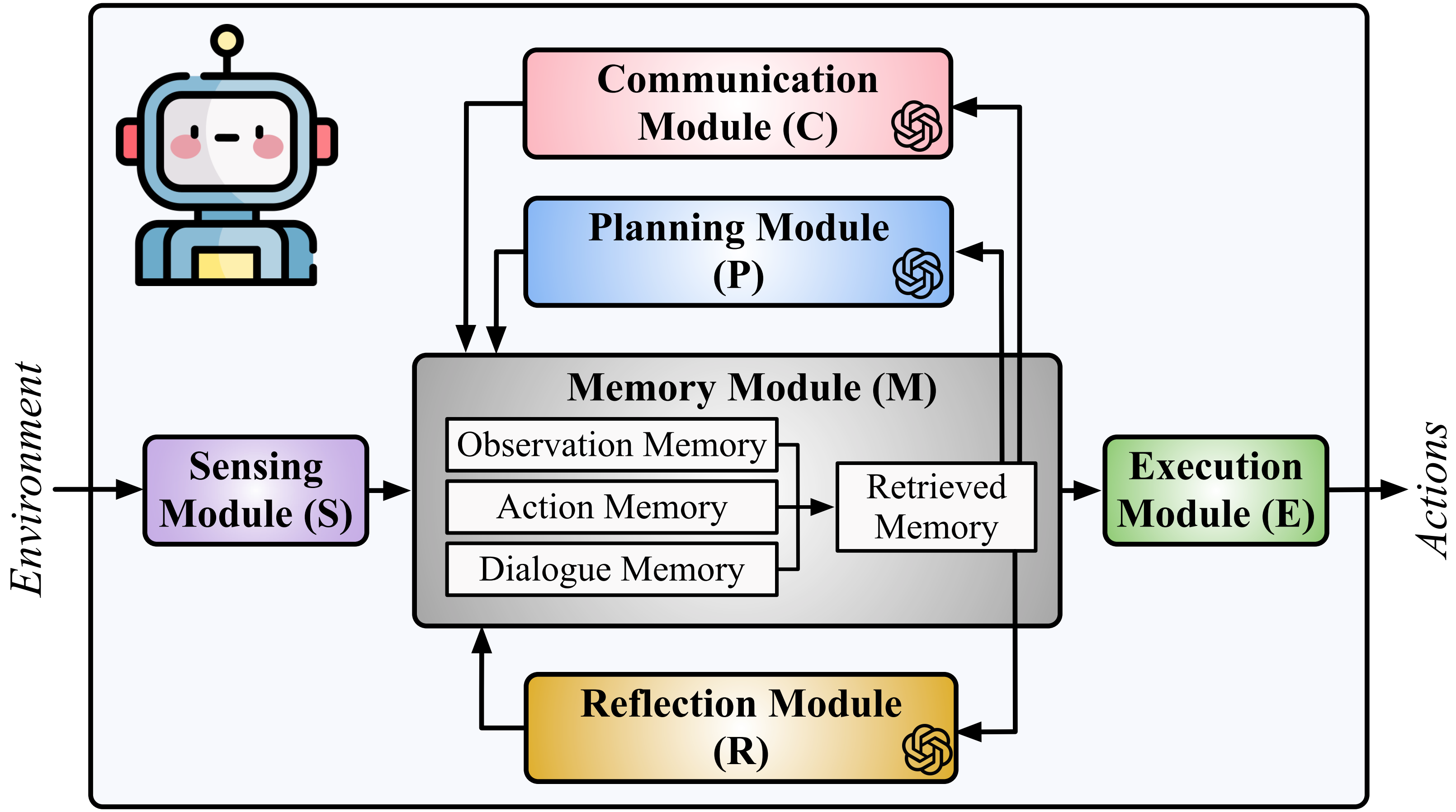
Embodied Agent System Paradigm



Embodied Agent System Paradigm



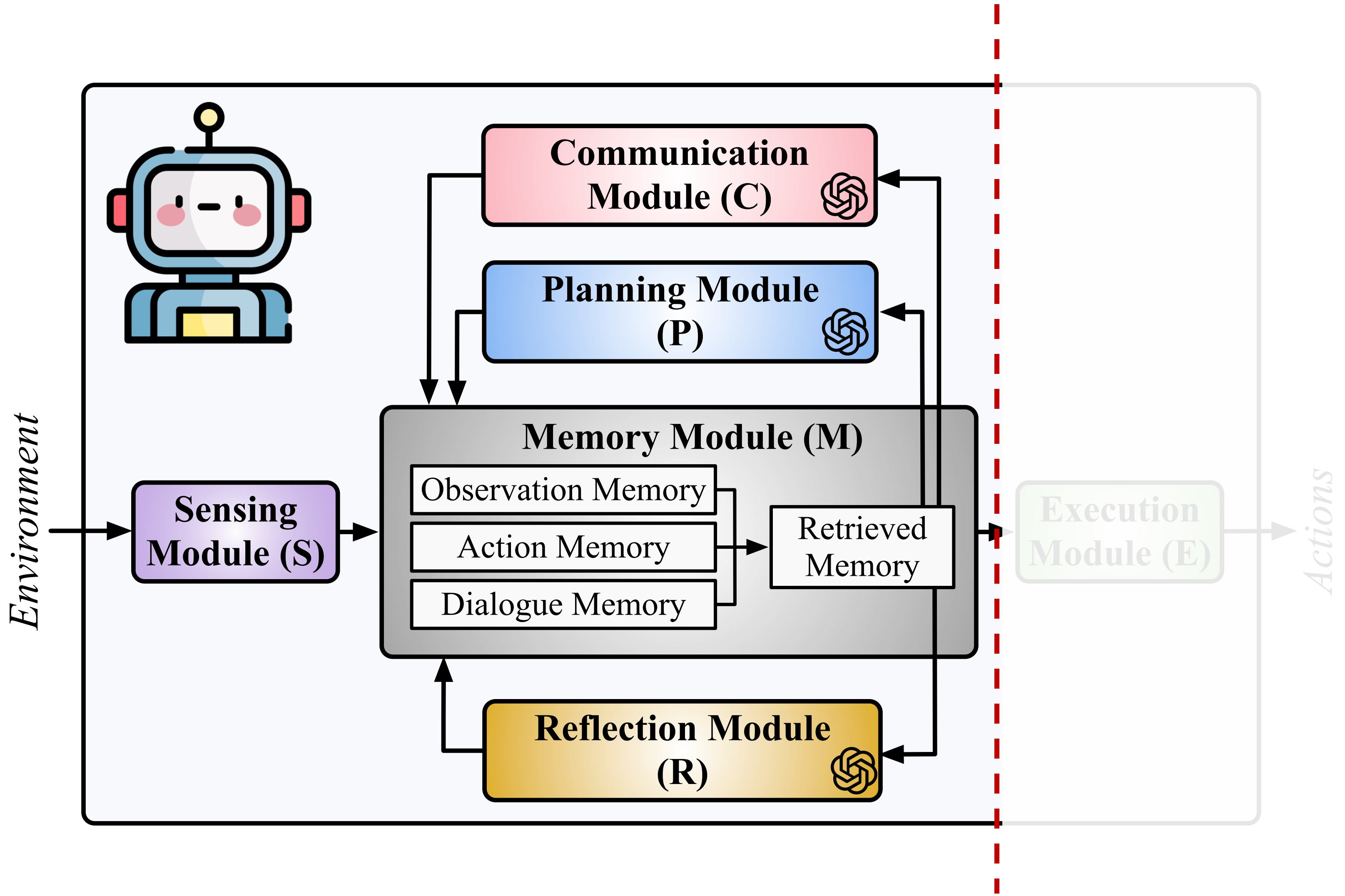
Embodied Agent System Paradigm



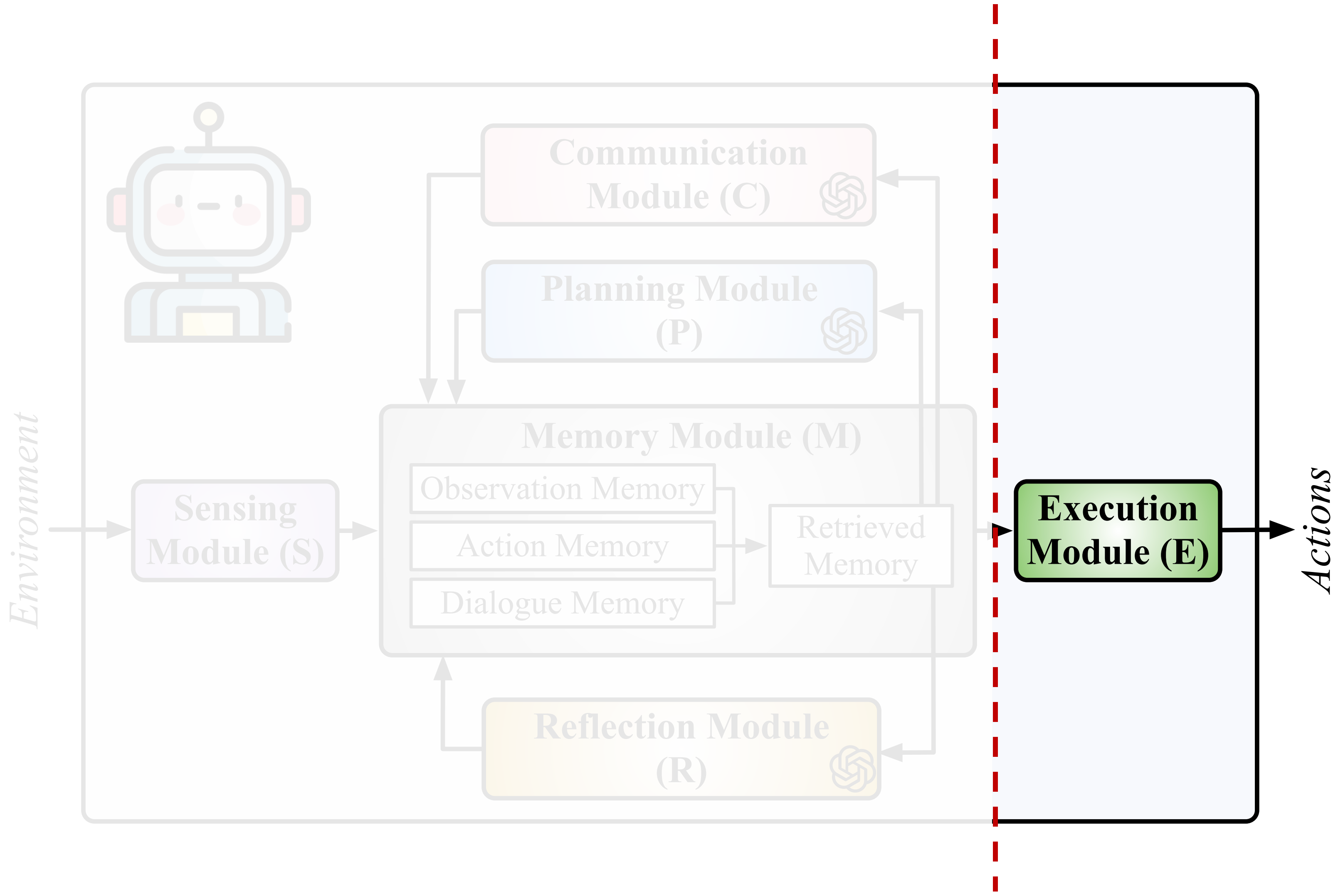
Embodied Agent System Paradigm

System 2

(Cognition Capability)



Embodied Agent System Paradigm

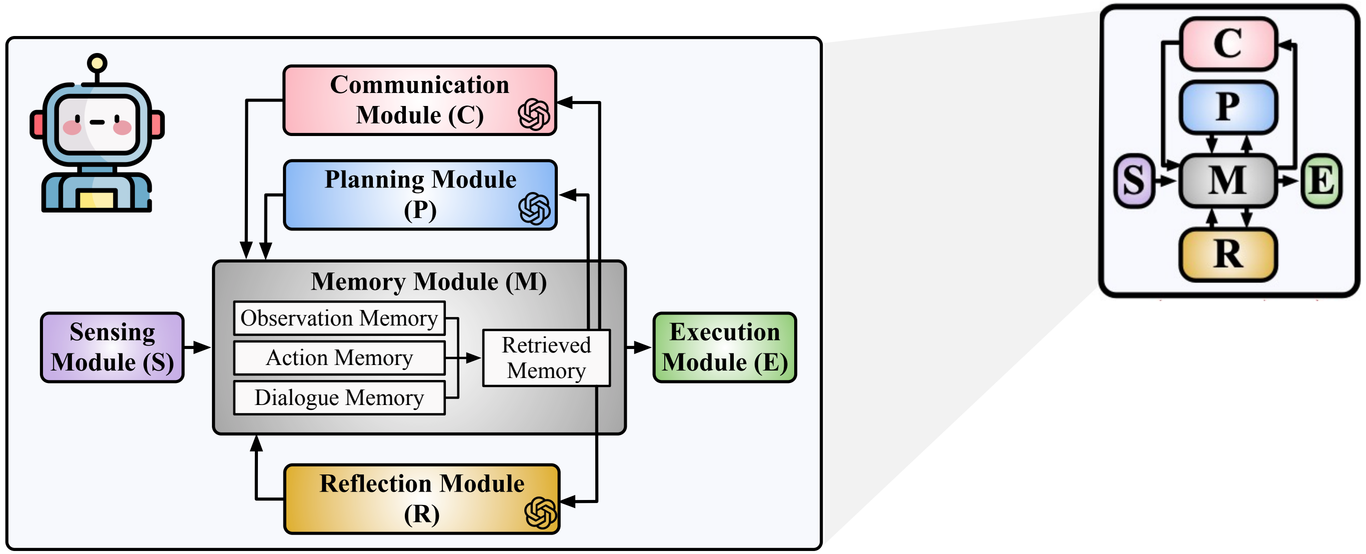


System 1

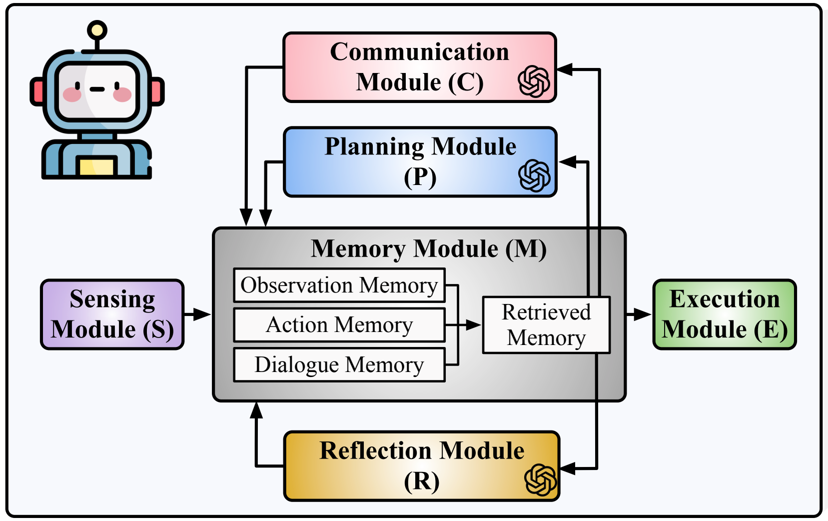
(Autonomy Capability)

Embodied Agent System Paradigm

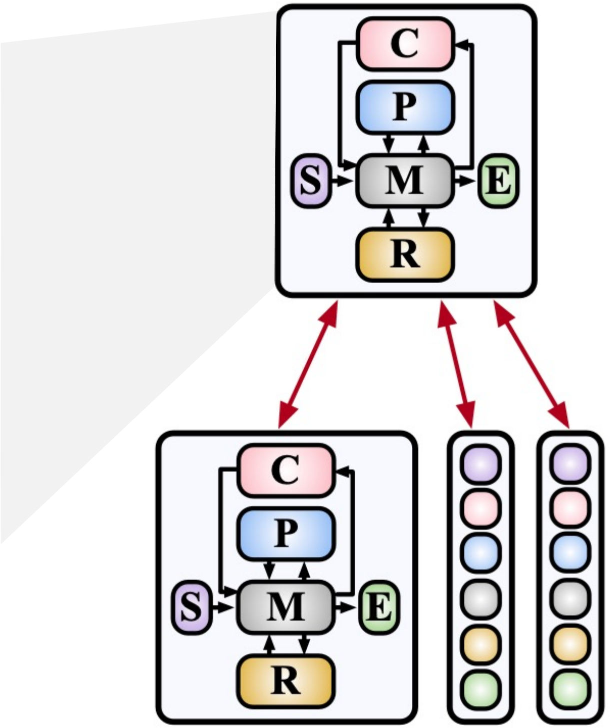
Cooperative Embodied AI Systems



Embodied Agent System Paradigm

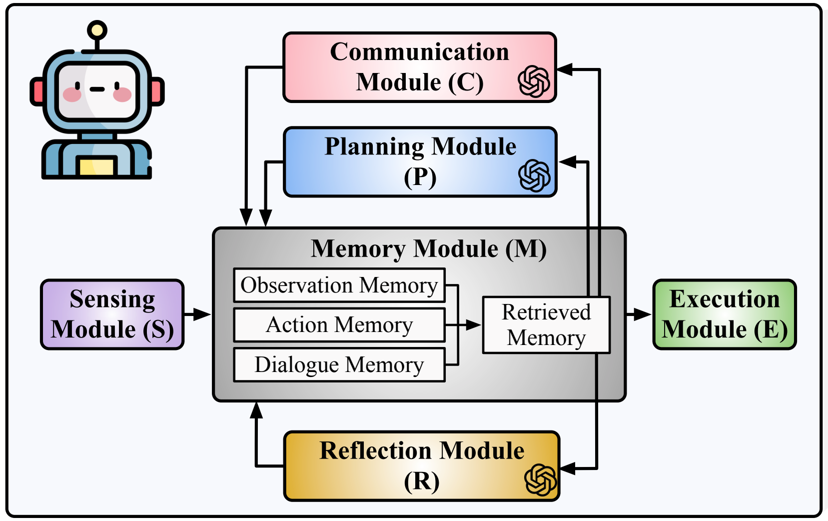


Cooperative Embodied AI Systems

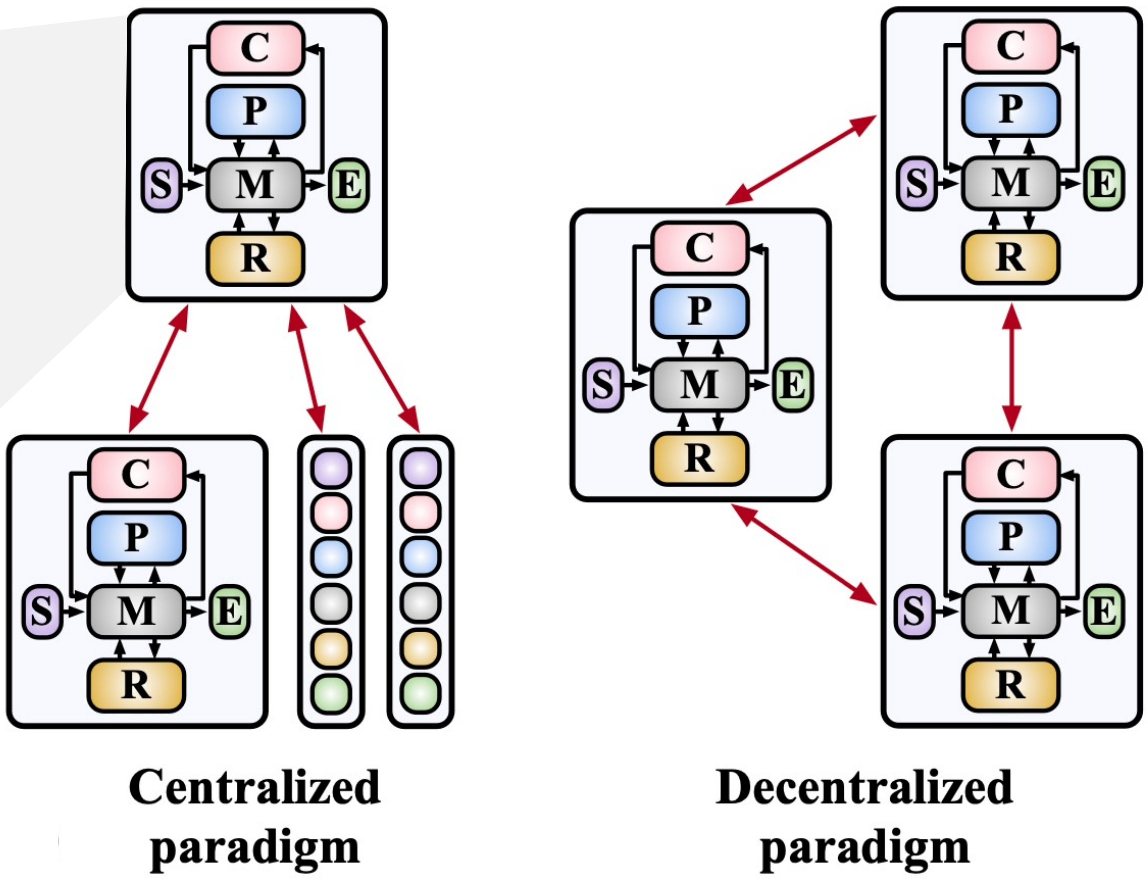


Centralized paradigm

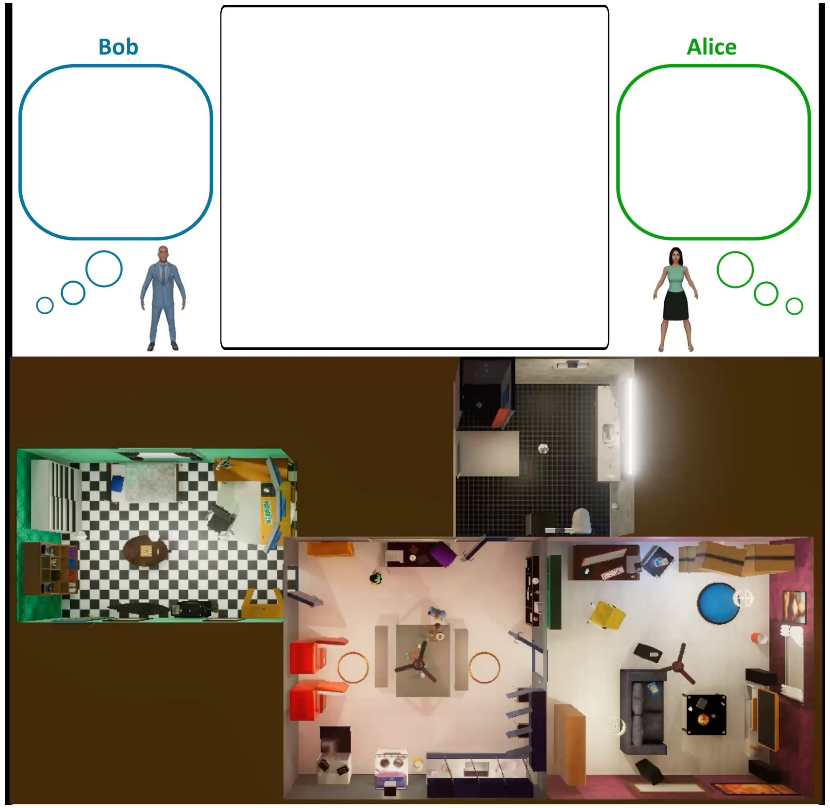
Embodied Agent System Paradigm

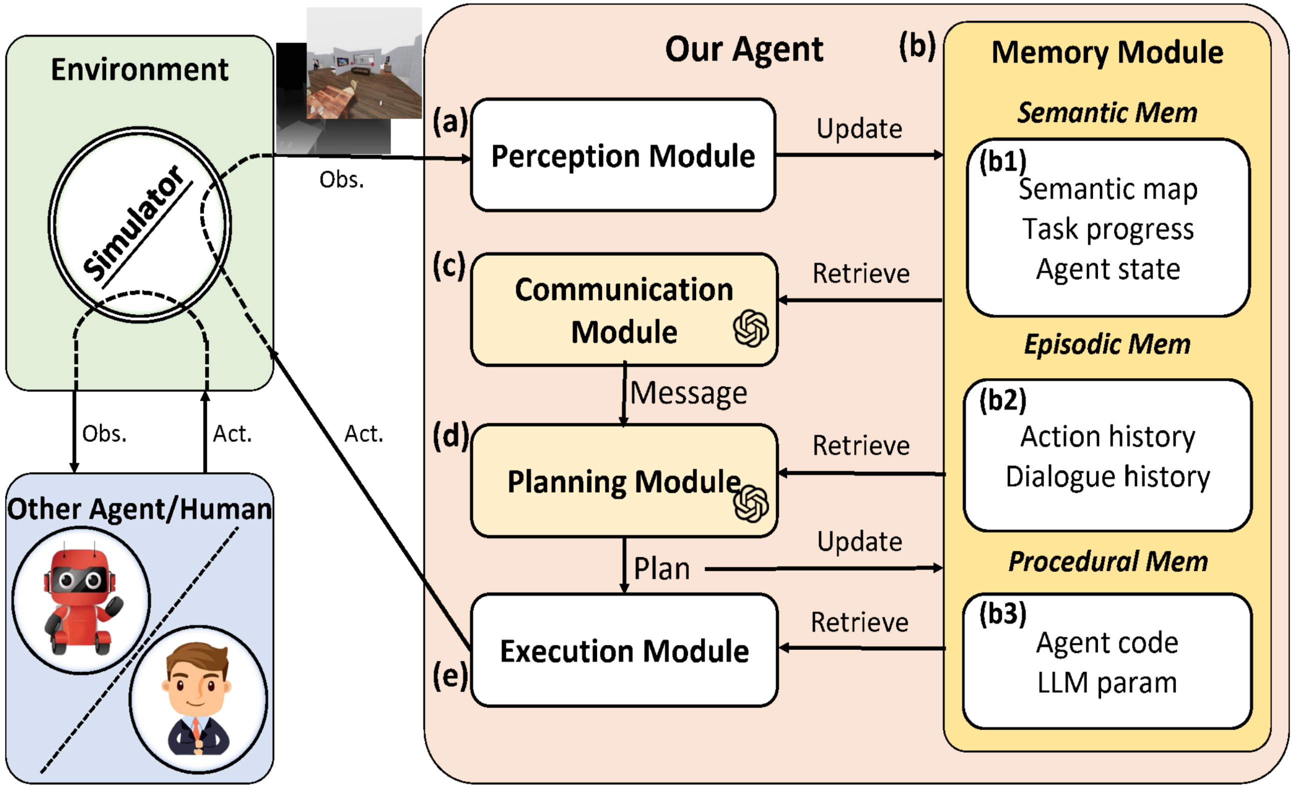


Cooperative Embodied AI Systems



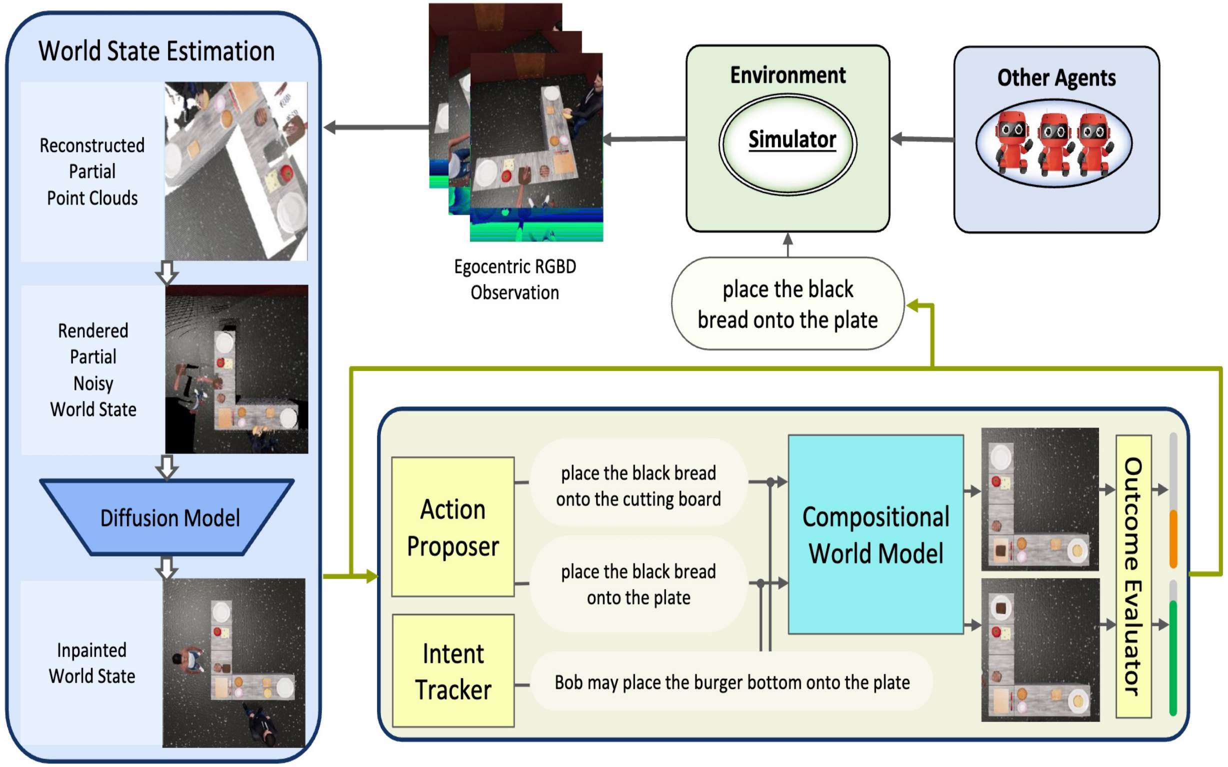
Embodied System Example: CoELA





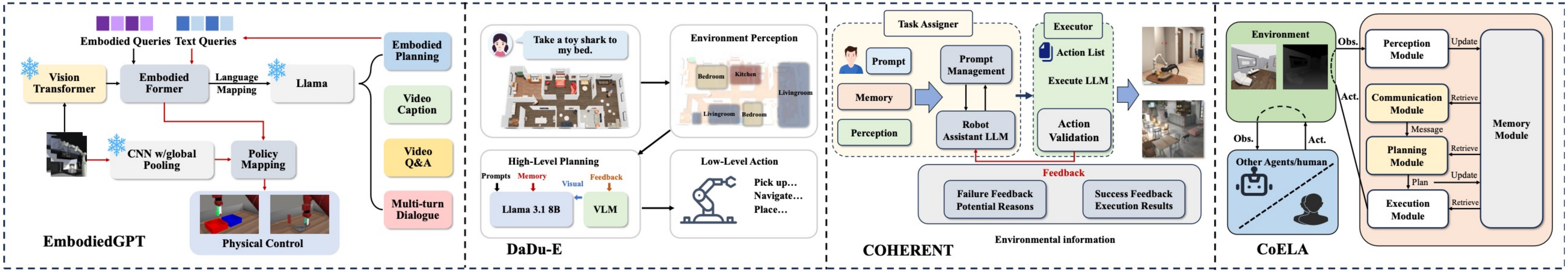
Zhang et al, “CoELA: Building Cooperative Embodied Agents Modularly with Large Language Models”, in ICLR 2024

Embodied System Example: COMBO



Zhang et al, "COMBO: Compositional World Models for Embodied Multi-Agent Cooperation", in ICLR 2025

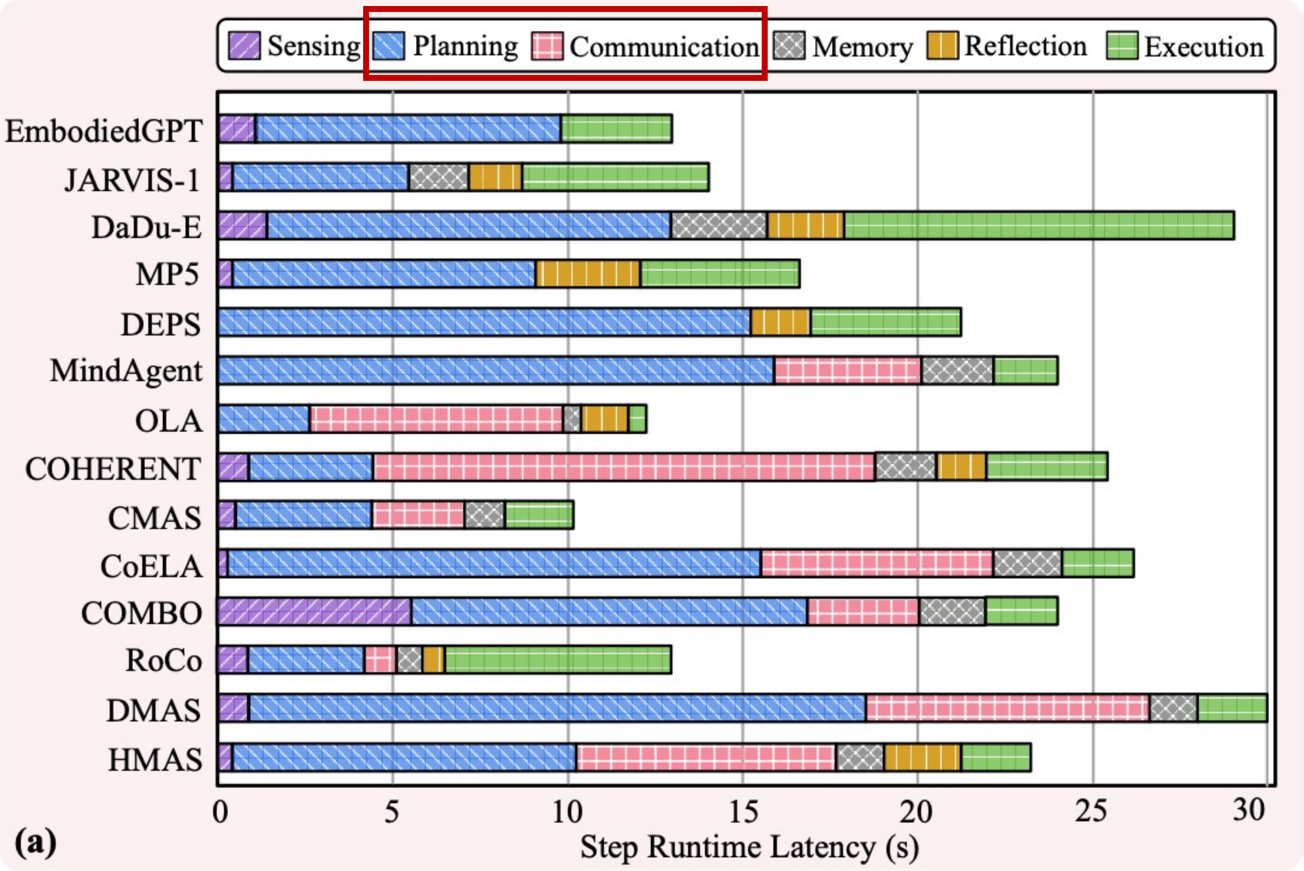
Representative Embodied Agent Workloads



Embodied AI Systems	System Module						Application	Datasets and Tasks
	Sensing	Planning	Communication	Memory	Reflection	Execution		
EmbodiedGPT [39]	ViT	Llama-7B	–	–	–	MLP	Embodied planning, visual captioning, VQA	Franka Kitchen [59], Meta-World [60], VirtualHome [61]
JARVIS-1 [24]	MineCLIP	GPT-4/Llama-13B	–	Ob., Act.	Llama-13B	Action list	Embodied planning (e.g, obtain diamond pickaxe)	Minecraft [62]
DaDu-E [40]	PointCloud	Llama-8B	–	Ob., Act.	LLaVA-8B	AnyGrasp	Object transport, Autonomous decision-making	Self-designed four-level tasks
MP5 [36]	MineCLIP	GPT-4	–	–	GPT-4	MineDojo	Object transport, Situation-aware long-term planning	Minecraft [62]
DEPS [15]	Symbolic info	GPT-4	–	–	CLIP	MineDojo	Embodied planning (e.g, obtain diamond pickaxe)	Minecraft [62], MineRL [63], ALFWorld [64]
MindAgent [6]	–	GPT-4	GPT-4	Ob., Act., Dx.	–	Action list	Collaborative planning, gaming, housework	CuisineWorld [6], Minecraft [62]
OLA [21]	–	GPT-4/Llama-70B	GPT-4	Ob., Act., Dx.	GPT-4	Action list	Collaborative planning, object transport	VirtualHome [61], C-WAH [65]
COHERENT [28]	DINO	GPT-4	GPT-4	Ob., Act., Dx.	GPT-4	RRT/A-star	Collaborative planning, Robot arm manipulation	BEHAVIOR-1K [66]
CMAS [20]	ViLD	GPT-4	GPT-4	Ob., Act., Dx.	–	Action list	Collaborative planning, manipulator, object transport	BoxNet1, BoxNet2, WareHouse, BoxLift [20]
CoELA [4]	Mask R-CNN	GPT-4	GPT-4	Ob., Act., Dx.	–	A-star	Collaborative object transporting, housework	TDW-MAT [67], C-WAH [65]
COMBO [5]	Diffusion	LLaVA-7B	LLaVA-7B	Ob., Act., Dx.	–	A-star	Collaborative gaming, housework	TDW-Game [68], TDW-Cook [68]
RoCo [27]	ViT	GPT-4	GPT-4	Ob., Act., Dx.	GPT-4	RRT	Robot arm motion planning, manipulation	RoCoBench [27]
DMAS [20]	ViLD	GPT-4	GPT-4	Ob., Act., Dx.	–	Action list	Collaborative planning, manipulator, object transport	BoxNet1, BoxNet2, WareHouse, BoxLift [20]
HMAS [20]	ViLD	GPT-4	GPT-4	Ob., Act., Dx.	GPT-4	Action list	Collaborative planning, manipulator, object transport	BoxNet1, BoxNet2, WareHouse, BoxLift [20]

Embodied Agent System Characterization

Runtime Analysis:

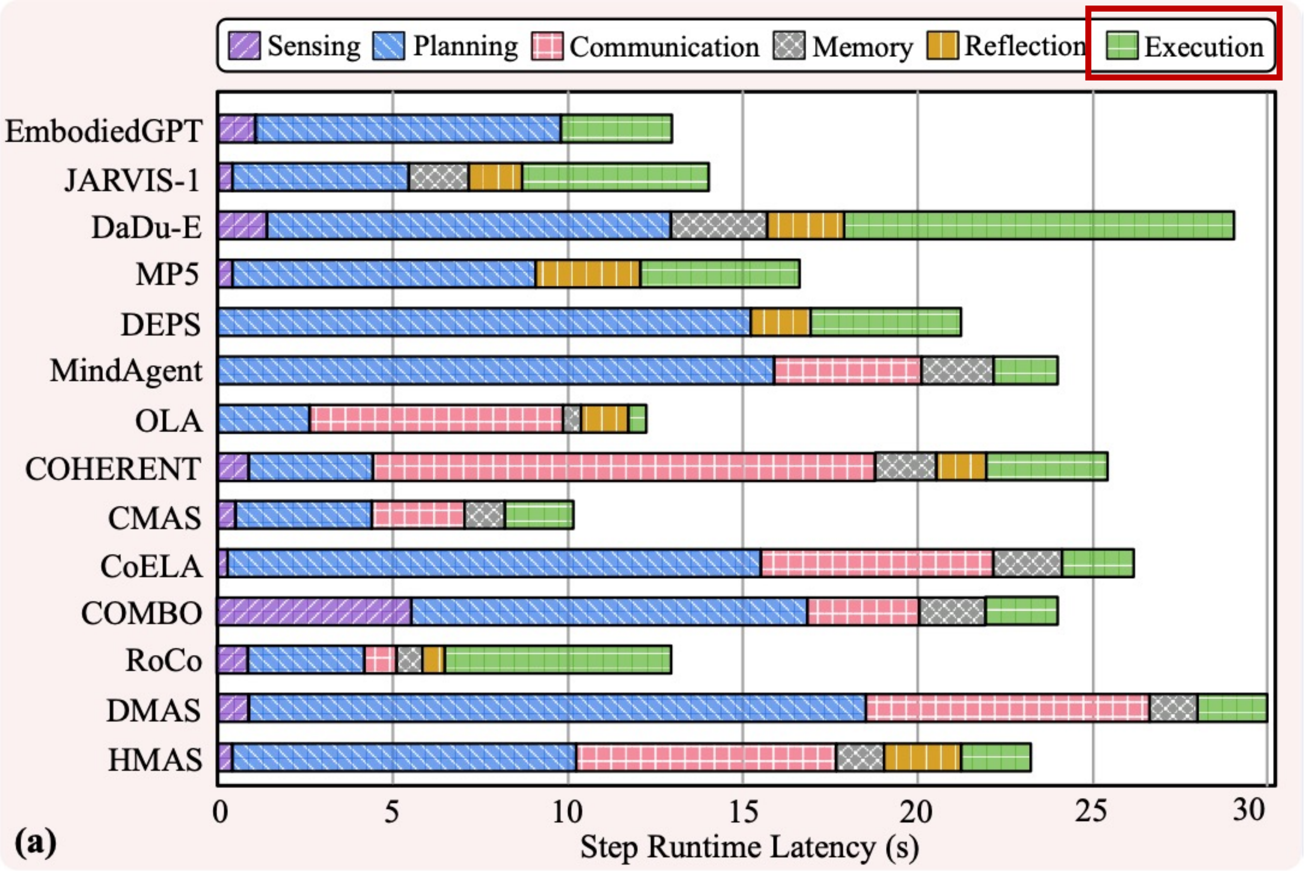


Takeaway:

- End-to-end latency in long-horizon embodied tasks is significant.
- LLM-based planning and communication dominate the latency due to repeated runs.

Embodied Agent System Characterization

Runtime Analysis:

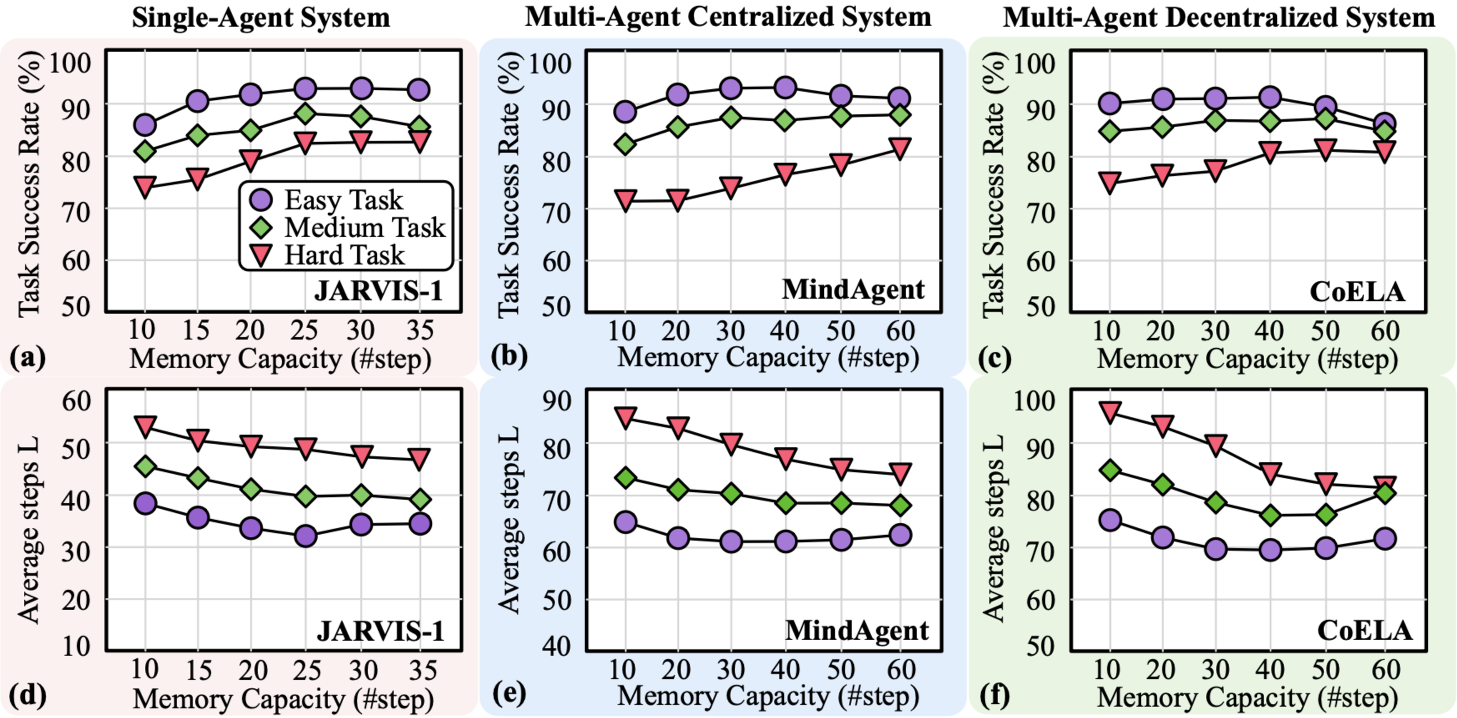


Takeaway:

- End-to-end latency in long-horizon embodied tasks is significant.
- LLM-based planning and communication dominate the latency due to repeated runs.
- Low-level planning and execution also contribute notable delays due to multiple executions and computational complexity.

Embodied Agent System Characterization

Memory Analysis:

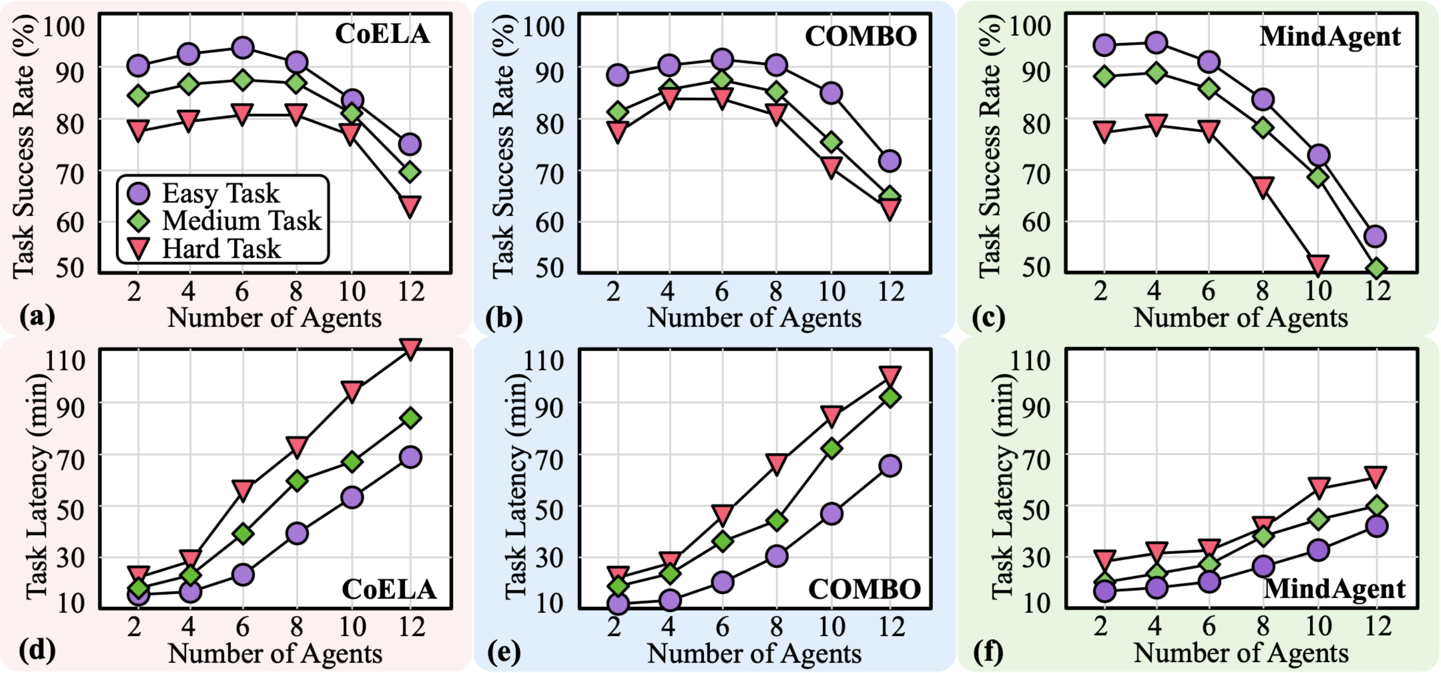


Takeaway:

- Increasing memory module capacity improves success rates and reduces #steps, especially for complex tasks.
- However, excessively large memory introduces inconsistencies and increases retrieval time per step.

Embodied Agent System Characterization

Scalability Analysis:



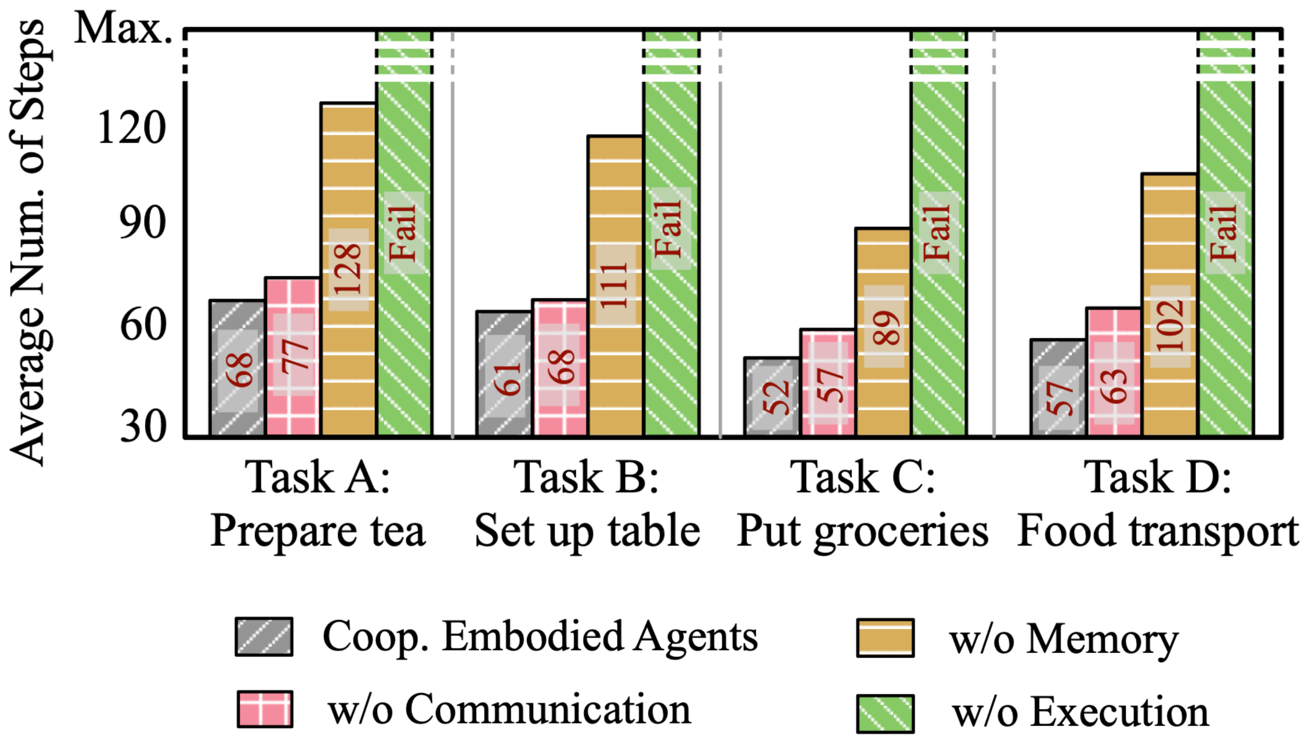
Decentralized | Decentralized | Centralized

Takeaway:

- Multi-agent embodied systems face scalability challenges as the number of agents increases.
- Centralized vs. decentralized:
  - Centralized systems: success rate challenge
  - Decentralized systems: latency challenge

Embodied Agent System Characterization

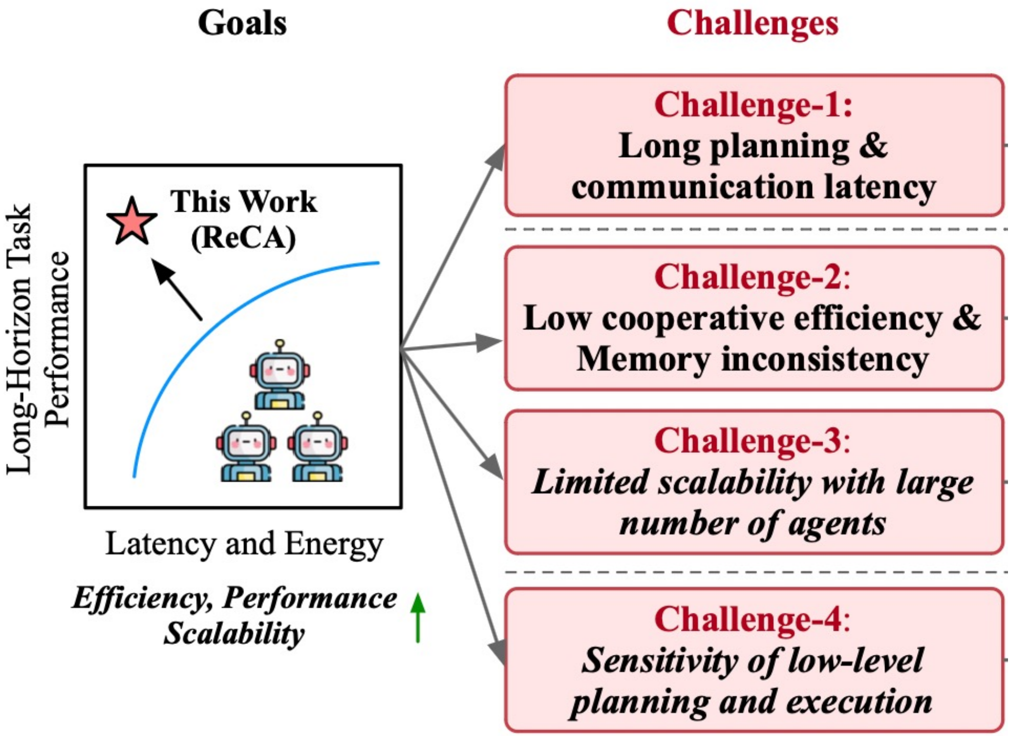
Module Sensitivity Analysis:



Takeaway:

- Memory module is critical for tracking agent status and task success.
- Low-level execution module plays an indispensable role in system functionality.

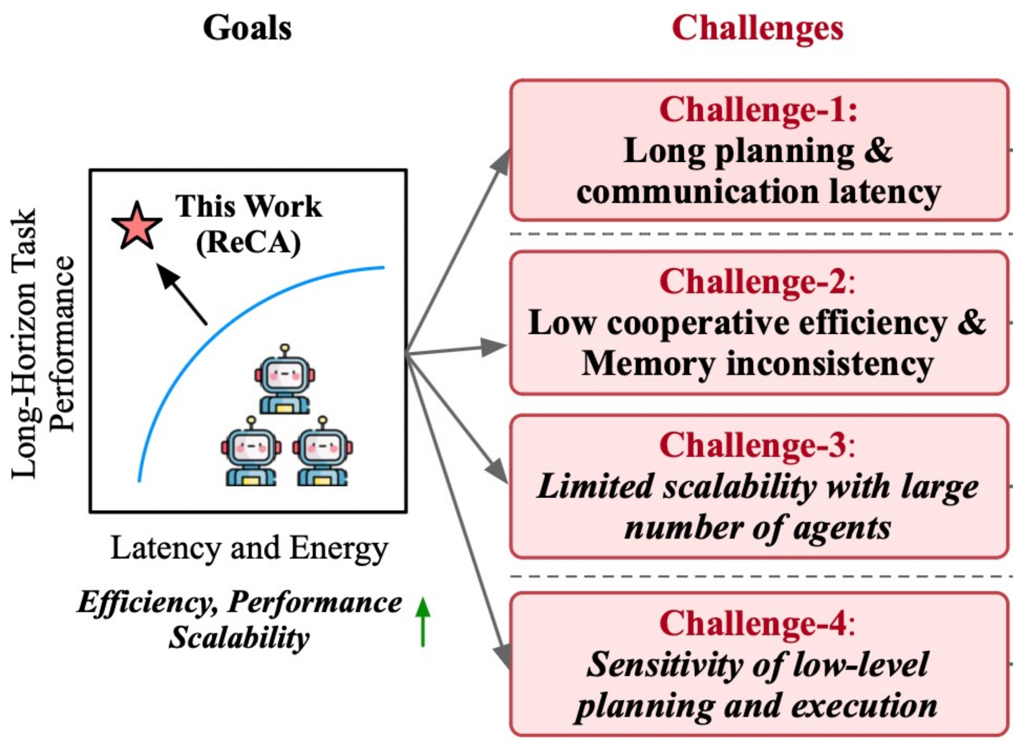
Challenges of Embodied Agent Systems



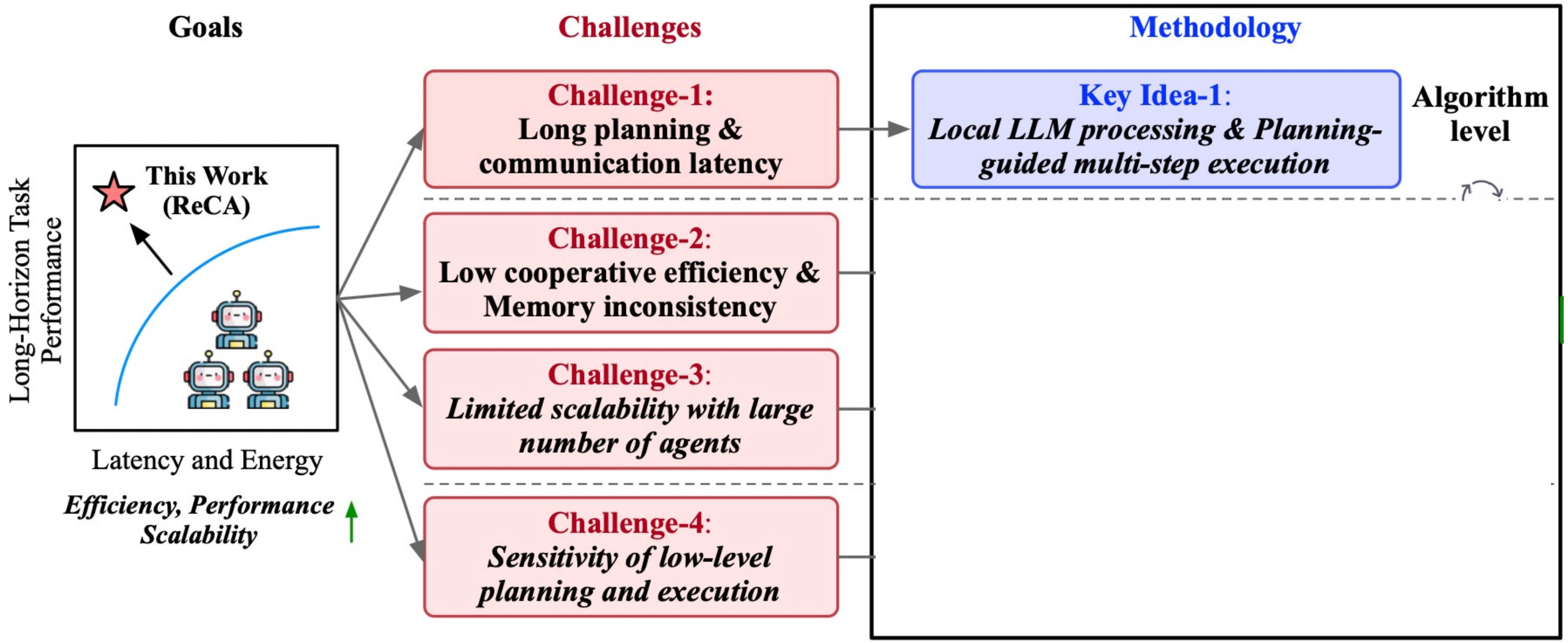
Research Question:

How to enhance the **efficiency and scalability** of cooperative embodied systems?

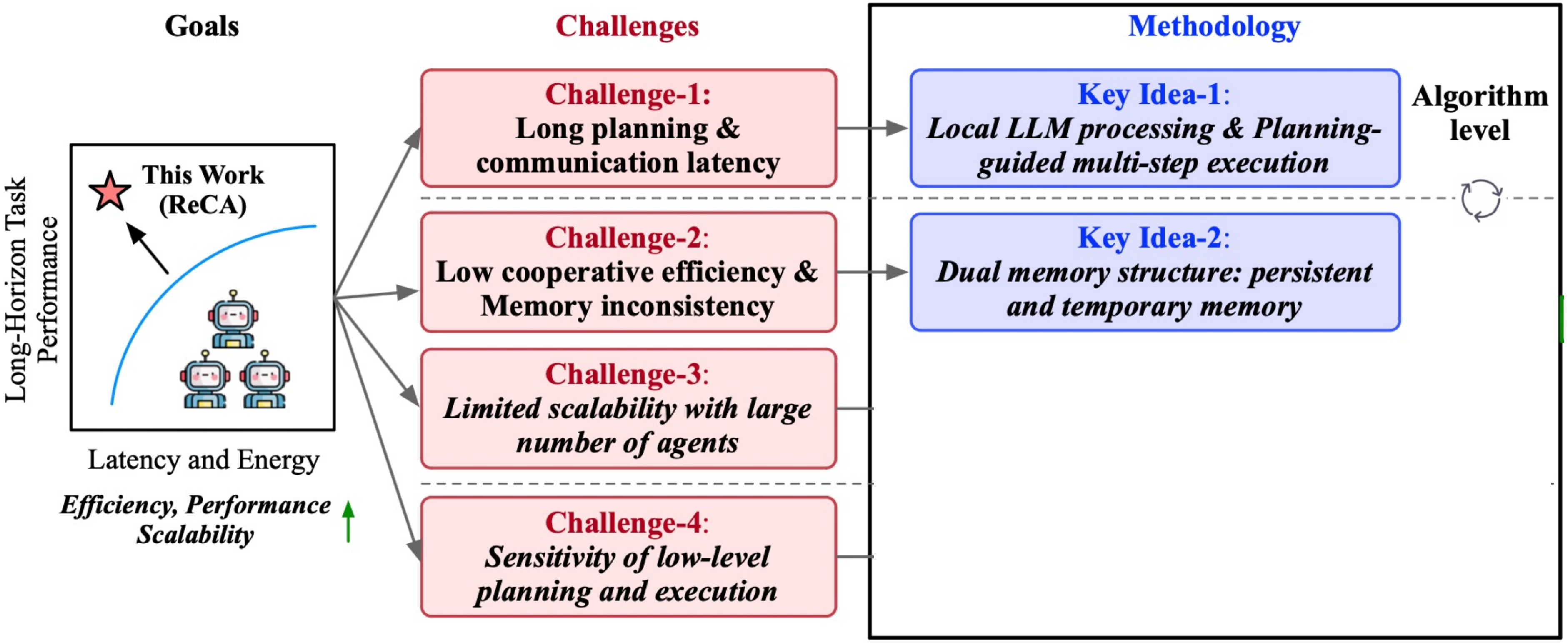
Our Methodology



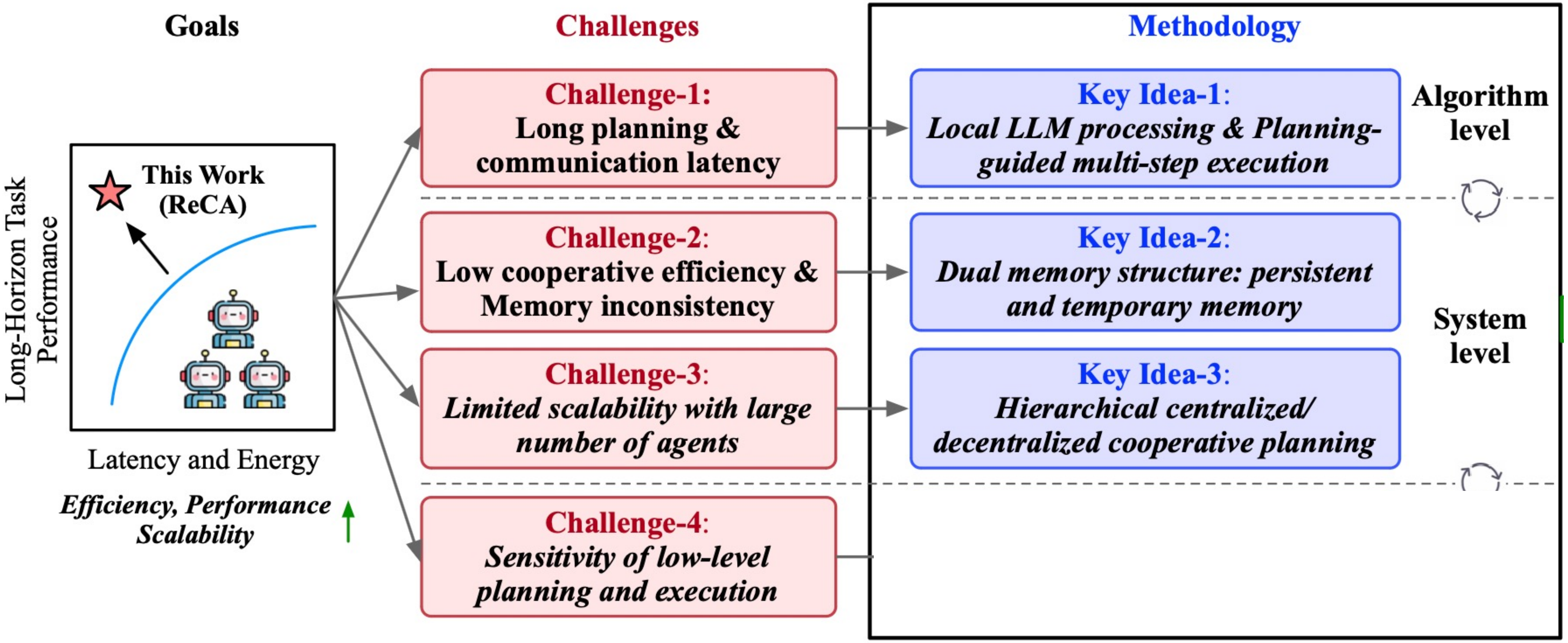
Our Methodology



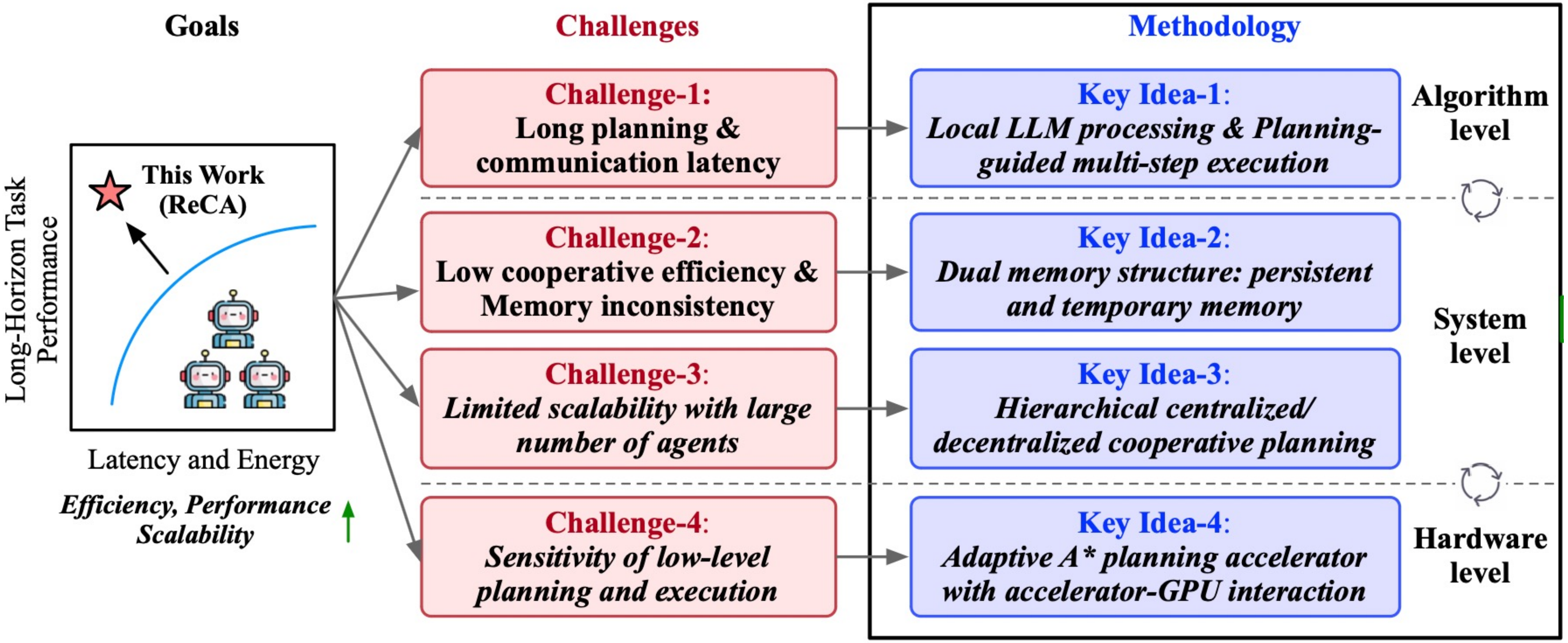
Our Methodology



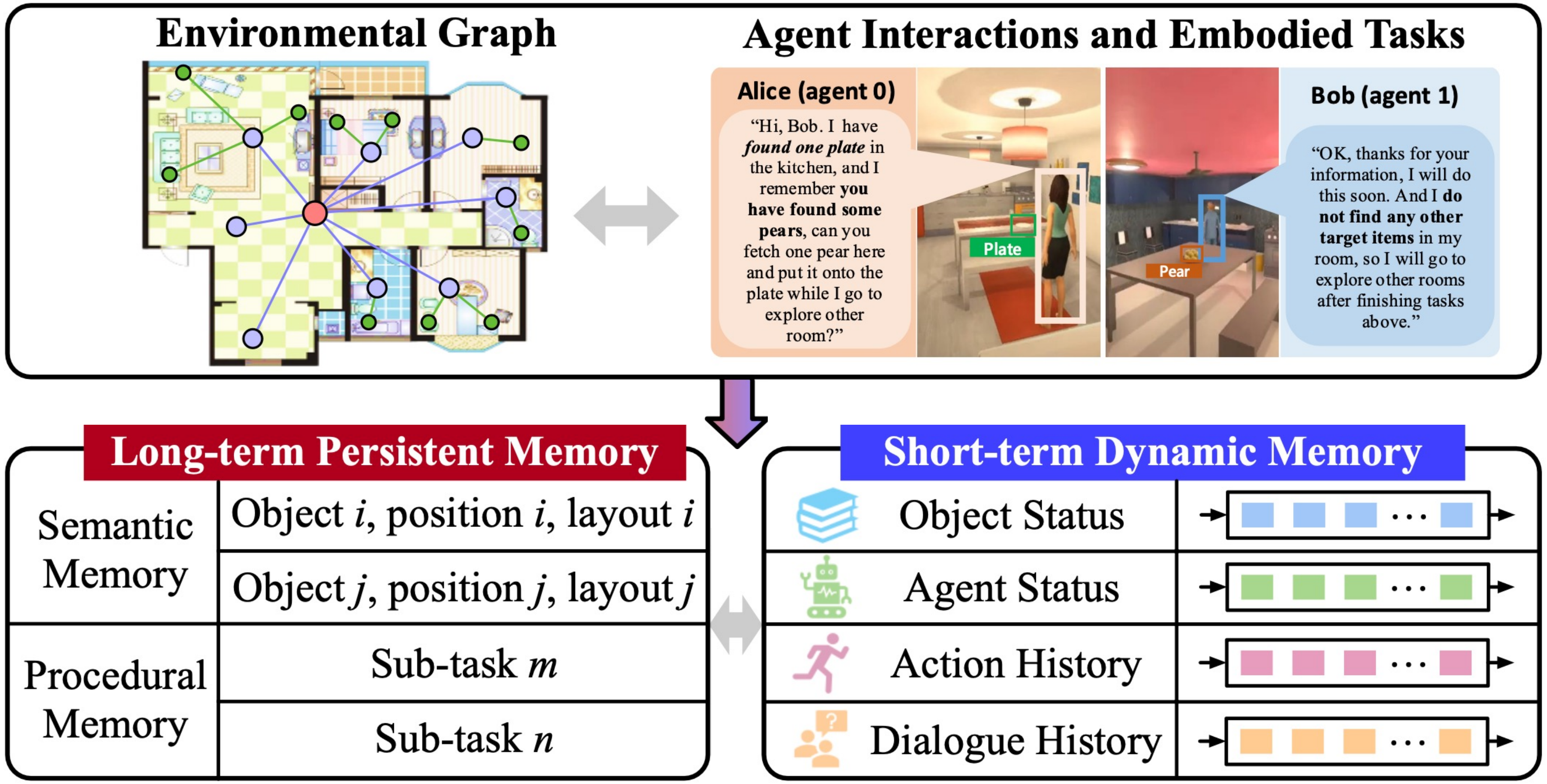
Our Methodology



Our Methodology

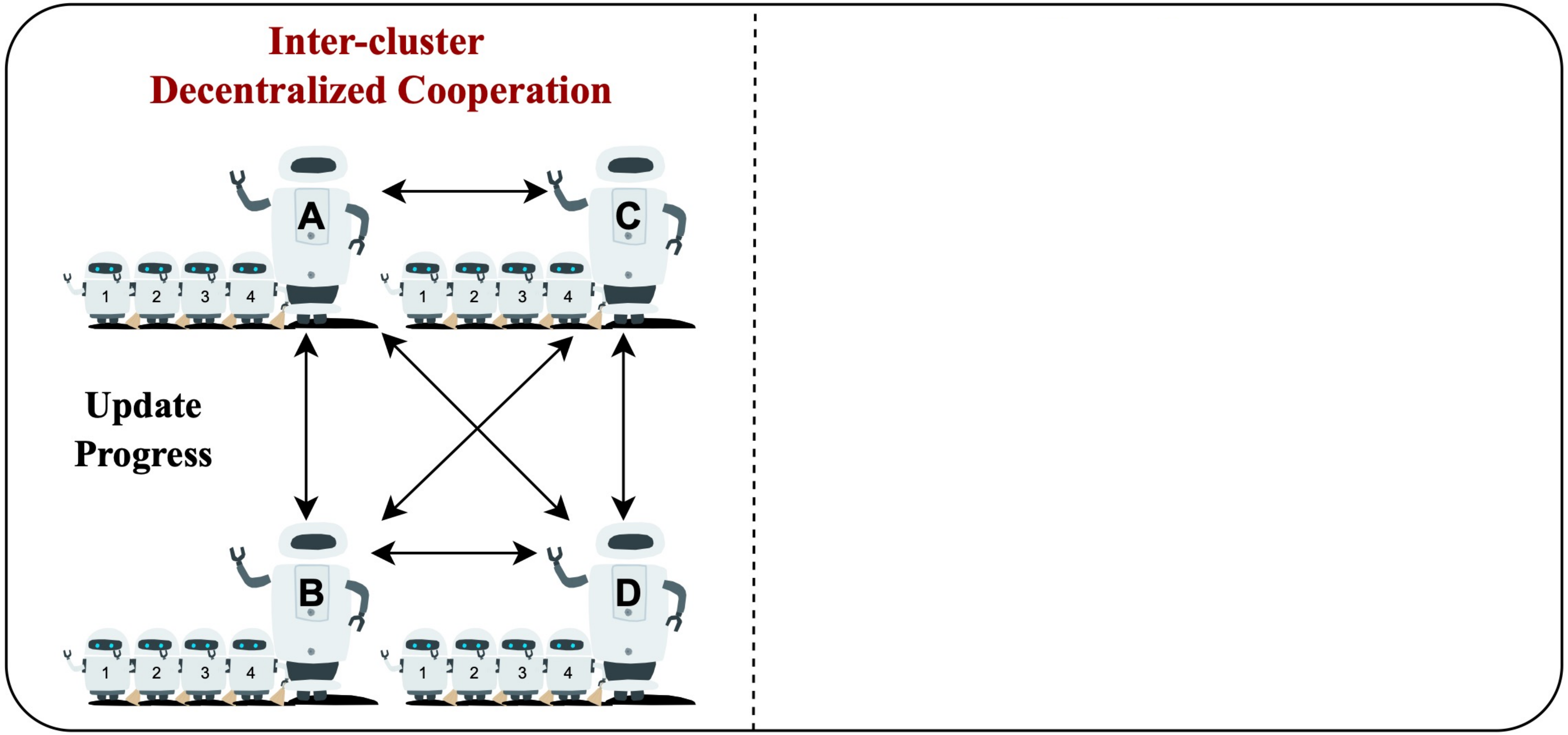


System Optimization – Dual Memory Structure



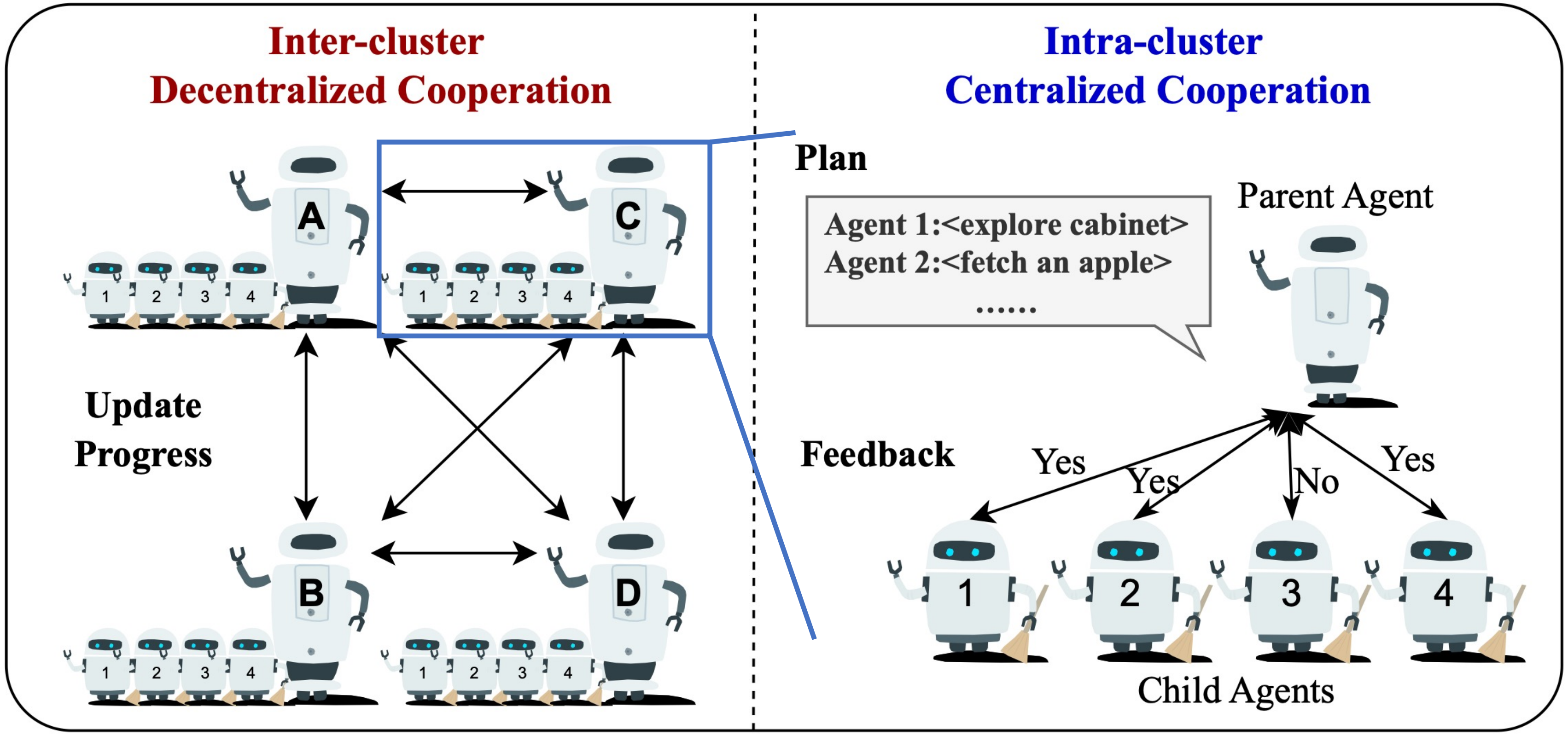
- ❑ Dual-memory structure for agentic systems:
  - ❑ Long-term memory: subtask and environment info
  - ❑ Short-term memory: action, dialog, agent history (periodically update)

System Optimization - Hierarchical Cooperative Planning



- ❑ Hierarchical cooperative planning for agentic systems:
  - ❑ Inter-cluster decentralized cooperation

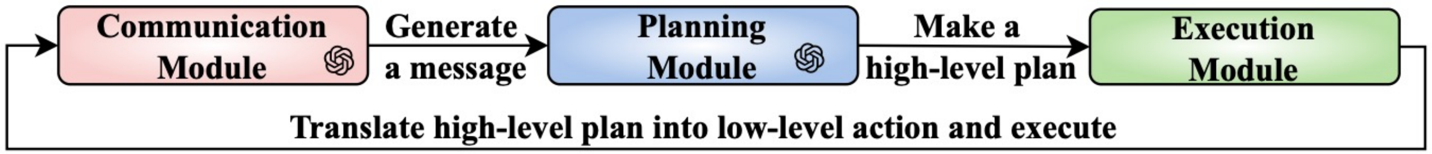
System Optimization - Hierarchical Cooperative Planning

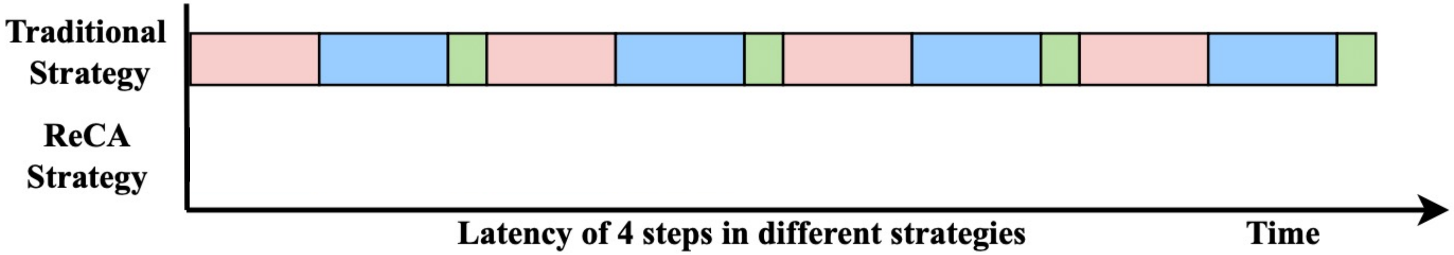


- ❑ Hierarchical cooperative planning for agentic systems:
  - ❑ Inter-cluster decentralized cooperation
  - ❑ Intra-cluster centralized cooperation

System Optimization – Execution Pipeline

Baseline embodied system pipeline

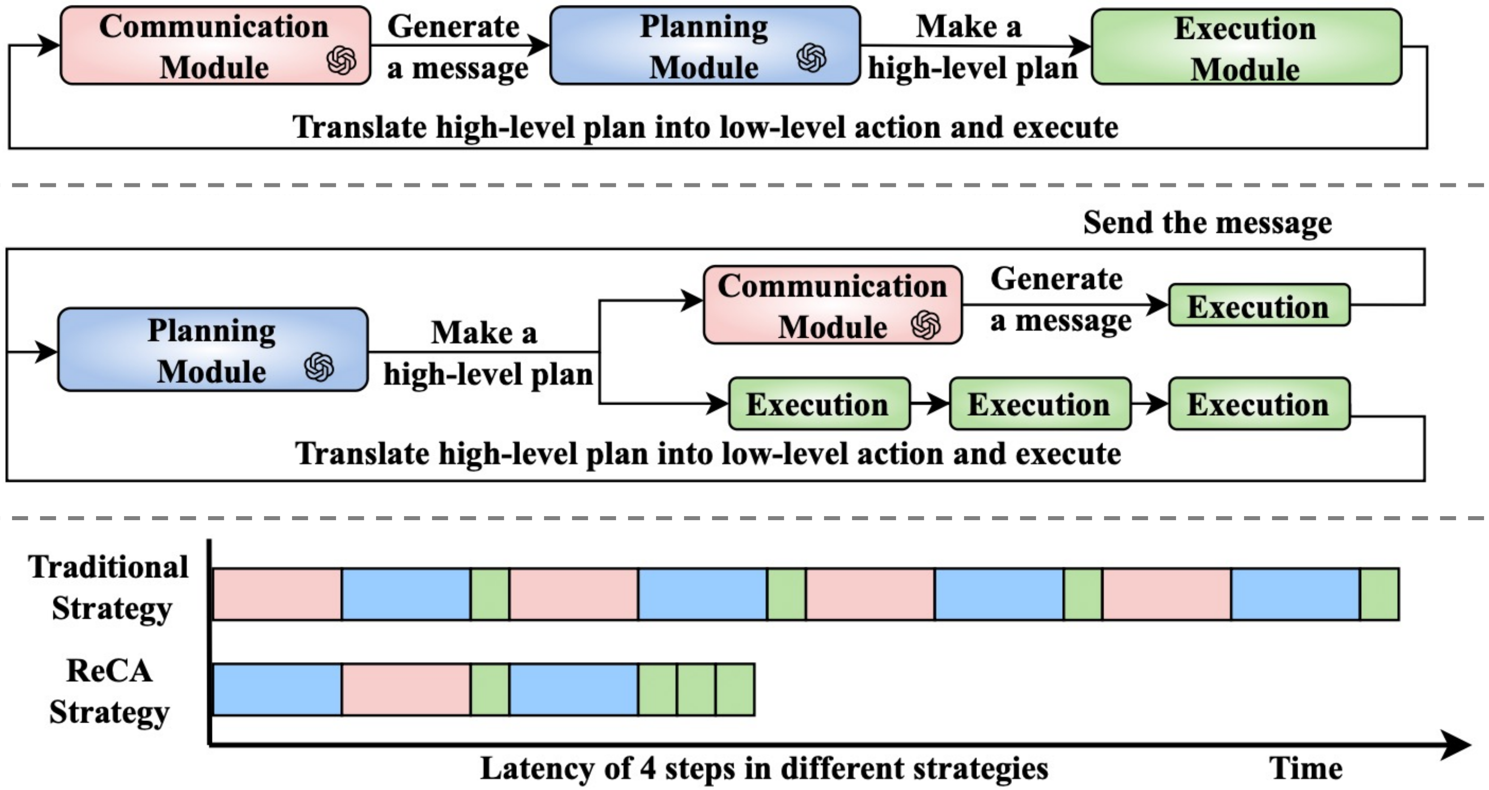




System Optimization – Execution Pipeline

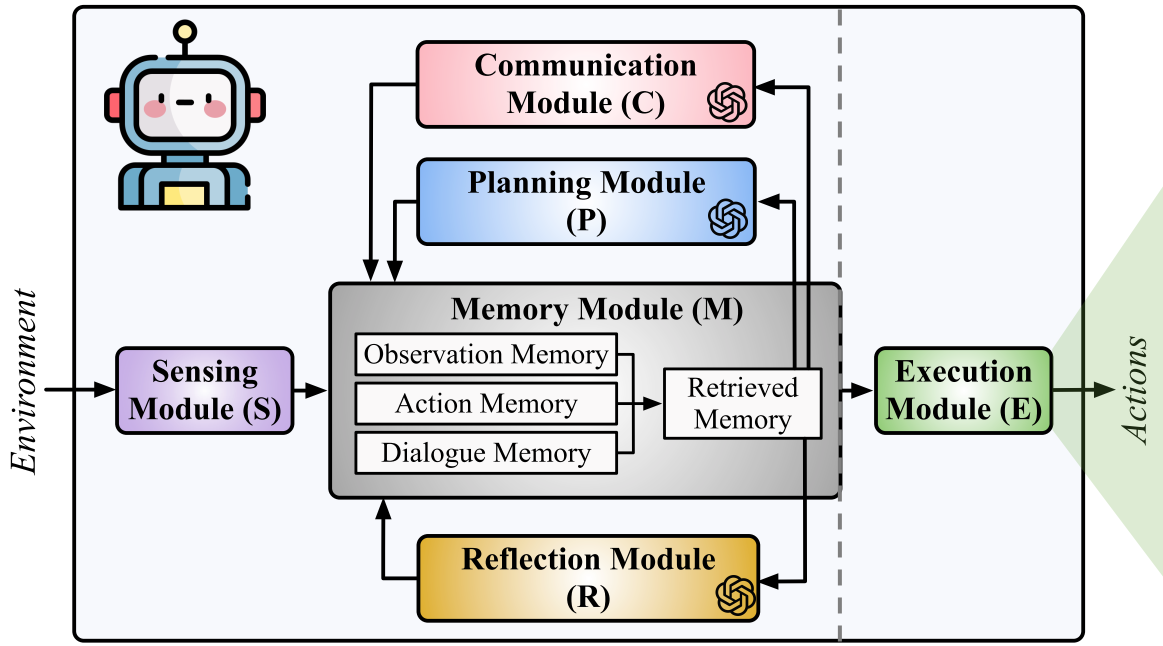
Baseline embodied system pipeline

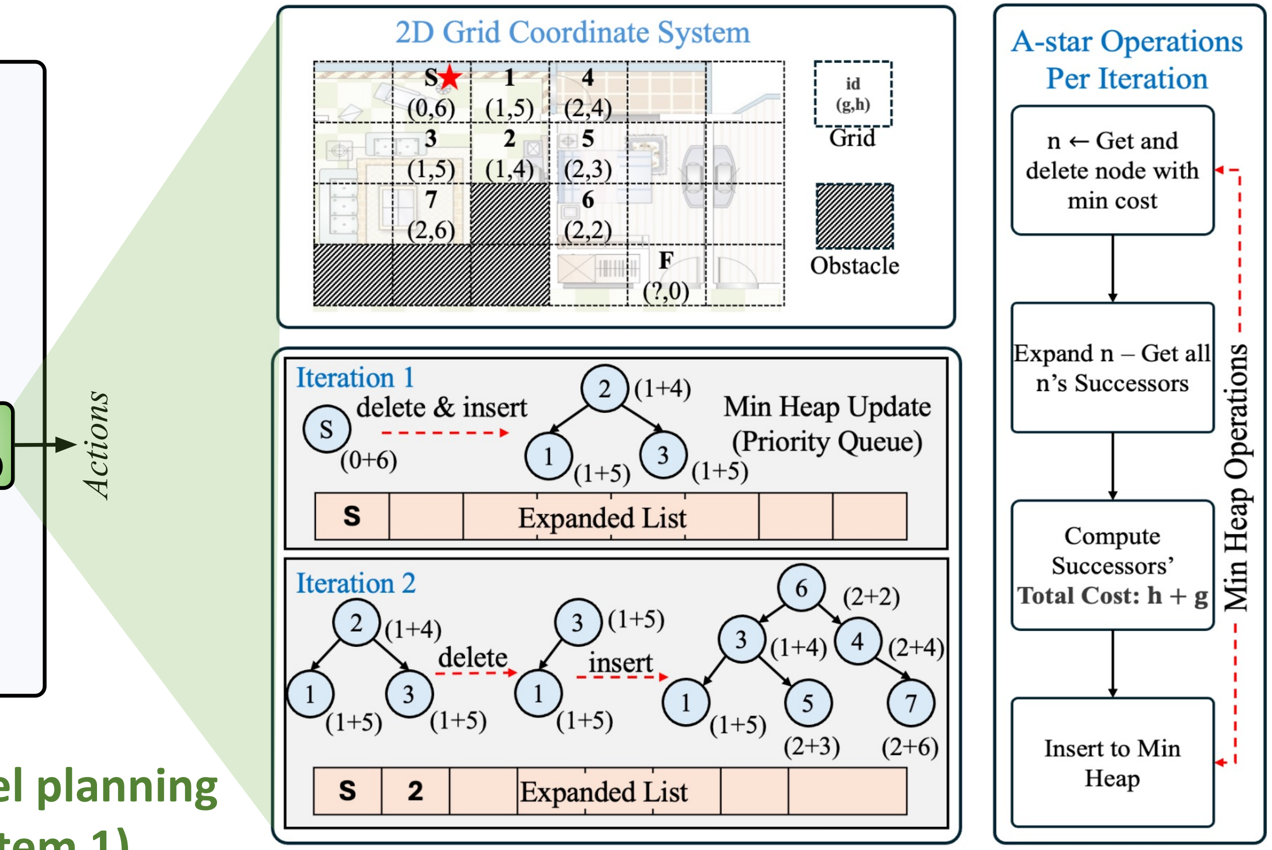
Optimized embodied system pipeline



- ❑ Efficient execution pipeline
  - ❑ Planning-then-communication strategy
  - ❑ Planning-guided multi-step execution

Hardware Optimization – Heterogenous SoC





High-level reasoning (System 2)

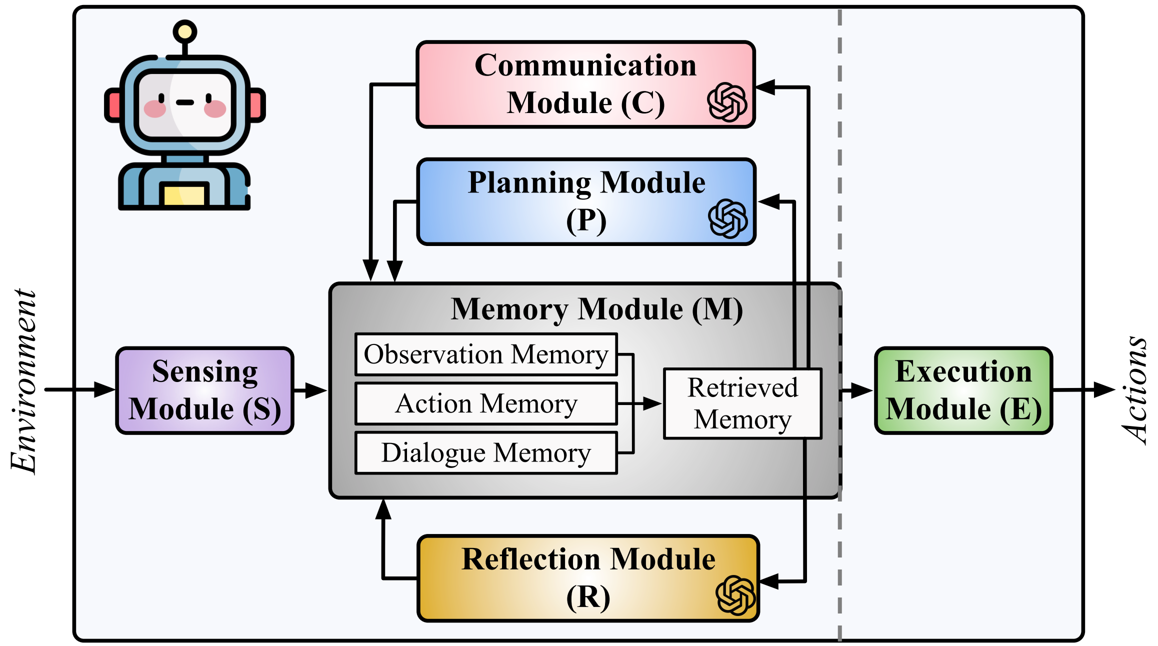
LLM-based operations

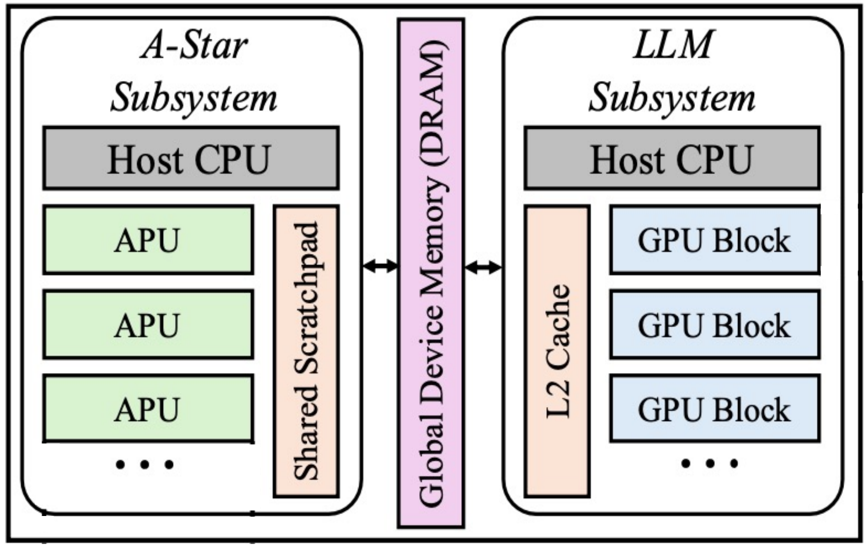
Low-level planning (System 1)

Search/solver-based operations

Example: A-based path planning*

Hardware Optimization – Heterogenous SoC





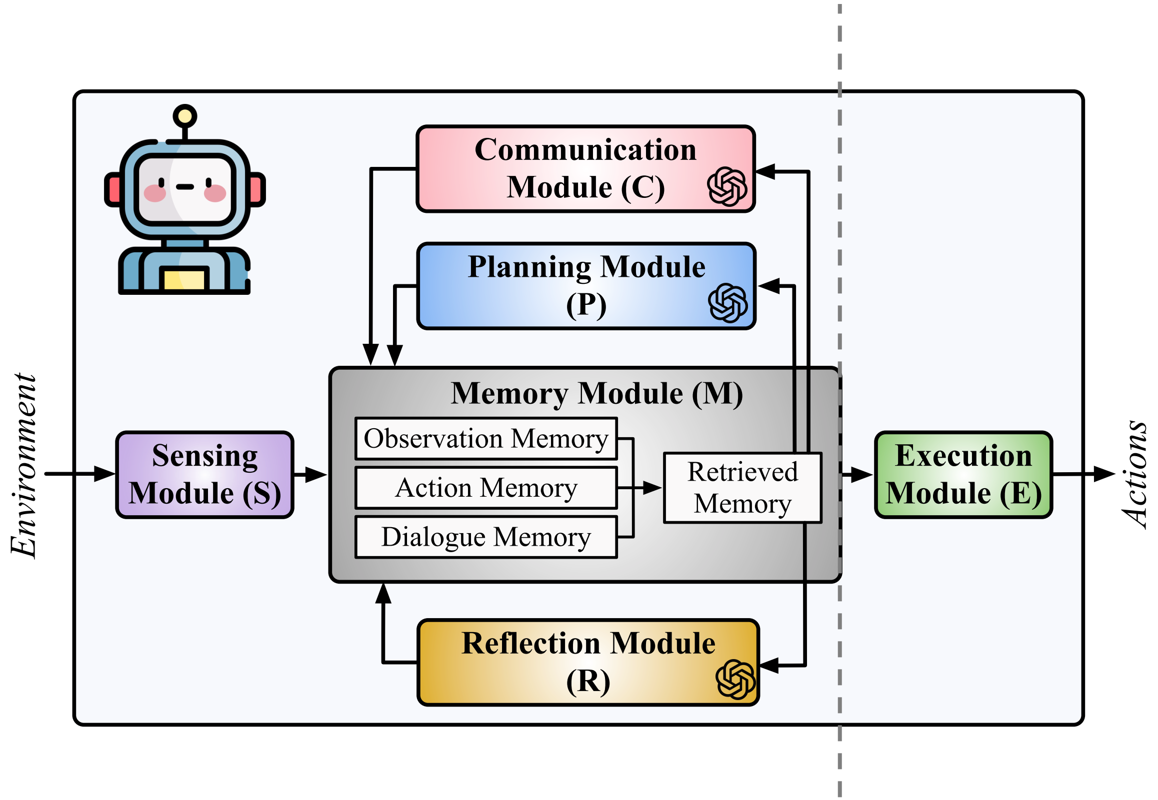
High-level reasoning (System 2)

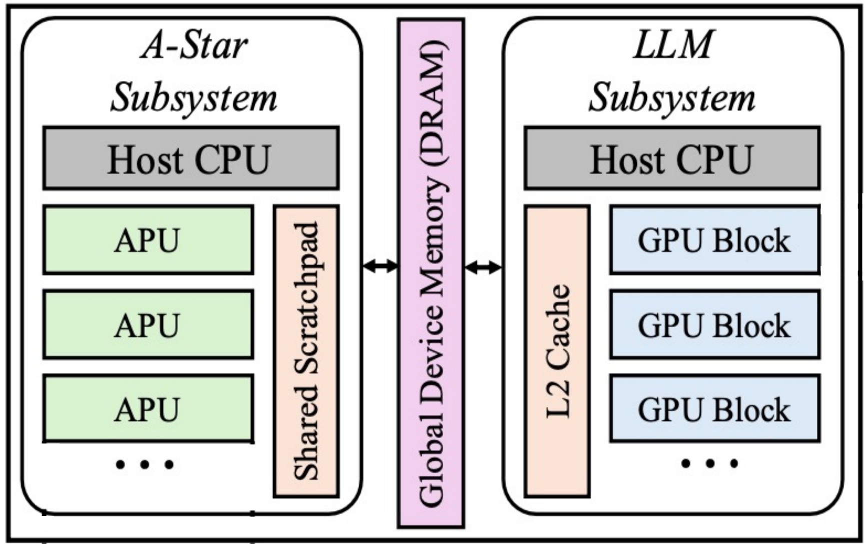
Low-level planning (System 1)

A*-Subsystem (System 1)

LLM-Subsystem (System 2)

Hardware Optimization – Heterogenous SoC



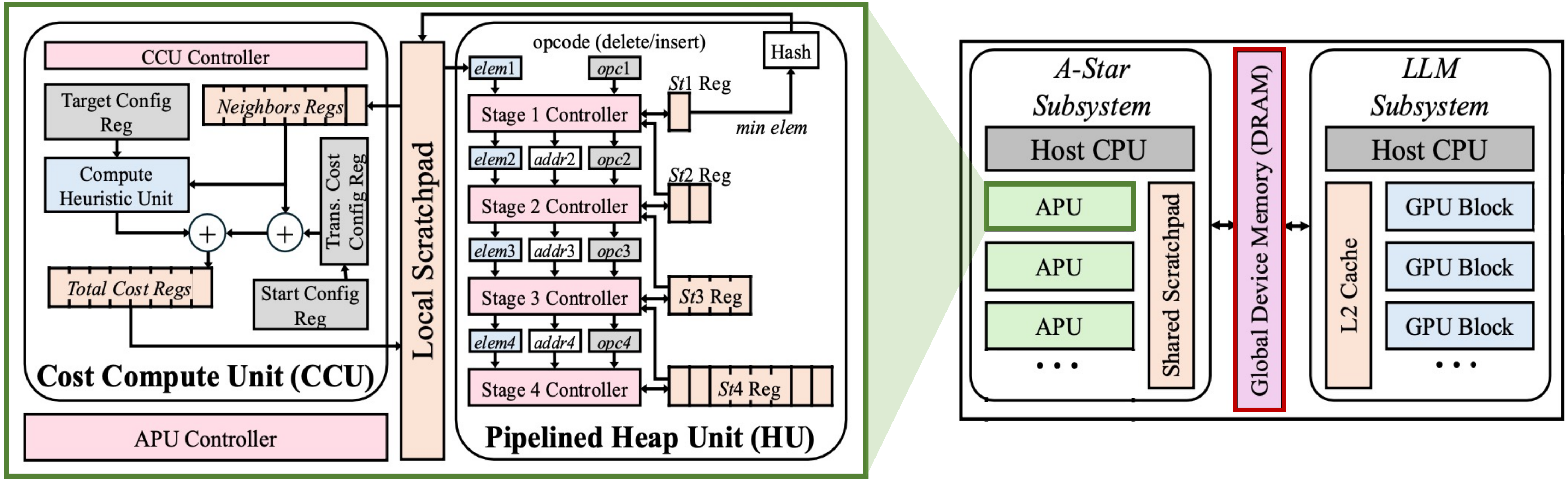


A*-Subsystem (System 1)

LLM-Subsystem (System 2)

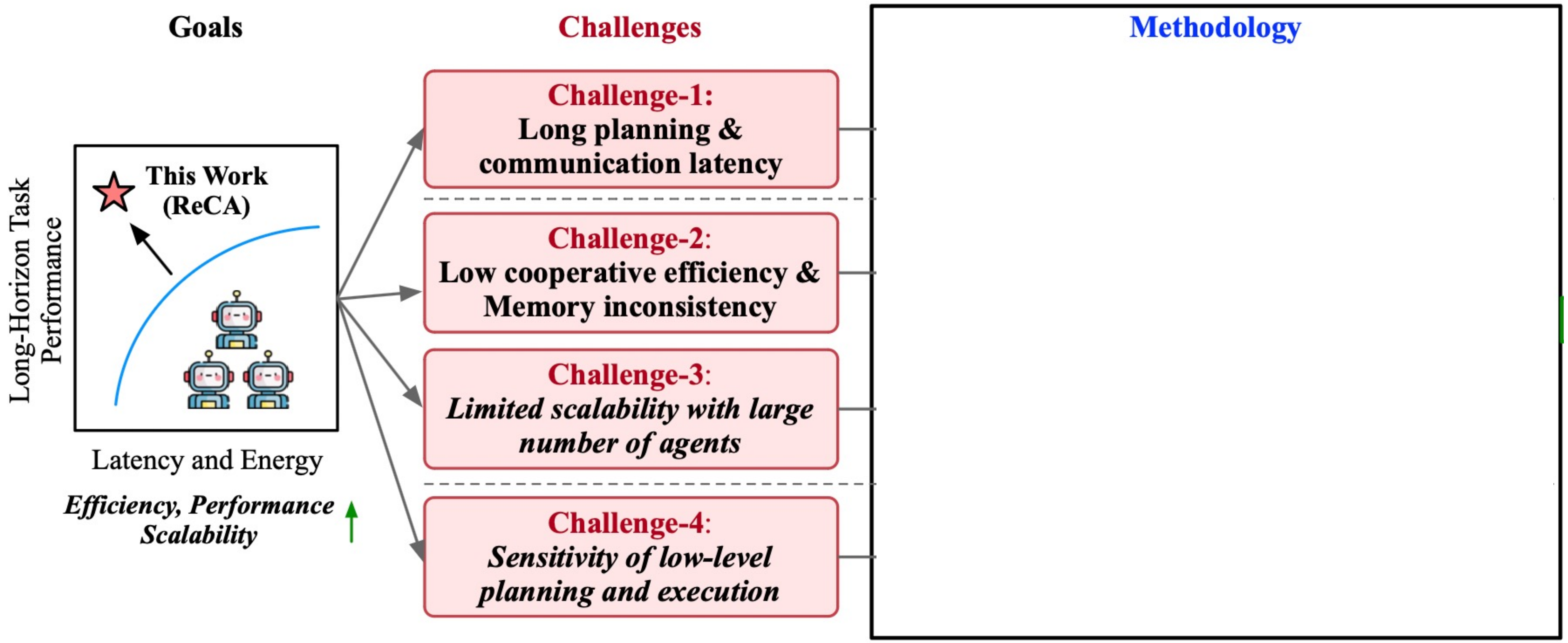
- ❑ Hardware system for embodied agent systems:
  - ❑ LLM Subsystem: for high-level decision making and communication
  - ❑ Control Subsystem: for low-level planning and action

Hardware Optimization – Heterogenous SoC

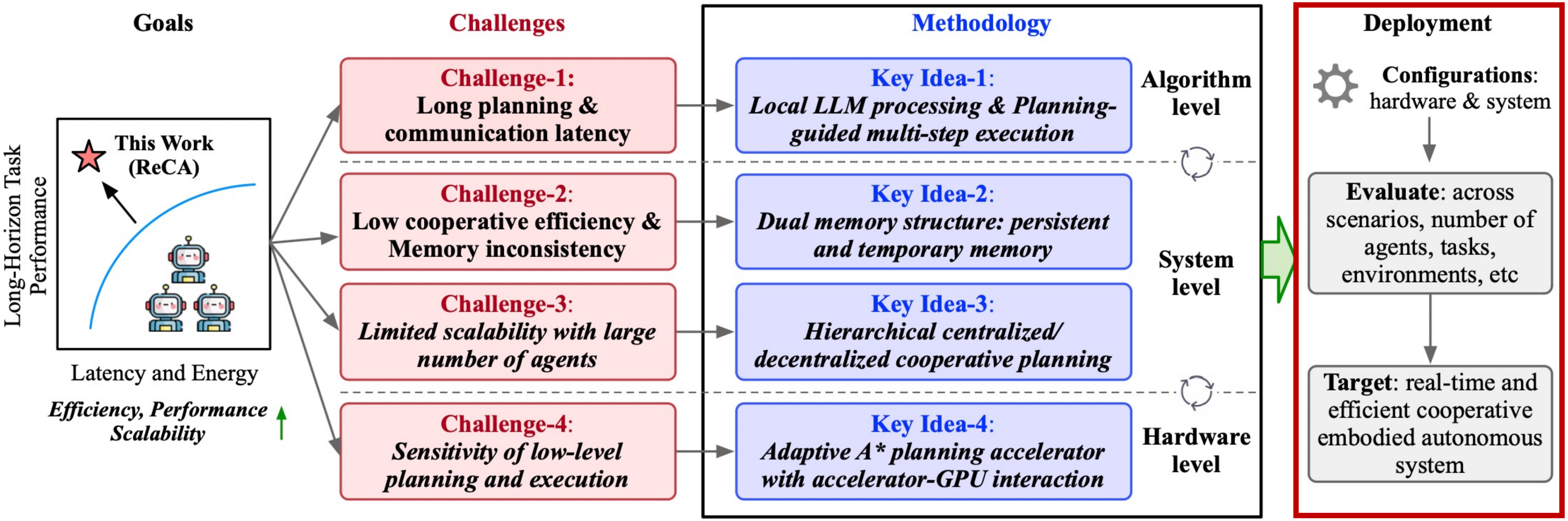


- ❑ Microarchitecture of low-level subsystem:
  - ❑ **Cost Compute Unit (CCU)**: for cell cost evaluation
  - ❑ **Pipelined Heap Unit (HU):** for priority queue management
  - ❑ **Scratchpad memory**: for storing neighboring cell during node expansion

Optimizations of Embodied Agent Systems

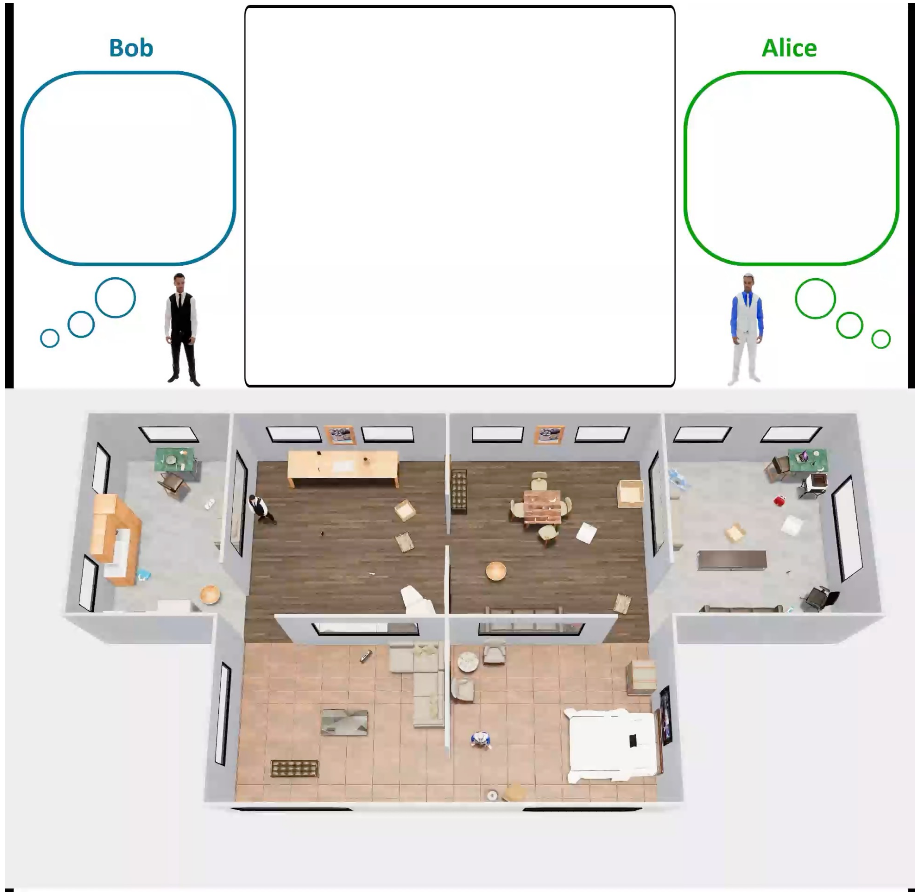


Evaluation


Goals
Challenges
Methodology
Deployment
This Work (ReCA)
Long-Horizon Task Performance
Latency and Energy
Efficiency, Performance Scalability
Challenge-1:
Long planning & communication latency
Challenge-2:
Low cooperative efficiency & Memory inconsistency
Challenge-3:
Limited scalability with large number of agents
Challenge-4:
Sensitivity of low-level planning and execution
Key Idea-1:
Local LLM processing & Planning-guided multi-step execution
Key Idea-2:
Dual memory structure: persistent and temporary memory
Key Idea-3:
Hierarchical centralized/ decentralized cooperative planning
Key Idea-4:
Adaptive A* planning accelerator with accelerator-GPU interaction
Algorithm level
System level
Hardware level
Configurations: hardware & system
Evaluate: across scenarios, number of agents, tasks, environments, etc
Target: real-time and efficient cooperative embodied autonomous system

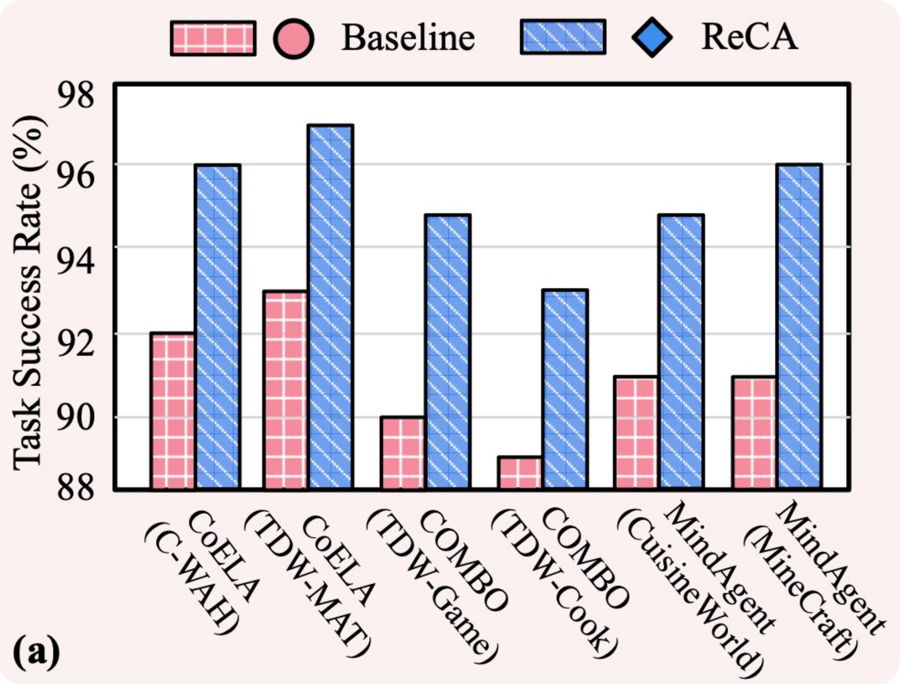
Evaluation - Setup

- **Embodied Workloads**:
  - CoELA, COMBO, MindAgent
- **Long-horizon Tasks**:
  - TDW-MAT, TDW-Cook, TDW-Game, CuisineWorld, C-WAH, MineCraft
- **Metrics**:
  - Task success rate, Number of steps, End-to-end runtime
- **Hardware**:
  - NVIDIA A6000 GPU (for LLM-subsystem)
  - Xilinx Zynq-7000 ZC706 FPGA (for control-subsystem)



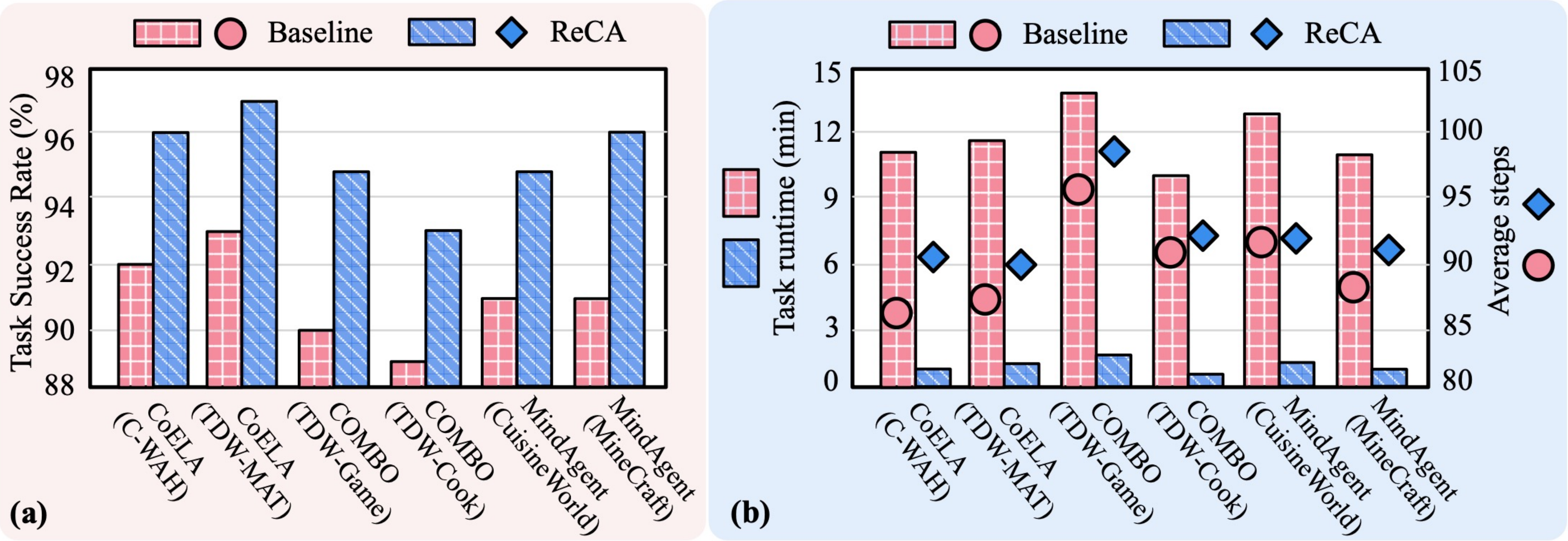
Example: C-WAH task

Evaluation – Success Rate and Efficiency Improvement



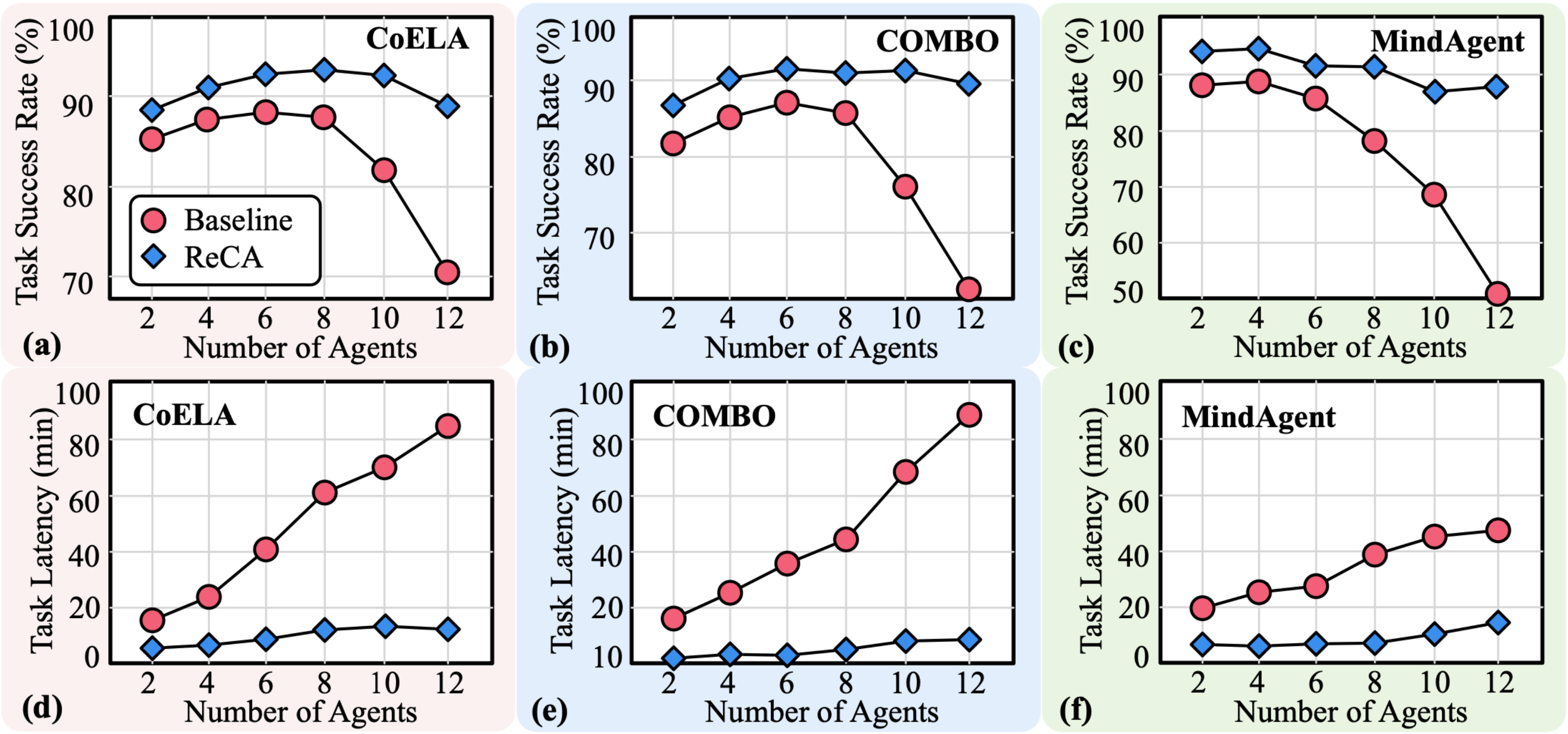
Improved success rate: ReCA increases task success rate by 4% on average.

Evaluation – Success Rate and Efficiency Improvement



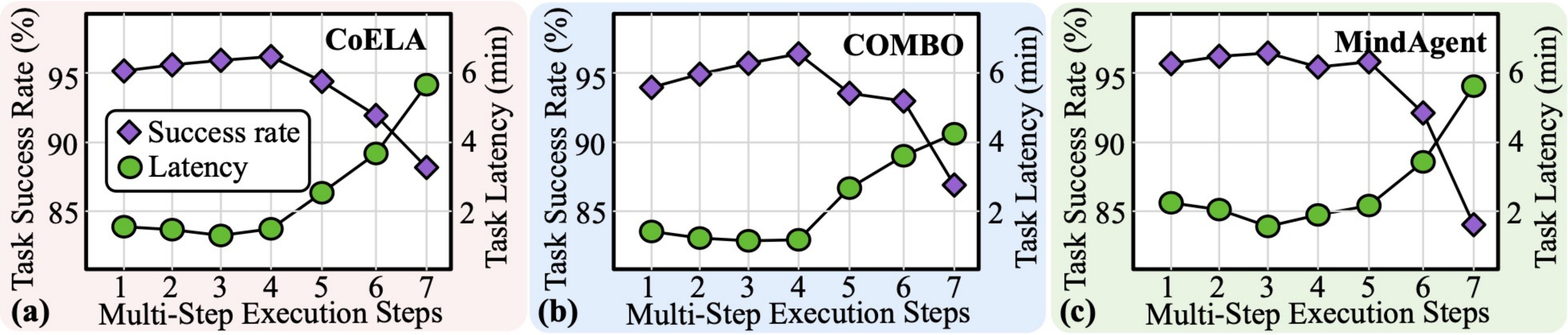
Improved success rate: ReCA increases task success rate by 4% on average.
Improved efficiency: ReCA reduces end-to-end task runtime by 8.4x on average.

Evaluation – Scalability Improvement



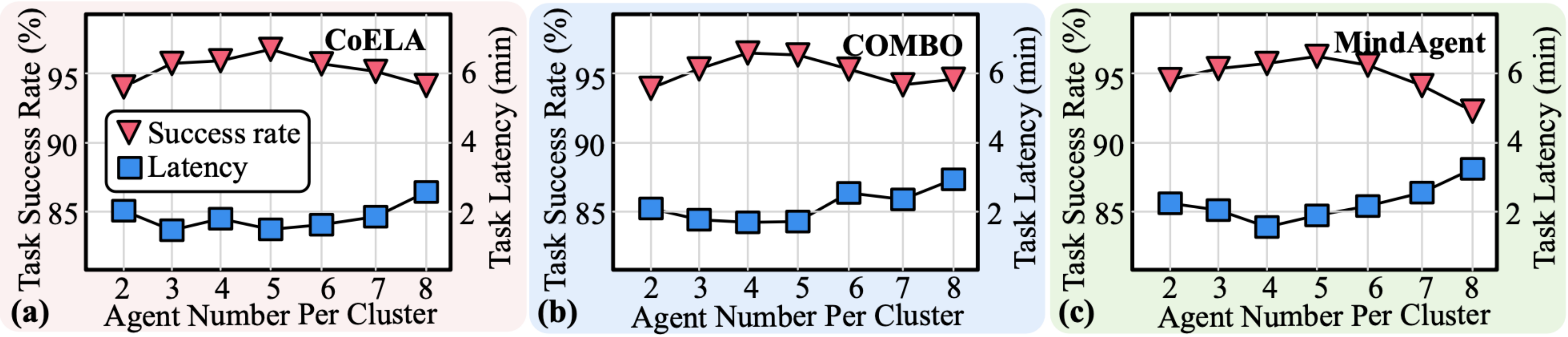
Improved scalability: ReCA **scales well** in both decentralized embodied systems (CoELA, COMBO) and centralized embodied systems (MindAgent).

Evaluation – Sensitivity across Multi-Step Execution



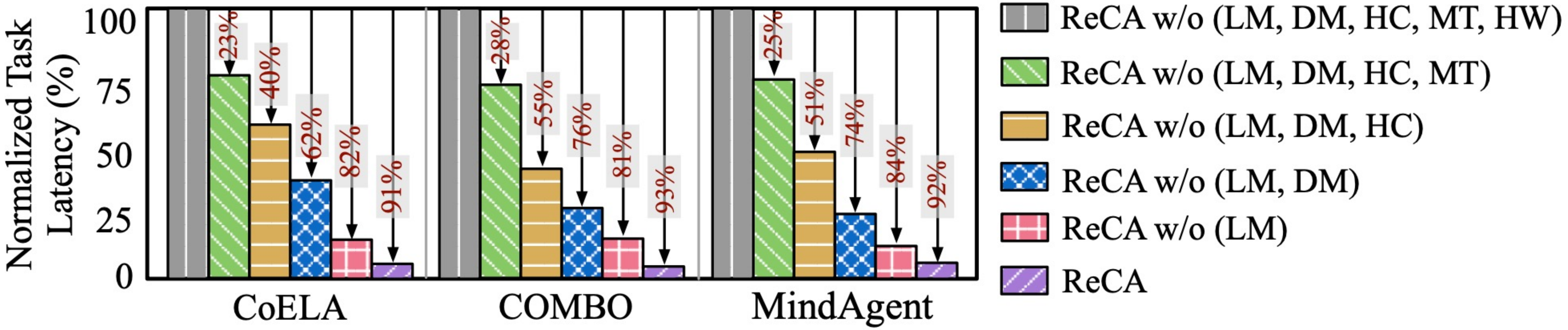
Multi-Step Execution Steps: ReCA exhibits optimal task performance and efficiency under **4-5 action steps per LLM reasoning run**.

Evaluation – Sensitivity across Hierarchical Planning



Hierarchical Cooperative Planning: ReCA exhibits optimal task performance and efficiency under **5-agent per cluster**.

Evaluation – Ablation Study



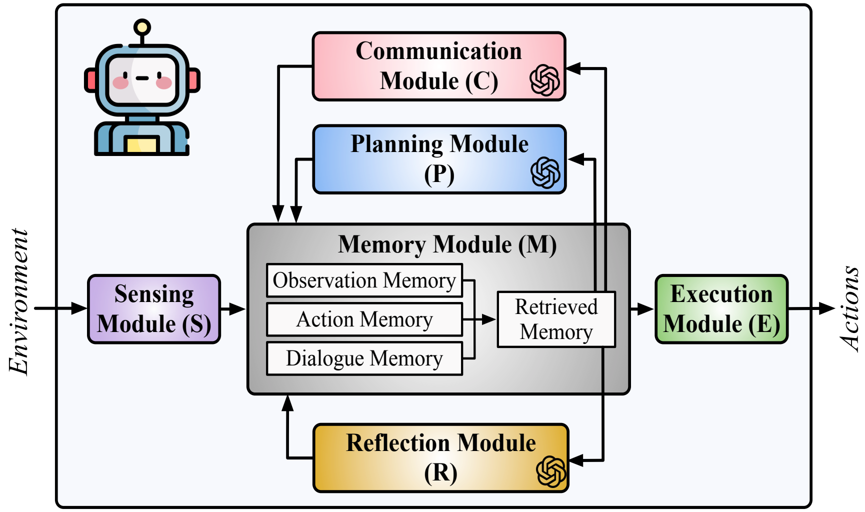
LM: local model (Sec.5.1) DM: dual memory (Sec.5.2) HC: hierarchical cooperation (Sec.5.3)
MT: multi-step execution (Sec.5.4) HW: A-star/GPU heterogenous hardware system (Sec.6)

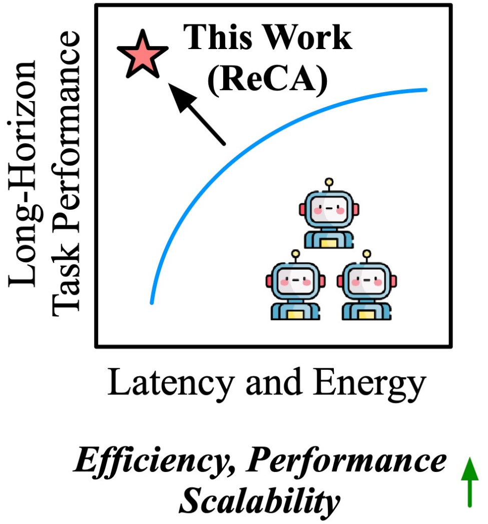
Proposed dual-memory, hierarchical cooperation, multi-step execution, and heterogenous architecture **optimizations are effective**.

Model-system-hardware co-design is critical for system performance.

ReCA Summary

- **Embodied agents** integrate perception, cognition, and physical action to conduct long-horizon tasks
- In this work,
  - Characterize **system implications**
  - Leverage **co-design intelligence**
    - **Algorithm**: efficient local LLM deployment
    - **System**: dual-memory structure, hierarchical planning, and planning-guided multi-step execution
    - **Hardware**: heterogenous architecture for high-level reasoning and low-level control
  - Achieve **efficient and scalable embodied AI systems** across cooperative long-horizon multi-objective tasks





ReCA: Integrated Acceleration for Real-Time and Efficient Cooperative Embodied Autonomous Agents

Zishen Wan[1], Yuhang Du[2], Mohamed Ibrahim[1], Jiayi Qian[1], Jason Jabbour[3], Yang (Katie) Zhao[2], Tushar Krishna[1], Arijit Raychowdhury[1], Vijay Janapa Reddi[3]

Email: *zishenwan@gatech.edu*
Webpage: *https://zishenwan.github.io*

ACM Inter Conf on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2025