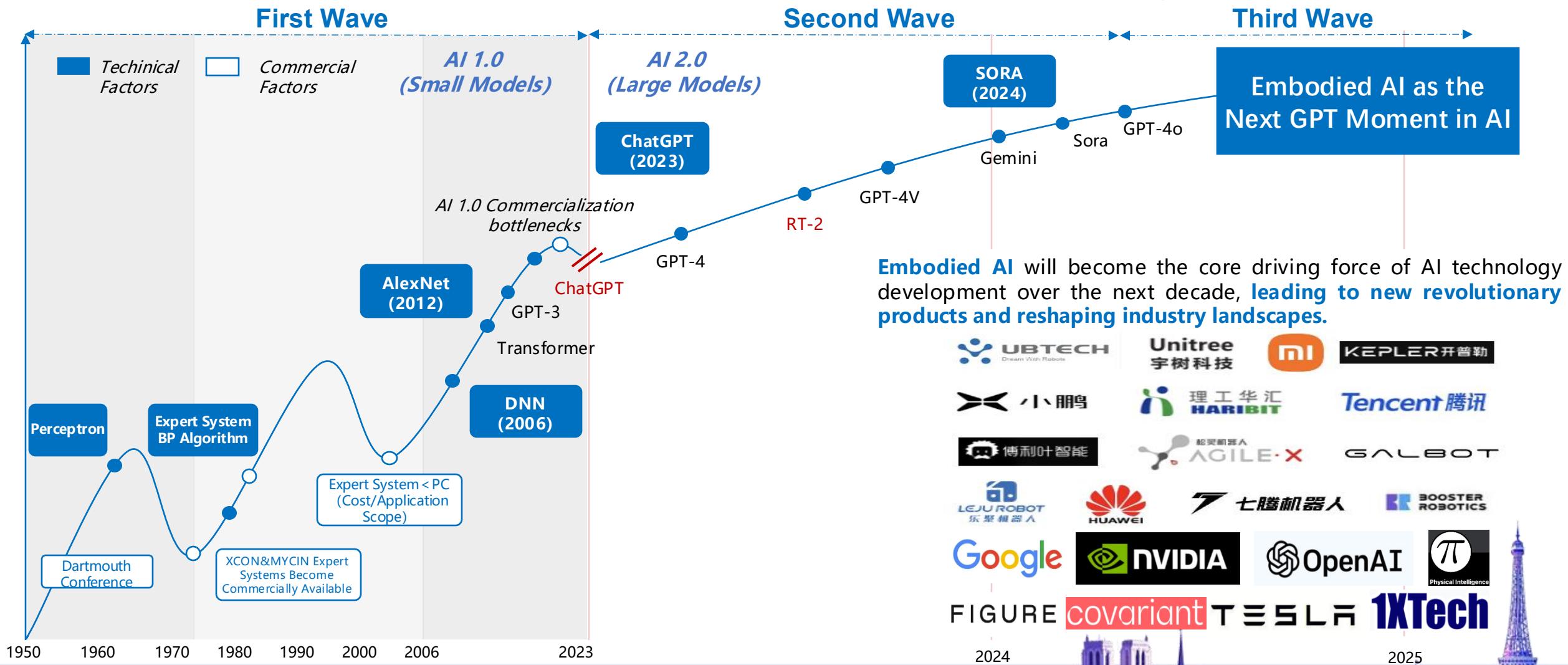


RoboBrain: A Unified Brain Model for Robotic Manipulation & RoboOS: A Hierarchical Collaborative Framework for RoboBrain and Robot Agents

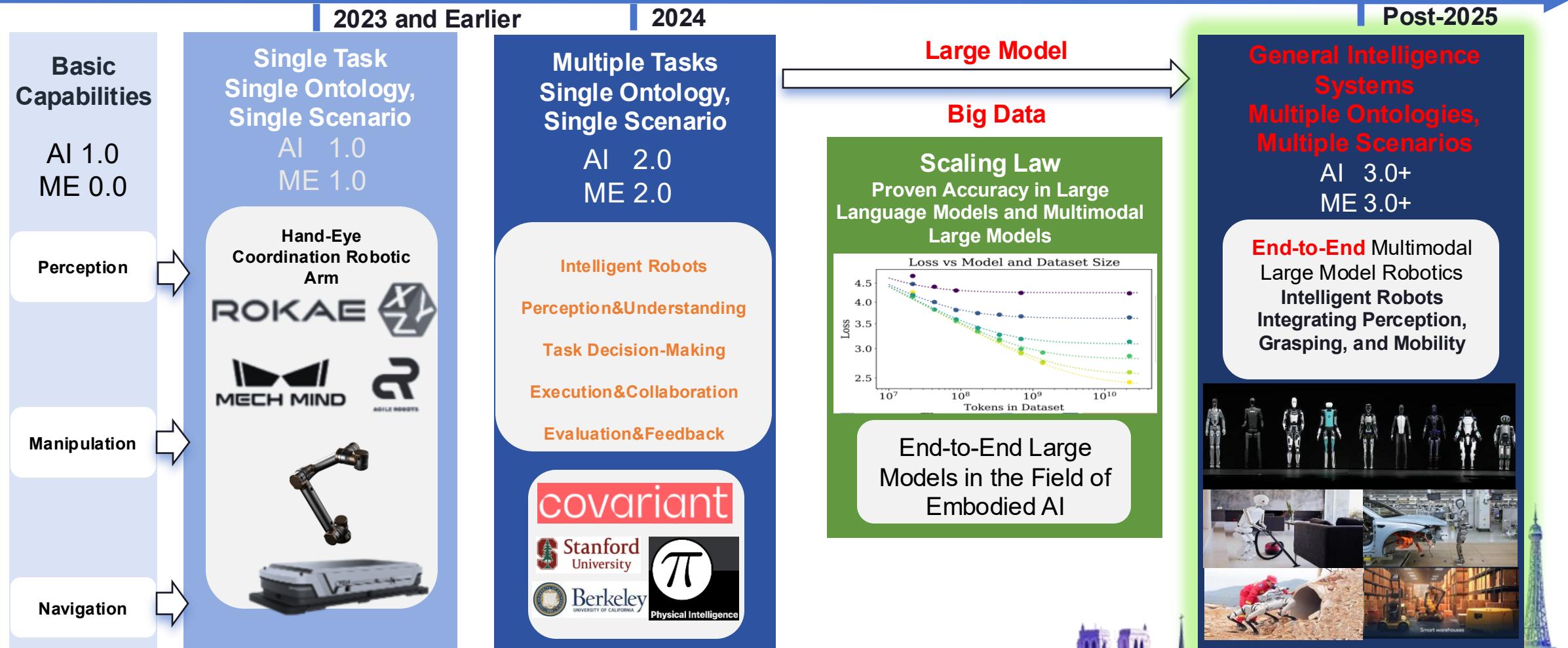
Minglan Lin

Embodied Intelligence Researcher, BAAI

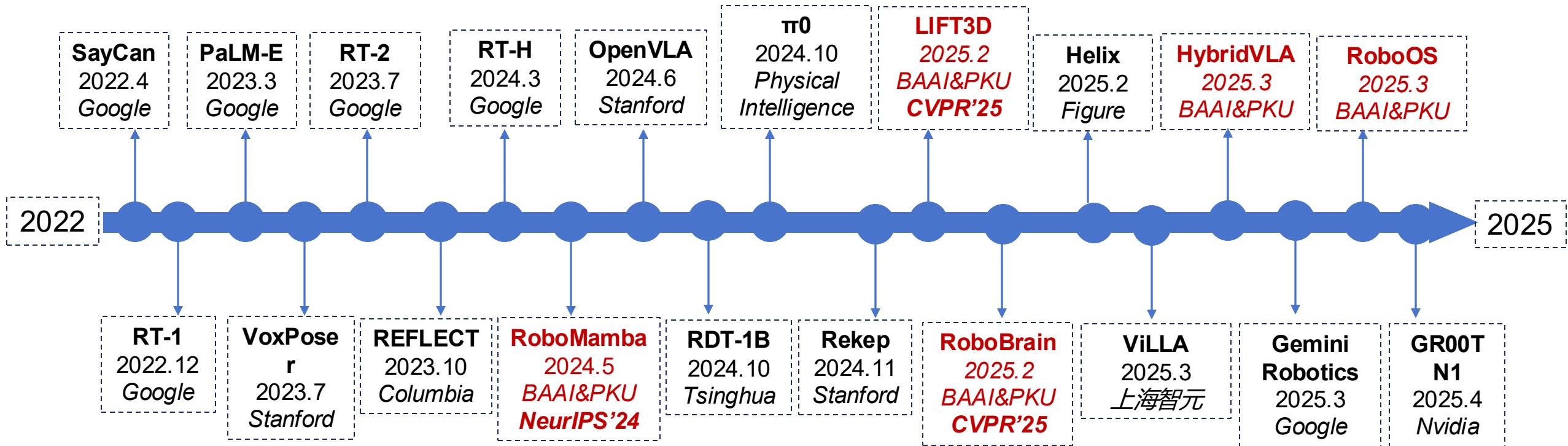
Trends of Artificial Intelligence



Trends of Embodied Intelligence



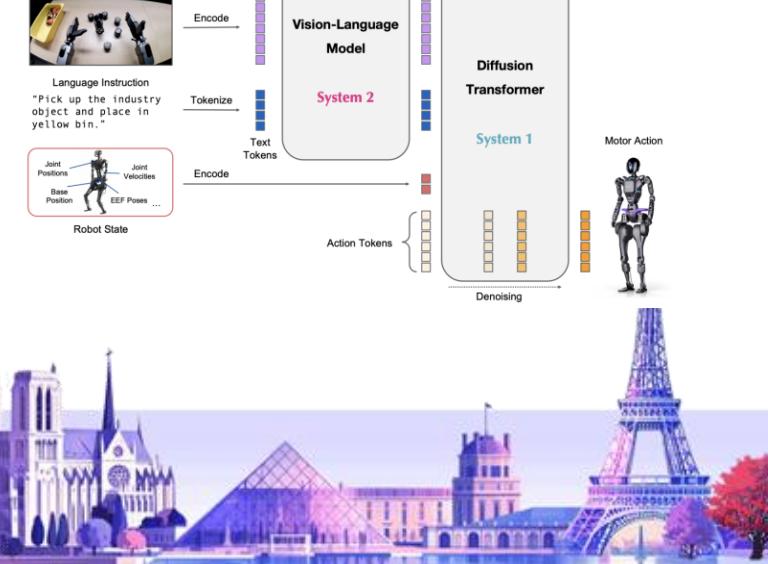
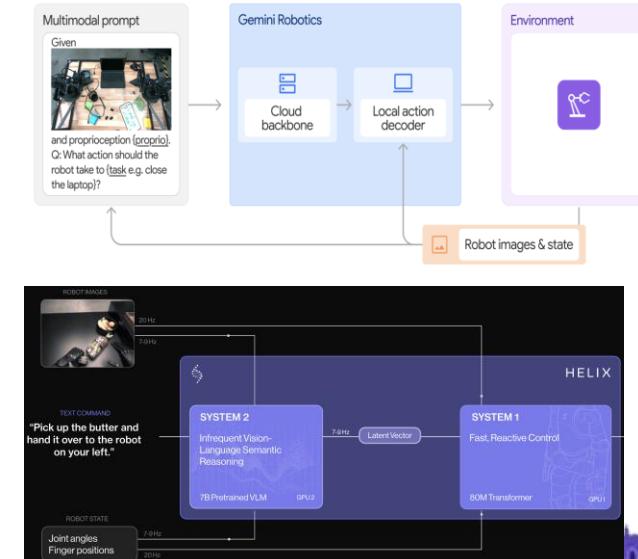
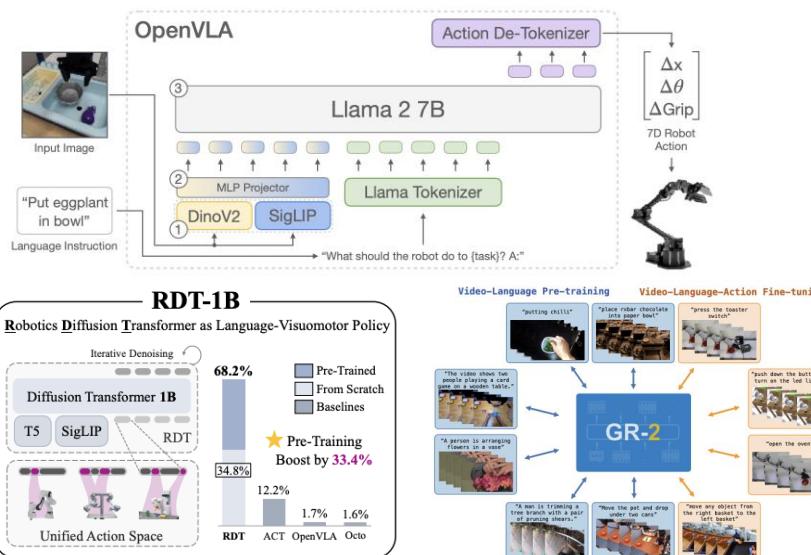
Milestone in Embodied Intelligence



Pathways to Embodied Intelligent



- The realization of embodied intelligent agents are categorized into two major technical approaches:
 - End-to-End Vision-Language-Action Models:
 - OpenVLA, PI0, RDT, RT Series, GR Series, etc.
 - Brain-Cerebellum Collaborative Frameworks:
 - RoboOS (BAAI)**, Helix, Gemini Robotics, ReKep, etc.



Why Choose a Brain-Cerebellum Collaborative Framework ?



- While End-to-End models offer efficient decision-making, they suffer from **limited generalization and scalability** due to constraints in **environment interaction and hardware adaptation**, making it difficult to handle diverse real-world scenarios. In contrast, modular Brain-Cerebellum collaborative framework --- known for their **strong generalization and interpretability** --- are emerging as a key focus in both academic and industrial research.

- 01** / **Modularization**: The Brain-Cerebellum collaborative framework equips embodied intelligent agents with modular advantages, featuring scalable architecture, efficient development, and strong adaptability.
- 02** / **Generalization**: The VLM-based brain model exhibits strong generalization with rich multimodal cognitive abilities, independent of cerebellar model constraints.
- 03** / **Interpretability**: A more transparent decision-making process enhances human-machine collaboration efficiency.

The Brain-Cerebellum collaborative framework represents a simpler and more effective approach to realize embodied intelligence.



RoboBrain: **A Unified Brain Model for** **Robotic Manipulation**



Can traditional VLMs serve as the "brain"?



Although multimodal large language models (VLMs, such as GPT-4o) excel in general tasks, their application in embodied scenarios faces significant challenges, particularly in **the long-horizon operational capabilities** required for robots to perform complex tasks.



If the yellow robot gripper follows the yellow trajectory, what will happen?

- A. Robot puts the soda on the wooden steps.
- B. Robot moves the soda in front of the wooden steps.
- C. Robot moves the soda to the very top of the wooden steps.
- D. Robot picks up the soda can and moves it up.



GPT-4o performs poorly in embodied tasks.



E.g. "putting the pot in the drawer"

This task involves multi-step, long-horizon interactions, including moving, grasping, and placing, while requiring continuous interaction with objects such as the pot and the drawer.

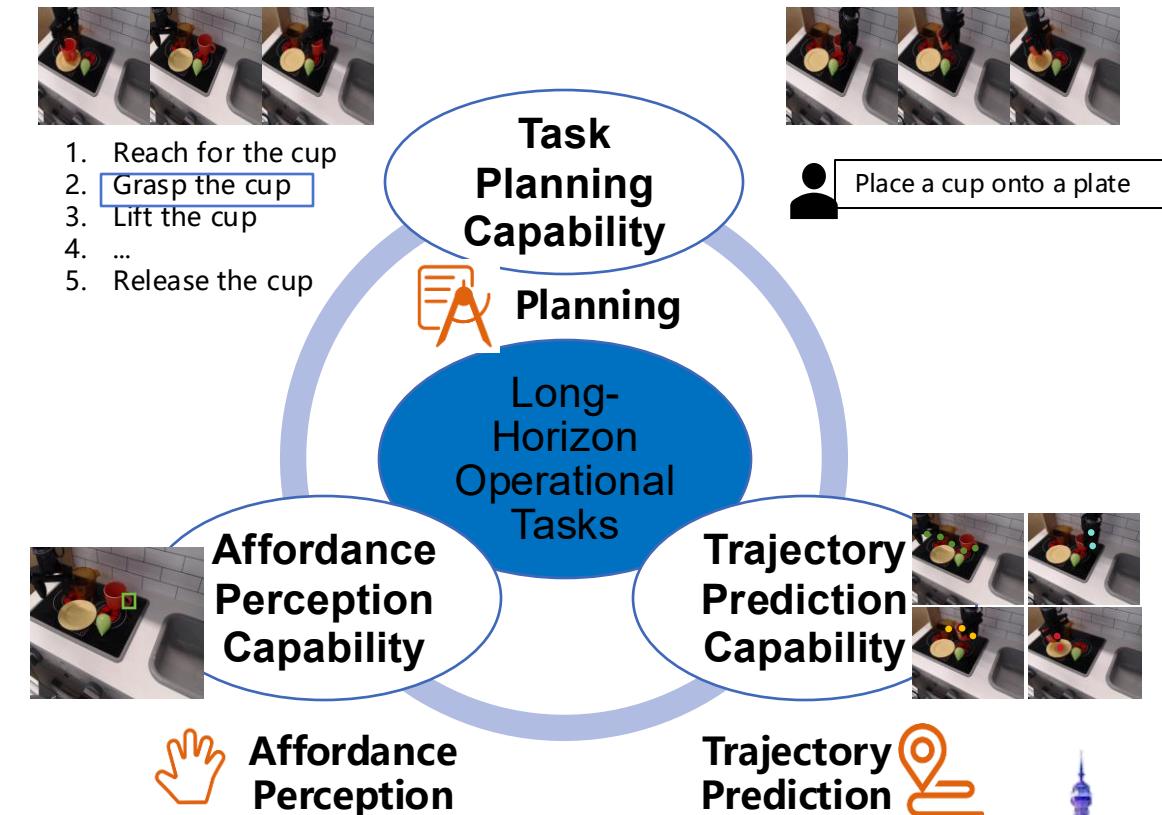
Given the objects in the image, if you are required to complete the task "**Put the pot in the drawer**", what is your detailed plan? Write your plan and explain it in detail, using the following format: Step1: xxx\n Step2: xxx\n ... \n

Step1: Move towards the pot on the counter. Move the robotic arm towards the pot that is placed on the counter next to the sink. The arm should move in a straight line towards the pot.
Step2: Pick up the pot. Grasp the pot with the gripper of the robotic arm and lift it from the counter. Make sure to hold the pot securely to avoid dropping it.
Step3: Move the pot towards the drawer. Move the pot towards the open drawer below the counter. The arm should move in a smooth motion, keeping the pot close to the body for stability.
Step4: Place the pot in the drawer. Place the pot inside the drawer and adjust its position if necessary. Ensure that the pot is fully inside the drawer and not touching the edges or the lid.
Step5: Close the drawer. This can be done by moving the arm away from the drawer and then pushing the drawer closed manually or using a mechanism built into the robotic arm.
Step6: Move away from the drawer. Move the robotic arm away from the drawer to complete the task. The arm should move back to its original position or to a designated safe area.

Three Core Capabilities for Long-Horizon Operational Tasks



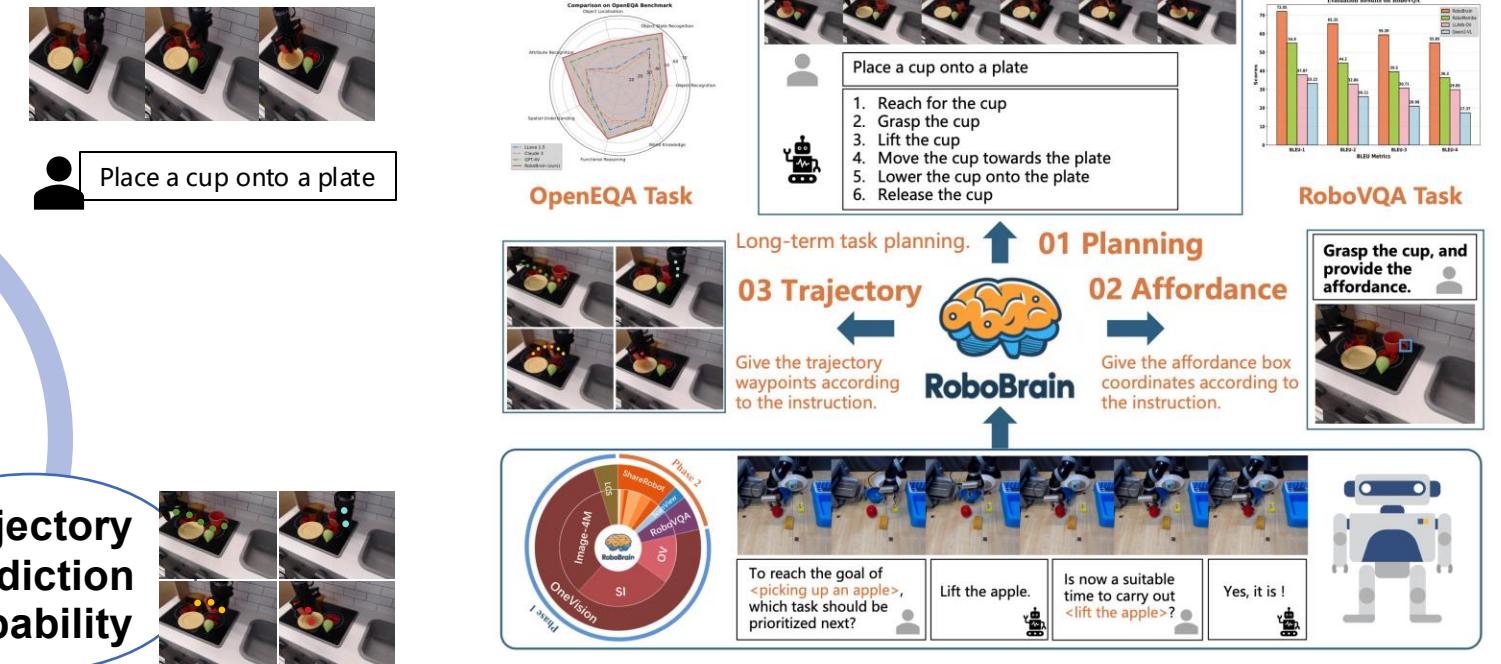
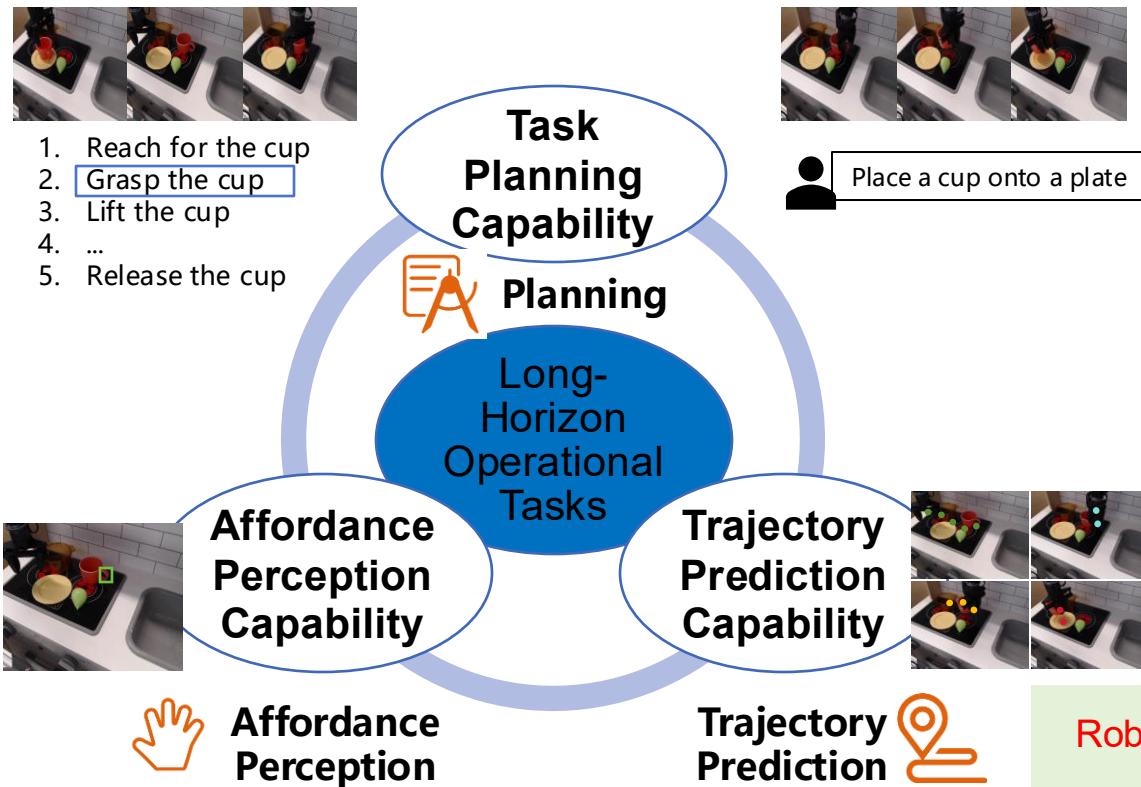
- **Task Planning Capability:** The robot must decompose abstract instructions into a sequence of executable subtasks. For example, to perform the task of "lifting a teapot and pouring water into a cup," the robot needs to sequentially complete action units such as "approaching and grasping the teapot," "aligning the spout with the cup," and "tilting the teapot to pour water."
- **Affordance Perception Capability:** The robot must accurately identify the operable regions of manipulated objects based on instructions, such as the handle or spout of a teapot, to ensure action precision.
- **Trajectory Prediction Capability:** The robot needs to predict the motion trajectory from the starting point to the target position based on task instructions.



RoboBrain: An Embodied Multimodal Brain Model for Long-Horizon Operational Tasks



Based on ShareRobot and internet data, we trained RoboBrain, an embodied multimodal brain model that bridges abstract instruction understanding to concrete action execution.

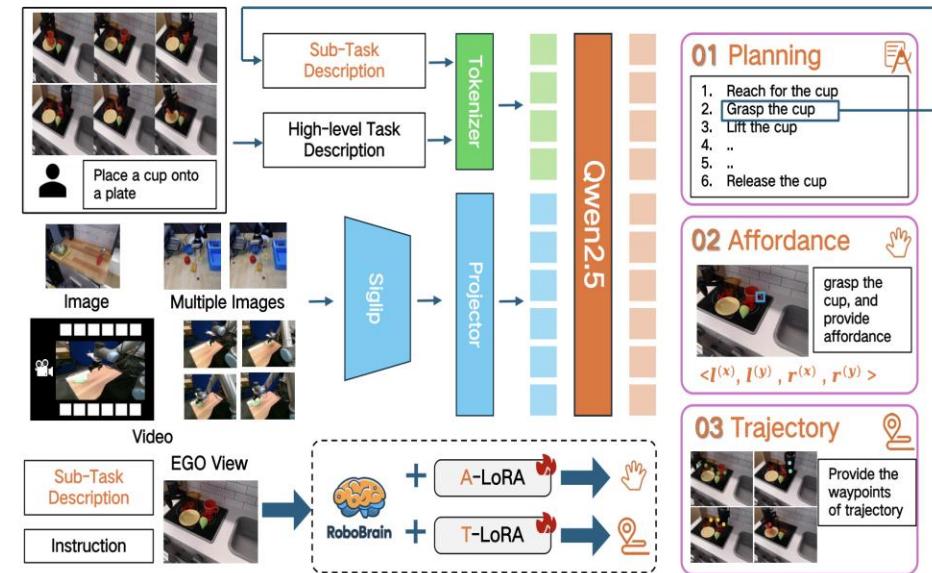
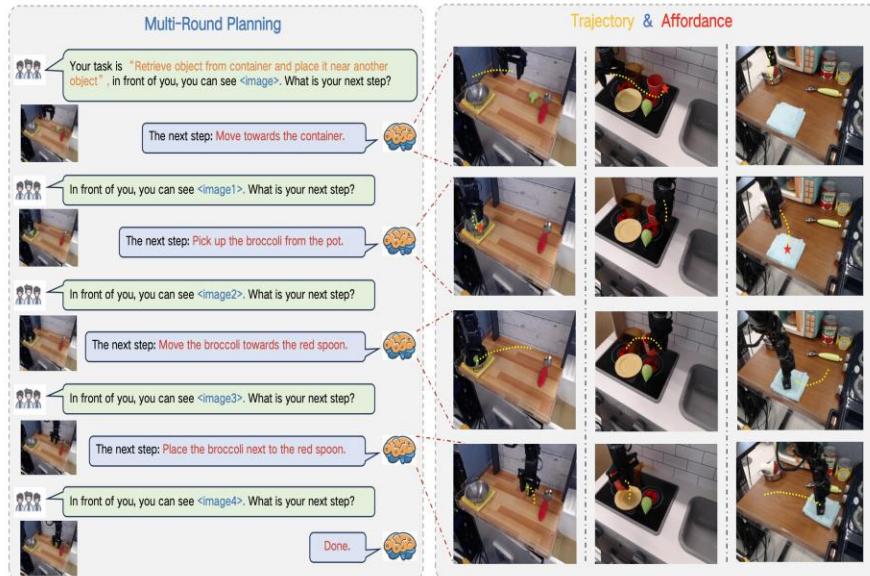


RoboBrain achieves closed-loop execution of long-horizon operational tasks through three core capabilities: task planning, affordance perception, and trajectory prediction.

RoboBrain: An Embodied Multimodal Brain Model for Long-Horizon Operational Tasks



RoboBrain processes visual information from high-resolution images and long video frames, comprehends abstract human language instructions, and outputs affordance regions and future action trajectories, enabling a cognitive leap from abstract instruction understanding to concrete action expression.



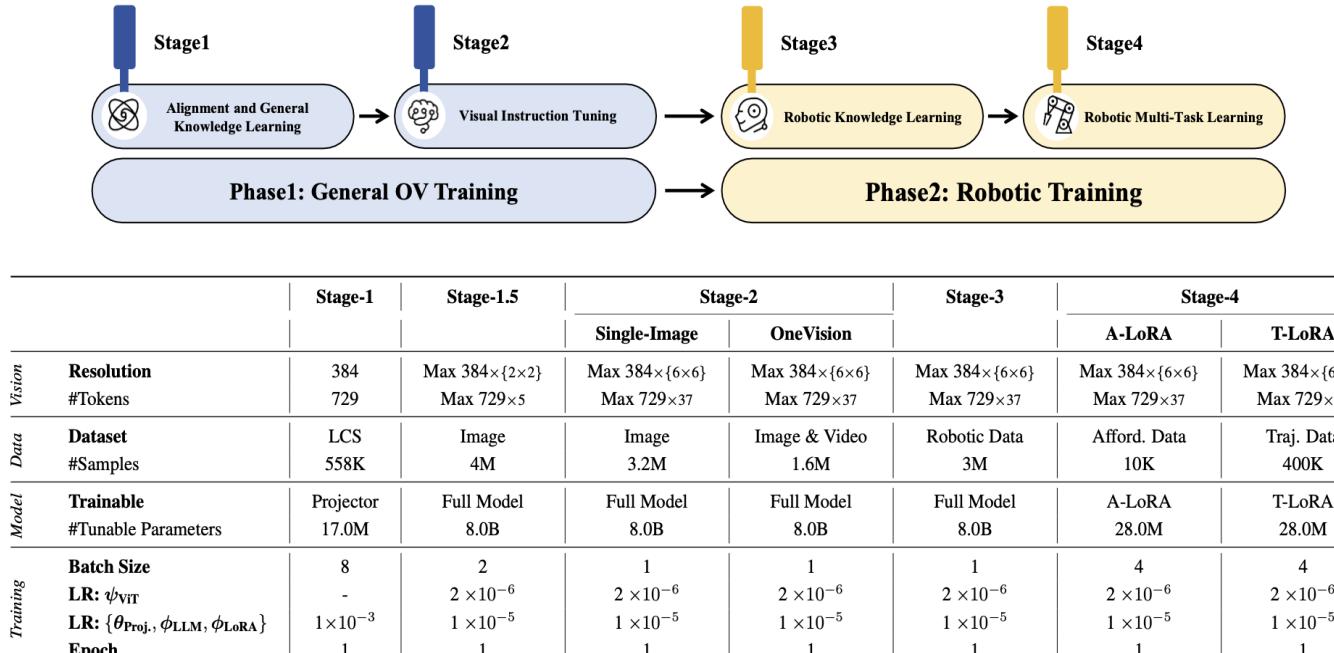
During inference, RoboBrain first processes visual inputs and decomposes instructions into a series of executable subtasks. Subsequently, it performs affordance perception and trajectory prediction.

Model Architecture Diagram: RoboBrain can receive visual information such as high-resolution images and long video frames, comprehend human language instructions, and output subtask planning, affordance regions, and operational trajectories.

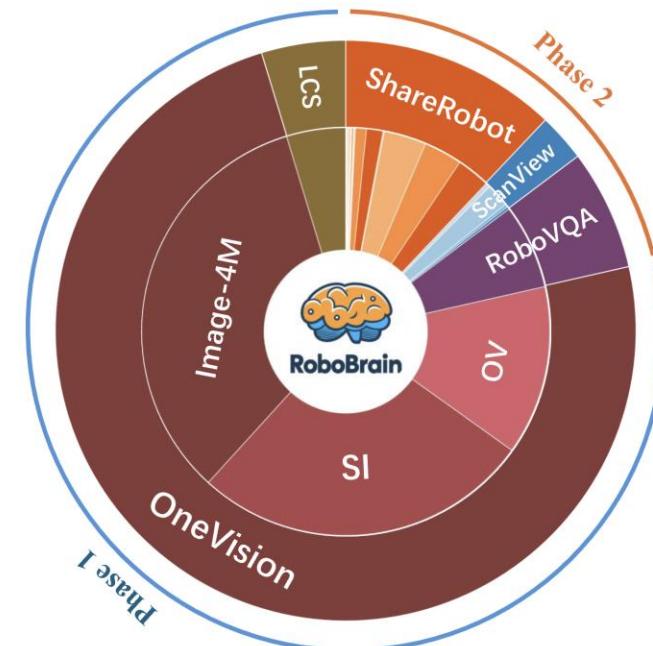
Training Strategy and Training Data



Developing an embodied MLLM with long-horizon task manipulation capabilities requires carefully designed training strategies and curated training data to enhance the model's core competencies while effectively mitigating catastrophic forgetting.



Details of the Four-Stage Training Process

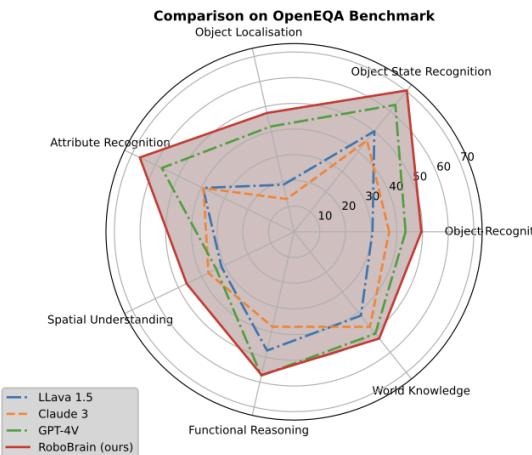


Multi-Stage Training Data Scale and Hybrid Strategy

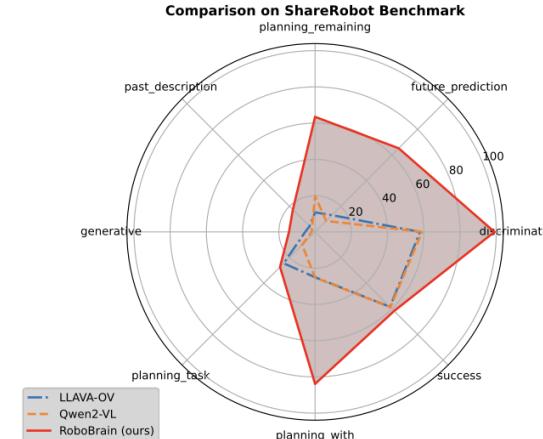
Experimental Results



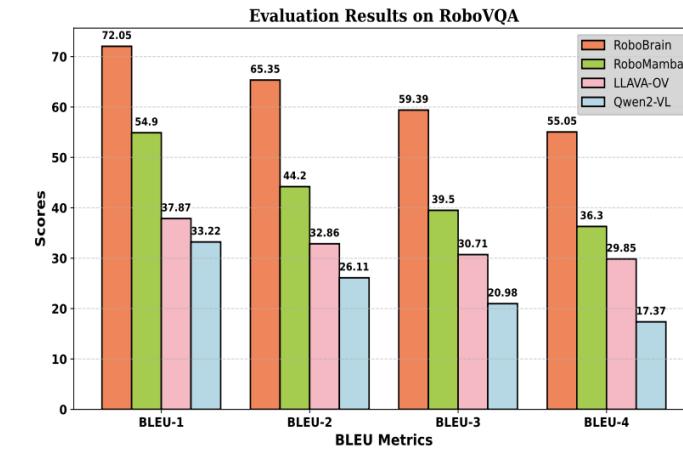
After four training stages and distributed large-scale training with nearly ten million data samples, RoboBrain demonstrated outstanding performance in multiple evaluation tasks, including task planning, operable region perception, and trajectory prediction, fully validating its robust capabilities in robotic operation planning.



(a) OpenEQA Benchmark



(b) ShareRobot Benchmark



(c) RoboVQA Benchmark

RoboBrain outperformed all baseline models across three embodied task planning benchmarks, demonstrating its superior capability in complex long-horizon task planning.

Experimental Results



RoboBrain outperformed all baseline models across three embodied task planning benchmarks, demonstrating its superior capability in complex long-horizon task planning.

Model	AP ↑
LLaVA-NeXT-7B [38]	9.8 %
Qwen2-VL-7B [5]	12.5 %
RoboBrain (Ours)	27.1 % (14.6↑)

RoboBrain's Performance in Operable Region Perception

Method	DFD ↓	HD ↓	RMSE ↓
RoboBrain (Base)	0.191	0.171	0.133
+ Start_Points	0.176	0.157	0.117
+ Max_Points	0.185	0.163	0.125
+ Spec-Token	0.109 (42.9%↓)	0.010 (94.2%↓)	0.091 (31.6%↓)

RoboBrain's Performance in Trajectory Prediction Tasks

Dataset	Split	RoboBrain (Ours)	GPT-4V [1]	LLaVA-OV-7B [25]	InternVL2-8B [9]	Qwen2-VL-7B [54]	GPT-4o [40]
A12D[18]	test	82.03	78.2	81.4	83.8	-	94.2
ChartQA[38]	test	80.48	78.5	80	83.3	83	85.7
DocVQA[39]	test	88	88.4	87.5	91.6	94.5	92.8
TextVQA[44]	val	75.85	-	71.07	77.4	84.3	-
MMMU[58]	val	49	56.8	48.8	51.8	54.1	69.1
MMStar[7]	test	61.23	57.1	61.7	61.5	60.7	63.9
OCRBench[31]	-	677	656	697	794	845	805
RealWorldQA[55]	test	68.89	61.4	66.3	64.4	70.1	58.6
SeedBench[21]	image	71.03	49.9	75.4	76.2	-	76.2
MMbench[32]	en-dev	81.52	81.3	83.2	-	-	83.4
MMbench[32]	en-test	80.44	75	80.8	81.7	83	-
MME[11]	test	2084	1926	1998	2210	2327	-

Performance of RoboBrain in General Multimodal Benchmarks



Visualization - Task Planning



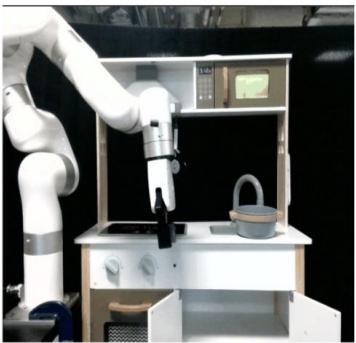
Given the objects in the image, if you are required to complete the task "**Water plants**", what is your detailed plan? Write your plan and explain it in detail, using the following format: Step1: xxx\n Step2: xxx\n ... \n



Step1: Move to the potted plant. Move towards the potted plant on the right side of the cardboard box.
Step2: Pick up the watering can. Pick up the white watering can from the cardboard box. Ensure that you have a firm grip on the handle for better control while watering.
Step3: Water the plants. Water the leaves and soil of the potted plant using the watering can. Start by gently pouring water over the top leaves, ensuring they are moist but not overly wet. Then, pour water into the soil around the base of the plant, making sure it reaches all areas without causing any water to spill out of the pot.
Step4: Place the watering can back. Put the watering can back into the cardboard box. Make sure it is placed securely so it doesn't tip over or cause any damage.
Step5: Move away from the plants. Step away from the plants to avoid any accidental damage while moving around. Ensure that there is no excess water pooling around the base of the plants, as this can lead to root rot.



(a) Embodied planning for Task [Water plants].



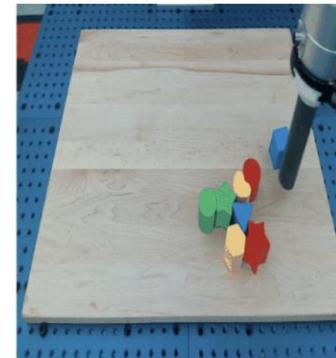
Given the objects in the image, if you are required to complete the task "**Put the pot in the drawer**", what is your detailed plan? Write your plan and explain it in detail, using the following format: Step1: xxx\n Step2: xxx\n ... \n



Step1: Move towards the pot on the counter. Move the robotic arm towards the pot that is placed on the counter next to the sink. The arm should move in a straight line towards the pot.
Step2: Pick up the pot. Grasp the pot with the gripper of the robotic arm and lift it from the counter. Make sure to hold the pot securely to avoid dropping it.
Step3: Move the pot towards the drawer. Move the pot towards the open drawer below the counter. The arm should move in a smooth motion, keeping the pot close to the body for stability.
Step4: Place the pot in the drawer. Place the pot inside the drawer and adjust its position if necessary. Ensure that the pot is fully inside the drawer and not touching the edges or the lid.
Step5: Close the drawer. This can be done by moving the arm away from the drawer and then pushing the drawer closed manually or using a mechanism built into the robotic arm.
Step6: Move away from the drawer. Move the robotic arm away from the drawer to complete the task. The arm should move back to its original position or to a designated safe area.



(b) Embodied planning for Task [Put the pot in the drawer].



Given the objects in the image, if you are required to complete the task "**Cluster blocks of the same color into different corners**", what is your detailed plan? Write your plan and explain it in detail, using the following format: Step1: xxx\n Step2: xxx\n ... \n



Step1: Identify the objects of the same color. - Green cubes: 2 - Blue cubes: 2 - Red cubes: 2 - Yellow cubes: 2
Step2: Group the objects by color. - Green cubes: [green cube, green cube] - Blue cubes: [blue cube, blue cube]
- Red shapes: [red shape, red shape] - Yellow shapes: [yellow shape, yellow shape]
Step3: Move the objects to form clusters.
- Place all green cubes next to each other in the top left corner.
- Place all blue cubes next to each other in the top right corner.
- Place all red shapes next to each other in the bottom left corner.
- Place all yellow shapes next to each other in the bottom right corner.
Step4: Verify the clusters. Check that each cluster contains only objects of the same color.



(c) Embodied planning for Task [Cluster blocks of the same color into different corners].



Given the objects in the image, if you are required to complete the task "**Clean the desk**", what is your detailed plan? Write your plan and explain it in detail, using the following format: Step1: xxx\n Step2: xxx\n ... \n



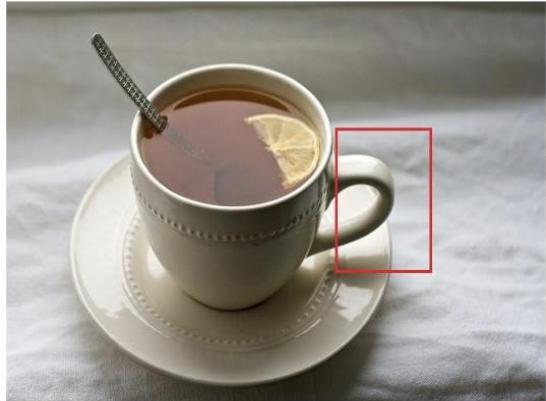
Step1: Move the coffee cup to a safe location. Make sure to keep the coffee cup away from the monitor.
Step2: Wipe down the desk with a **disinfectant wipe**. Start from one end of the desk and move to the other.
Step3: Clean the monitor screen using a **disinfectant wipe**. Be careful not to touch any other parts of the monitor.
Step4: Dust and clean the computer tower using a **disinfectant wipe**. Pay attention to any vents or crevices where dust may have accumulated.
Step5: Organize the cables and wires. Use the robotic arm to hold them in place while you use a cable organizer or ties to secure them.
Step6: Dispose of any trash or empty containers from the desk. Use the robotic arm to pick them up and throw them away.



(d) Embodied planning for Task [Clean the desk].



Visualization - Affordance Perception Capability



drink_with the cup



hold the cup



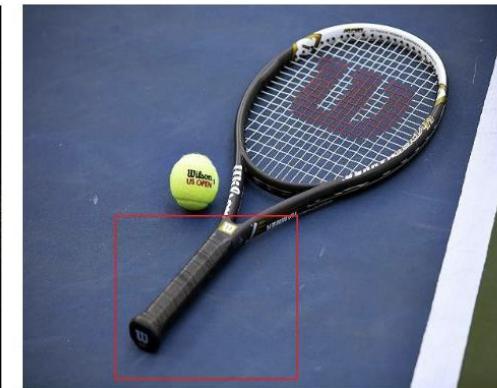
drink_with the bottle



open the microwave

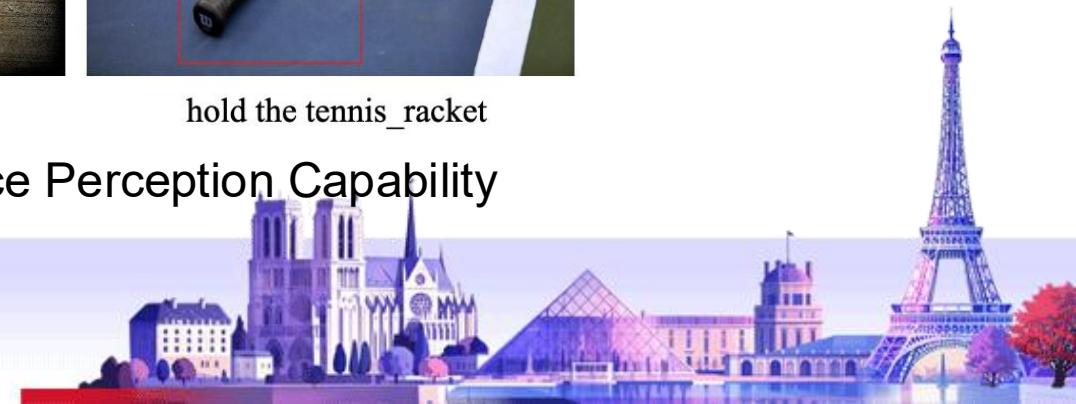


hold the hammer

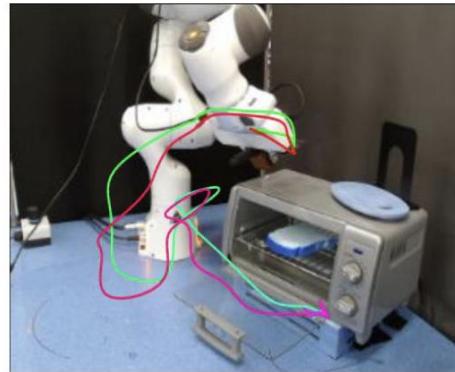


hold the tennis_racket

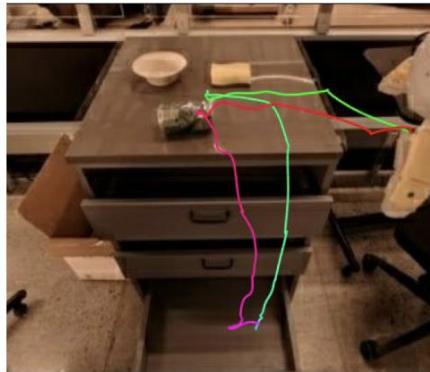
Application Cases of RoboBrain in Affordance Perception Capability



Visualization - Trajectory Prediction



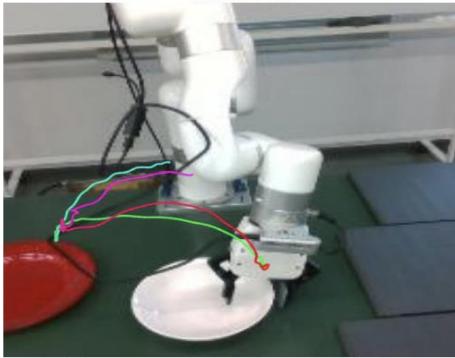
make a piece of toast with
the oven



place green rice chip bag into top
drawer



open bottom drawer



Pick up a white plate, and then
place it on the red plate

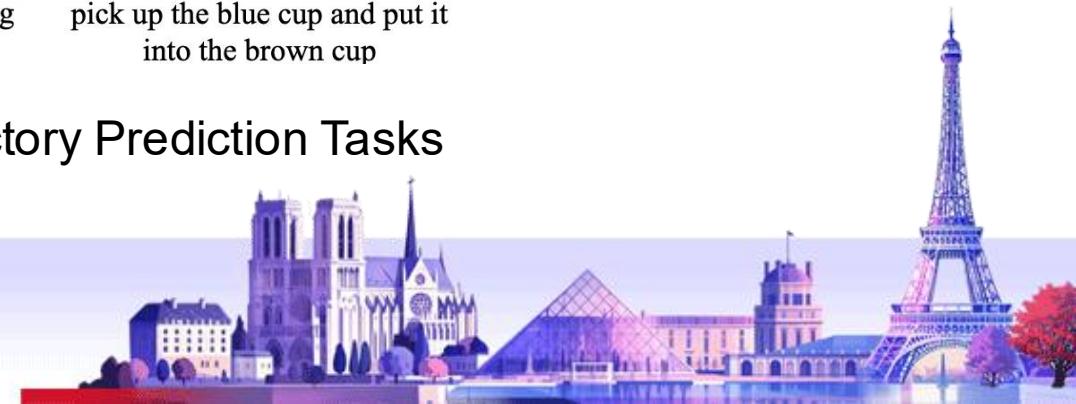


make a cup of coffee with keurig
machine



pick up the blue cup and put it
into the brown cup

Application Cases of RoboBrain in Trajectory Prediction Tasks



Four Ways Computer Vision Is Driving AI-Enhanced Robotics

Artificial intelligence (AI)-enhanced robotics has emerged as a primary area of growth in the field of computer vision. As the [IEEE CS 2025 Technology Predictions report](#) explains: "Embodied intelligence will enable robots to perceive, learn, and collaborate in dynamic environments, achieving unprecedented autonomy and human-like adaptability." Market research supports that sentiment: According to [Statista](#), the AI robotics market size is expected to show an annual growth rate (CAGR 2025-2030) of 23.37%, resulting in a market volume of US\$64.35bn by 2030.

Developments with large language models (LLMs), multimodal AI, and broader computer vision efforts have been among the primary forces behind this rapid acceleration, and as the industry gears up for its leading AI engineering event, the [Computer Vision and Pattern Recognition Conference \(CVPR\)](#), paper submissions reiterate the role that AI-enhanced robotics will play.

"A lot of people in computer vision are now interested in robotics," said Phillip Isola, CVPR 2025 Program Co-Chair and an associate professor at the Massachusetts Institute of Technology (MIT) in Boston, Mass., U.S. "They're first starting by modeling 3D scenes, and that will be more relevant to robotics. For example, they might be working on navigation of a house, and there's no robot. But that's where they're going."

CVPR 2025 Program Co-Chair Fuxin Li, an associate professor at Oregon State University in Corvallis, Ore., U.S., agreed, adding that the program has evolved accordingly. "It's an emerging trend. We see more and more convergence between the two areas. There are also a lot more real robots in CVPR papers than a few years ago."

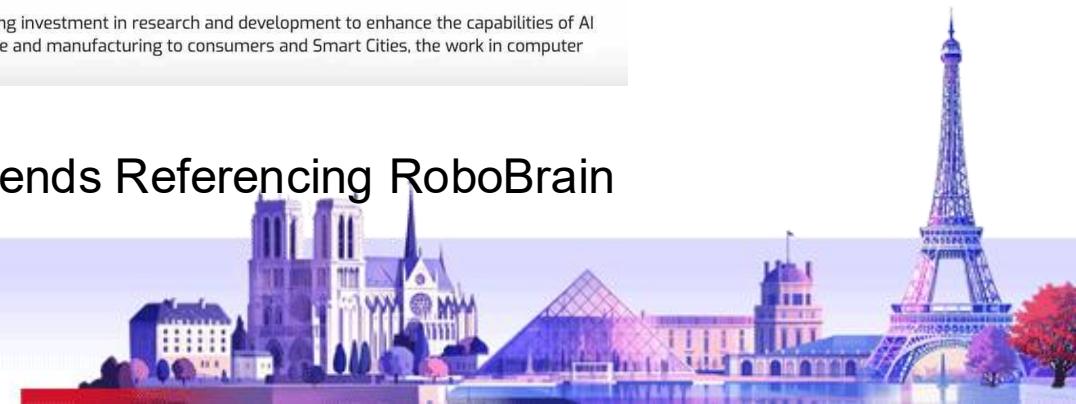
Current research focus areas

So, how exactly is this emerging trend shaping today's computer vision and pattern recognition landscape? From real-world, 3D data guiding robots in their tasks to datasets for navigating spatial relations to other datasets designed to maximize a robot's functional abilities to pick up, move, or shift elements in an environment, much of the work in the intersection of computer vision and robotics focuses on ways to train robot understanding and ensure they function efficiently and successfully in a variety of tasks. Specifically, current research focus includes:

1. **Expanding robotic intelligence.** To harness the full potential of robotic automation in a wide variety of landscapes, robots need to be able to address not just menial tasks but elevate their "thinking" to a more challenging cognitive level. In the CVPR paper "[RoboBrain: A Unified Brain Model for Robotic Manipulation from Abstract to Concrete](#)," researchers describe a dataset that labels multi-dimensional information such as task planning, object affordance, and end-effector trajectory.
2. **Addressing robotic dexterity and movement.** While higher-level tasks require superior thinking, more physical tasks require advanced precision. Thus, researchers are aiming efforts on angle manipulation, grip, and more. For example, "[3D-MVP: 3D Multiview Pretraining for Robotic Manipulation](#)," a paper to be presented at CVPR, will demonstrate a multi-view transformer to understand the 3D scene and predict gripper pose actions. Another paper, "[Let Humanoid Robots Go Hiking! Integrative Skill Development over Complex Trails](#)," seeks to train robots as hikers on complex trails via a universal learning framework that strives to address both visual perceptual awareness and body dynamics.
3. **Applying robotic automation to industrial and business environments.** Research efforts focus not only on the developmental side of robotics and AI but also on addressing their more rapid deployment for routine tasks. The CVPR [Workshop on Perception for Industrial Robotics Automation](#) will focus on a "bin picking" challenge, where robots are tasked with grabbing a randomized selection of objects from a bin at different angles/clasps. The competition awards up to US\$60,000 in prizes. In addition, CVPR exhibitor [Fourier](#) will be demonstrating its GR-1 robots, which according to its website, are being developed to be the next generation of bank reception manager, factory support, and rehabilitation diagnosis, training, and treatment.
4. **Identifying robotic opportunities in consumer functions.** Consumer applications offer a large area of potential, bringing forth visions of the housekeeper, Rosie, from the cartoon *The Jetsons* and far beyond. For instance, the CVPR paper, "[VidBot: Learning Generalizable 3D Actions from In-the-Wild 2D Human Videos for Zero-Shot Robotic Manipulation](#)," uses in-the-wild human videos to train robots on tasks such as closing the fridge, opening the cupboard, wiping the counter, and much more. Building on these concepts, exhibitor [Booster Robotics](#) has developed humanoid robots that play soccer—and offers a framework for training learning-based locomotion from the ground up.

While much work remains in the field of AI-enhanced robotics, just as much opportunity awaits. According to [analysts](#), "The growing investment in research and development to enhance the capabilities of AI robotics is expected to propel market growth in the coming years." From the implications in the industries ranging from healthcare and manufacturing to consumers and Smart Cities, the work in computer vision and pattern recognition today will have lasting influence on the future of robotics and its positive impact on humankind.

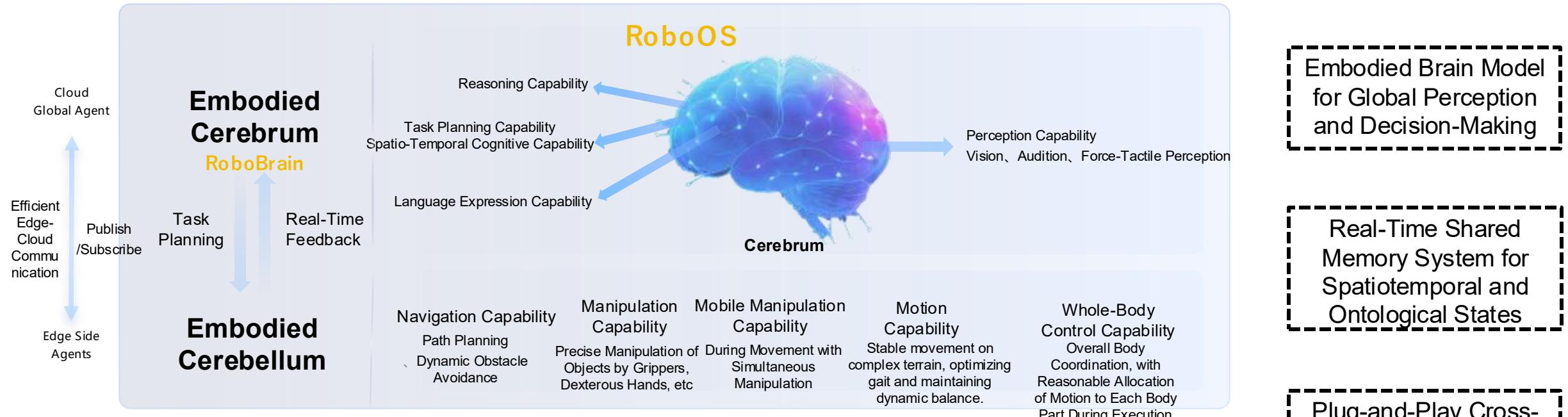
Official CVPR Review on Embodied Intelligence Trends Referencing RoboBrain



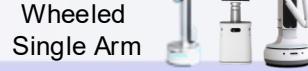
RoboOS: **A Hierarchical Collaborative Framework** **for RoboBrain and Robot Agents**



Cross-Ontology Embodied Cerebrum-Cerebellum Collaboration Framework



Ontology



GOSIM AI Paris 2025

Cross-Ontology Embodied Cerebrum-Cerebellum Collaboration Framework



Cloud-Based Brain Model RoboBrain (Slow System): Global Perception and Decision-Making

✓ Spatiotemporal Perception:

Fuses 3D scene reconstruction and state tracking for **dynamic spatiotemporal cognition**.

✓ Planning Guidance:

Enables **swarm task scheduling** and **fine-grained action guidance** (affordance/trajectory).

✓ Feedback and Error Correction:

Monitors execution (e.g., force anomalies) and performs **three-level dynamic replanning**.



Spatiotemporal Perception

Planning Guidance

Feedback and Error Correction

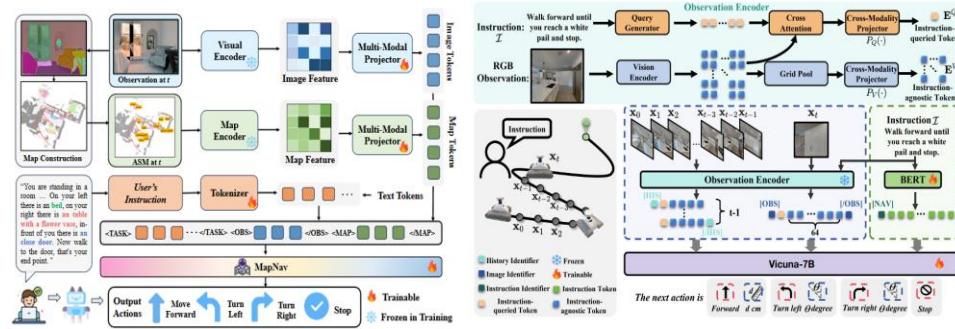
Cross-Ontology Embodied Cerebrum-Cerebellum Collaboration Framework



Cerebellum Skill Library (Fast System): Low-Latency Precise Execution

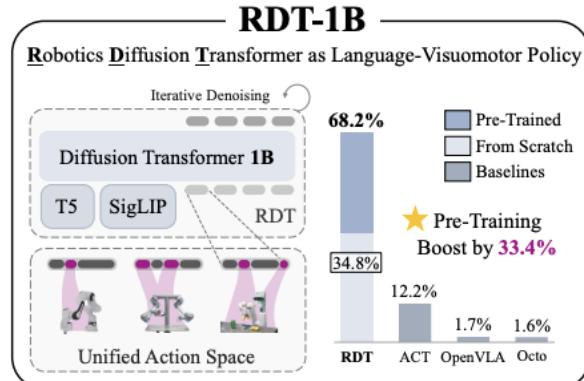
✓ Plug-and-Play Skills:

- Flexible Integration of Manipulation Skills (e.g., AnyGrasp, VLA)
- Various Traditional Control/Visual Navigation Skills (e.g., SLAM, VLN),
- Compatibility with Specialized Tools for Different Embodiments (e.g., Flexible Manipulation, Precise Manipulation)

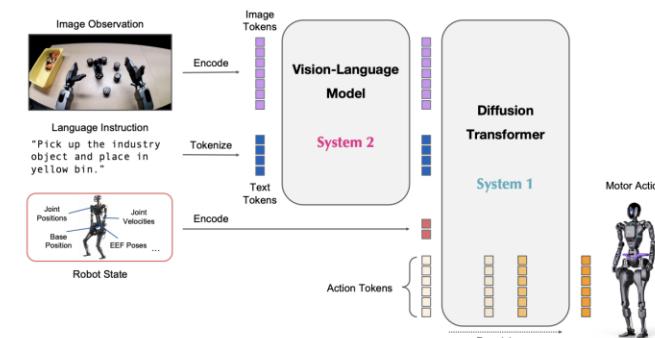


Navigation Skills

✓ Plug-and-Play Embodiment: Single-Arm, Dual-Arm, Wheeled, Humanoid



Manipulation Skills

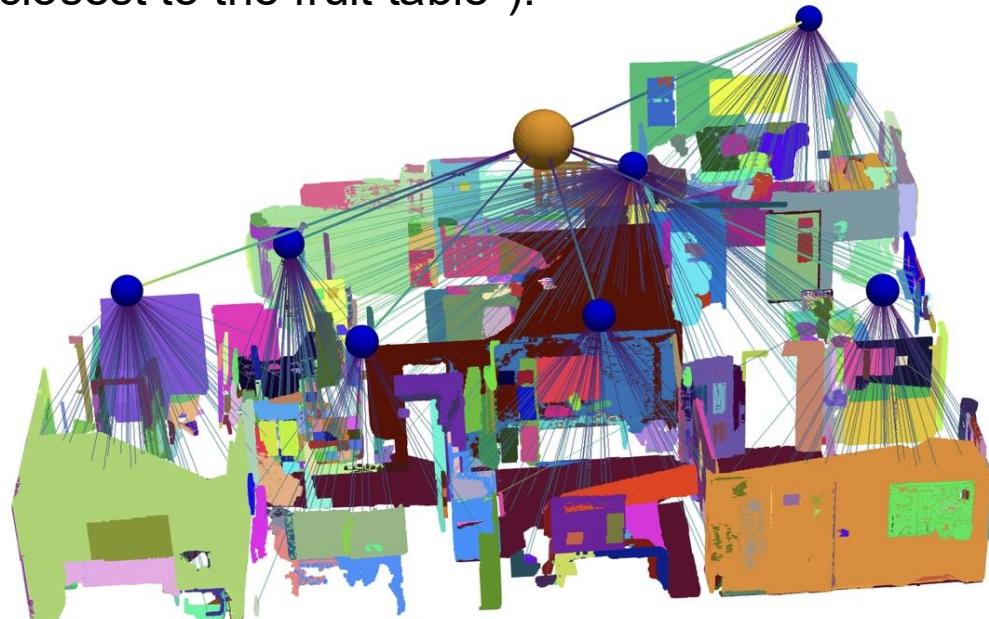


Cross-Ontology Embodied Cerebrum-Cerebellum Collaboration Framework



Shared Memory System: Cross-Ontology Multi-Robot Data Hub

- ✓ **Spatial Memory:** Real-time updates of object spatial relationships (e.g., "fruit knife located at table corner") and robot embodiment positions in the environment (e.g., "Unitree humanoid robot G1 is closest to the fruit table").



scene-graph

```
scene:  
  - recep_name: "kitchenTable"  
    recep_type: "table"  
    recep_object:  
      - 'apple'  
      - 'pear'  
      - 'banana'  
      - 'knife'  
  
  - recep_name: "customTable"  
    recep_type: "table"  
    recep_object:  
      - 'basket'  
      - 'plate'  
      - 'cup'  
  
  - recep_name: "servingTable"  
    recep_type: "table"  
    recep_object:  
      - 'bowl'  
      - 'fork'  
      - 'spoon'  
  
  - recep_name: "basket"  
    recep_type: "container"  
    recep_object:  
      - 'egg'
```

Format Example

Roboos: Customized Fine-Tuning Training Set

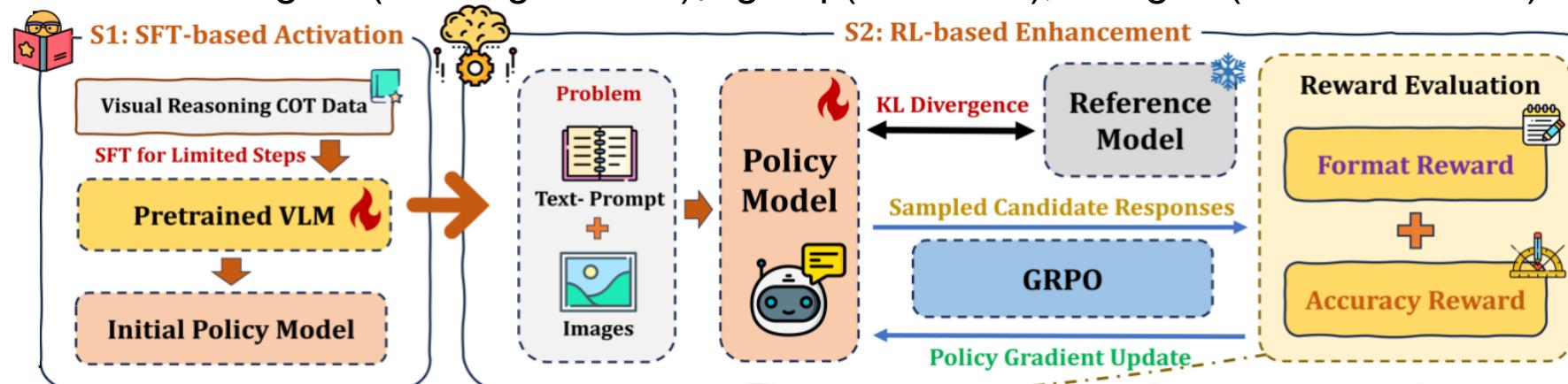
RoboBrain/VLM Fine-Tuning Training:

✓ Task Decomposition to SubTask (CoT-SFT)

- Task: I want to eat fruit
- Reasoning: Analyze fruit, location, and robot state based on Memory
- Predict: Robot1 retrieves fruit basket to kitchen, Robot2 places pear and fruit knife, Robot1 returns with basket

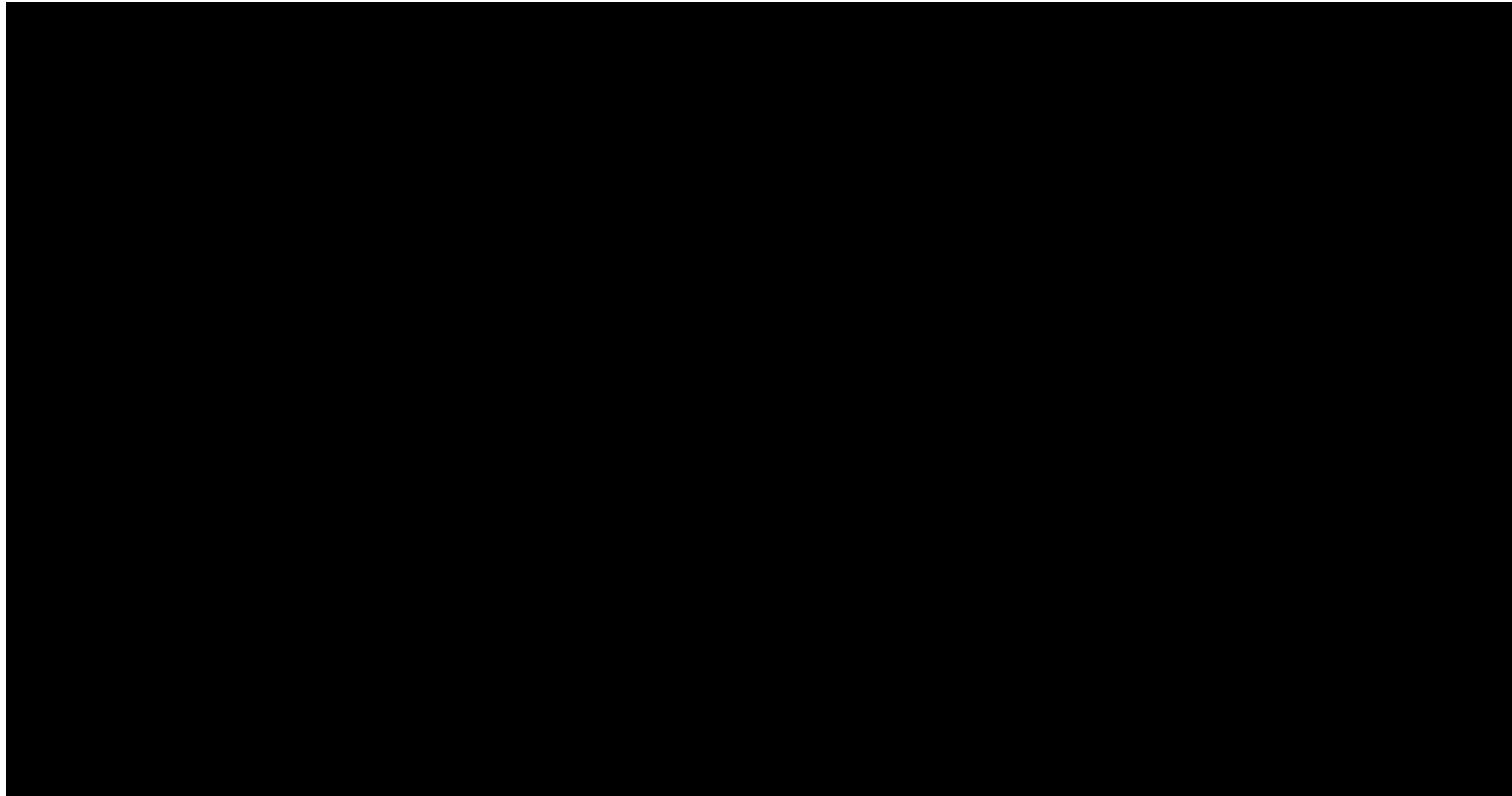
✓ SubTask --> Tool (R1)

- Task: Retrieve fruit basket to kitchen
- Predict: navigate('servingTable01), grasp(basket01), navigate(kitchenTable01)

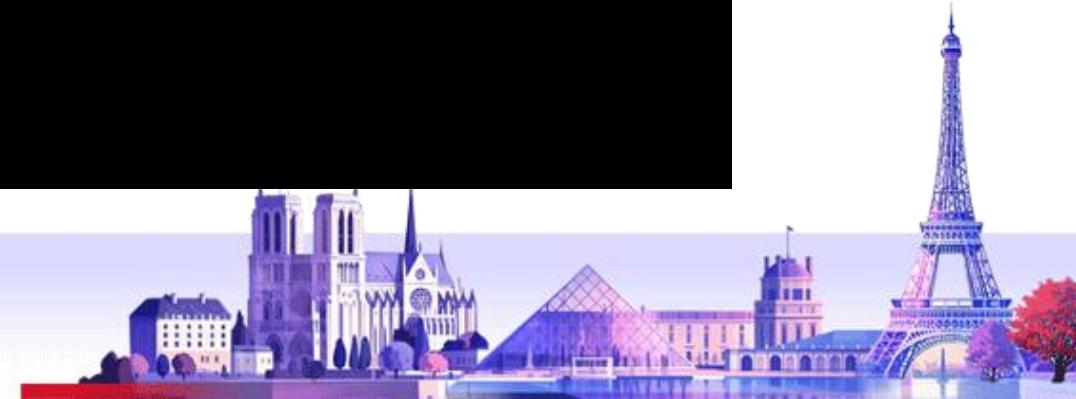


Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning, <https://arxiv.org/abs/2503.20752>

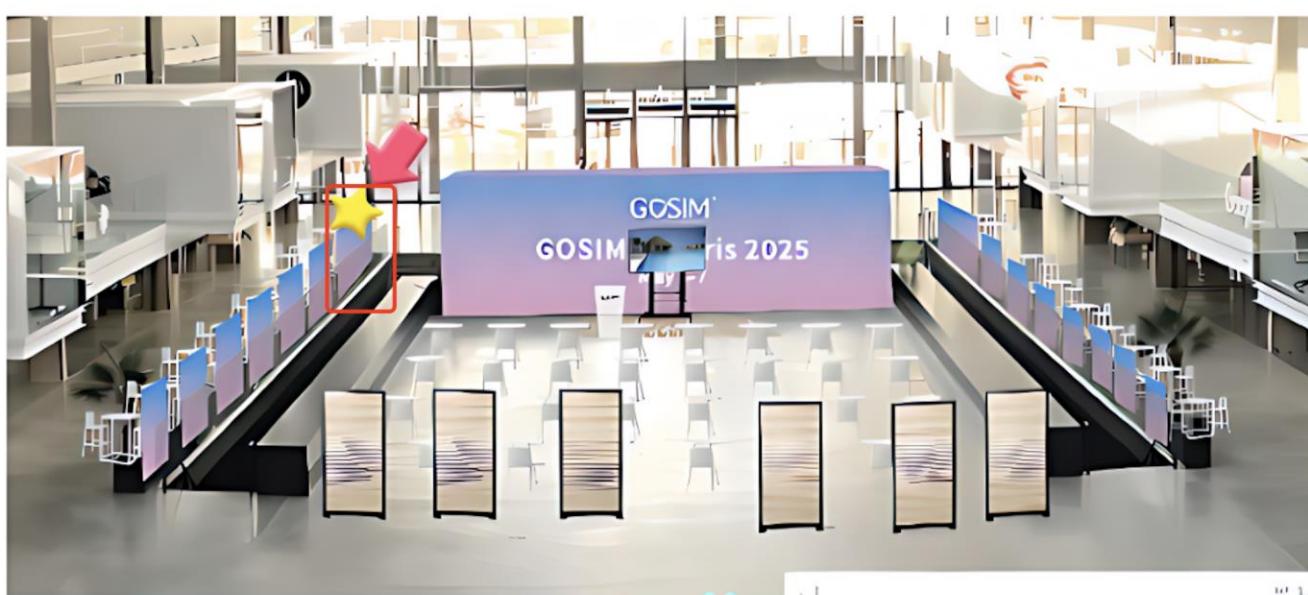
Demo Scenario: "I Want to Eat Some Fruit"



GOSIM AI Paris 2025

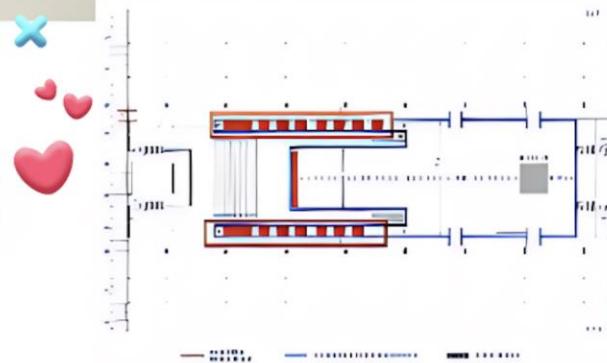


Welcome to our Booth!



OPEN PLATFORM

Welcome to our Booth!
Have A Nice Talk~



GOSIM AI Paris 2025



1st Floor, Open Platform Area

Entering the main gate, the first booth on the right side (next to the GOSIM main display board)

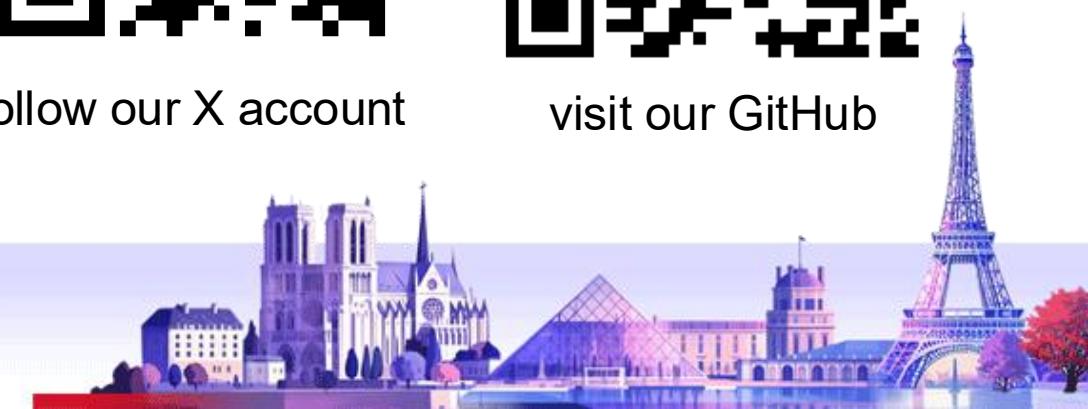
Visit us at **BAAI Booth** (with a shining Star in the left picture)



follow our X account



visit our GitHub



THANK YOU

