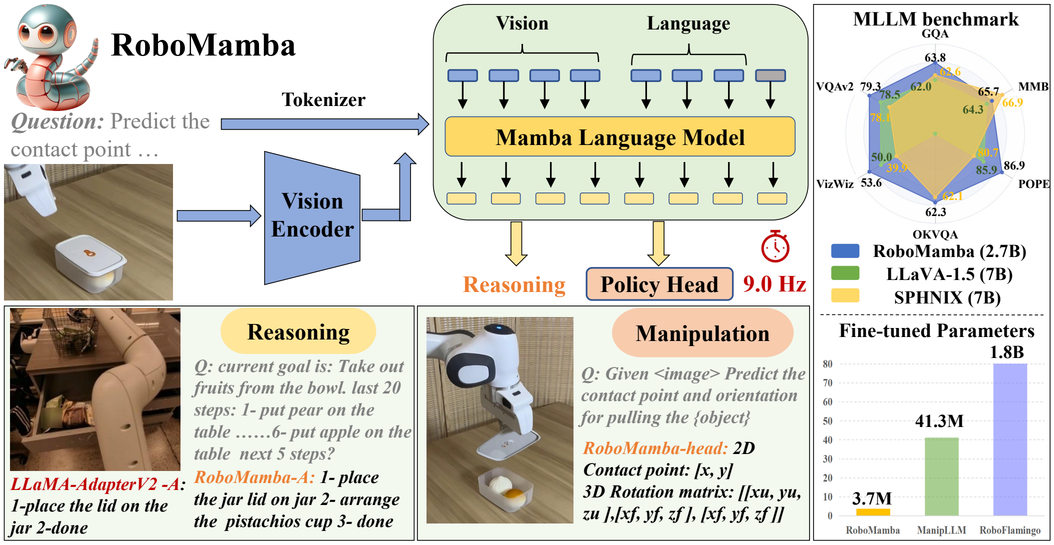


Q: current goal is: Take out fruits from the bowl. last 20 steps: 1- put pear on the table ……6- put apple on the table next 5 steps?

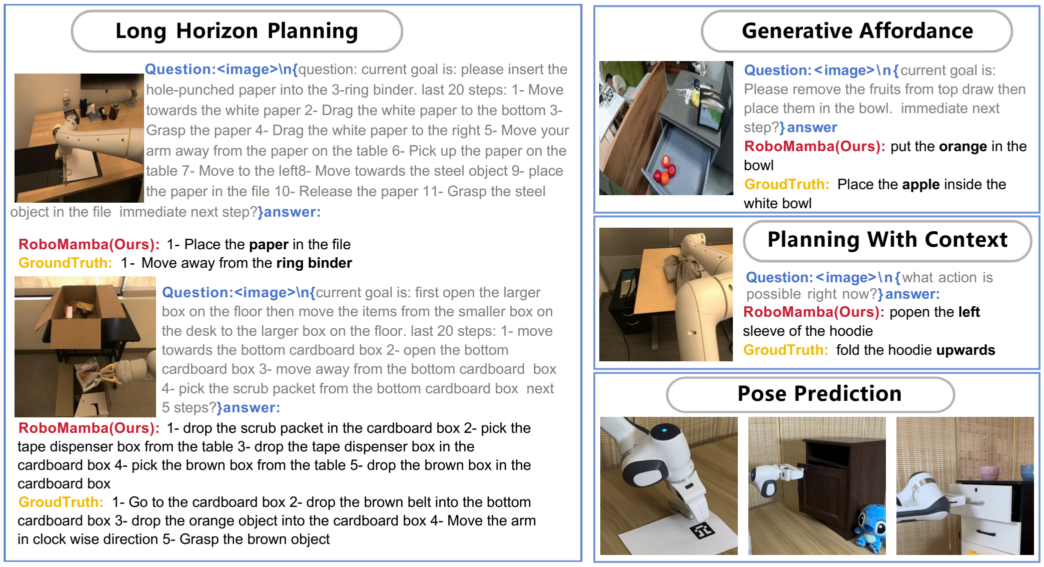
RoboMamba-A: 1- place the jar lid on jar 2- arrange the pistachios cup 3- done

LLaMA-AdapterV2 -A: 1-place the lid on the jar 2-done

Q: Given <image> Predict the contact point and orientation for pulling the {object}

RoboMamba-head: 2D Contact point: [x, y]
3D Rotation matrix: [[xu, yu, zu],[xf, yf, zf], [xf, yf, zf]]

Figure 1: **To Reviewer vV5n.** Detailed modifications of Figure 1.



Question:<image>\n{question: current goal is: please insert the hole-punched paper into the 3-ring binder. last 20 steps: 1- Move towards the white paper 2- Drag the white paper to the bottom 3- Grasp the paper 4- Drag the white paper to the right 5- Move your arm away from the paper on the table 6- Pick up the paper on the table 7- Move to the left8- Move towards the steel object 9- place the paper in the file 10- Release the paper 11- Grasp the steel object in the file immediate next step?}answer:

RoboMamba(Ours): 1- Place the **paper** in the file
GroundTruth: 1- Move away from the **ring binder**

Question:<image>\n{current goal is: first open the larger box on the floor then move the items from the smaller box on the desk to the larger box on the floor. last 20 steps: 1- move towards the bottom cardboard box 2- open the bottom cardboard box 3- move away from the bottom cardboard box 4- pick the scrub packet from the bottom cardboard box next 5 steps?}answer:

RoboMamba(Ours): 1- drop the scrub packet in the cardboard box 2- pick the tape dispenser box from the table 3- drop the tape dispenser box in the cardboard box 4- pick the brown box from the table 5- drop the brown box in the cardboard box
GroudTruth: 1- Go to the cardboard box 2- drop the brown belt into the bottom cardboard box 3- drop the orange object into the cardboard box 4- Move the arm in clock wise direction 5- Grasp the brown object

Question:<image>\n{current goal is: Please remove the fruits from top draw then place them in the bowl. immediate next step?}answer

RoboMamba(Ours): put the **orange** in the bowl
GroudTruth: Place the **apple** inside the white bowl

Question:<image>\n{what action is possible right now?}answer:

RoboMamba(Ours): popen the **left** sleeve of the hoodie
GroudTruth: fold the hoodie **upwards**

Figure 2: **To Reviewer 16K4.** The visualization of reasoning failure cases. In the bottom right corner of the image, we re-select the qualitative results from our real-world demonstration. Additionally, we replace the red dot and virtual end-effector with a physical Franka Panda Robot.

Table 1: **To Reviewer 16K4.** The Scalability Exploration of RoboMamba. For parameter scalability, since the official 7B model is not public, we compare the reasoning abilities of Mamba 1.4B and 2.7B on the same training dataset in the first two rows. The results indicate that increasing the model parameters effectively enhances reasoning ability under efficient LLM settings. For training data scalability, we introduce more training data during the co-training stage. Although there is a slight improvement on some MLLM benchmarks, the average reasoning did not improve.

Parameters	ShareGPT4V-SFT 665K	LVIS-Instruct-4V 220K	OKVQA	GQA	POPE
1.4B			28.5	40.8	66.8
2.7B			62.3	63.8	86.9
2.7B	+		56.8	57.5	85.9
2.7B		+	62.4	64.4	86.0