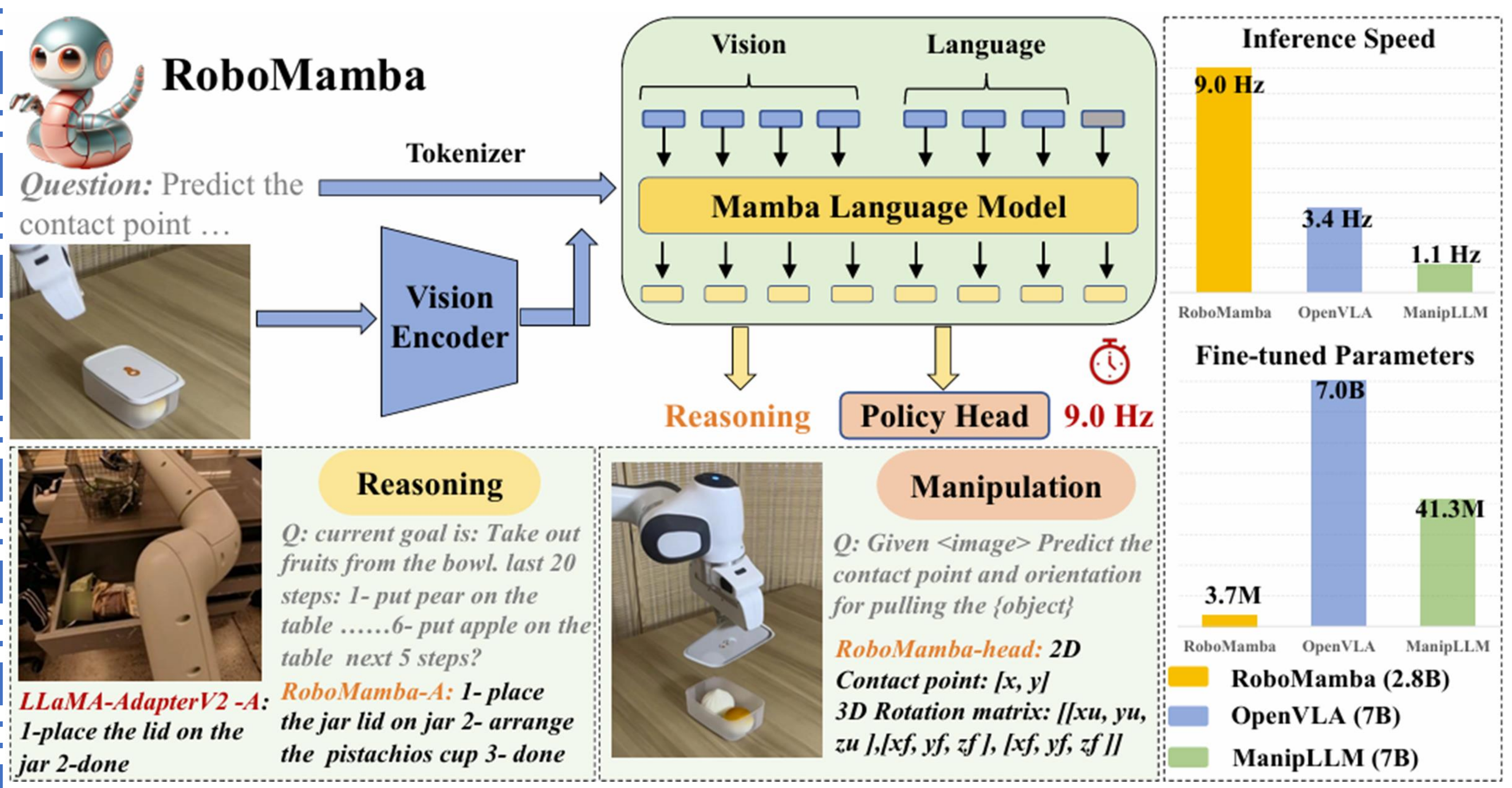
RoboMamba: Efficient Vision-Language-Action Model for Robotic Reasoning and Manipulation

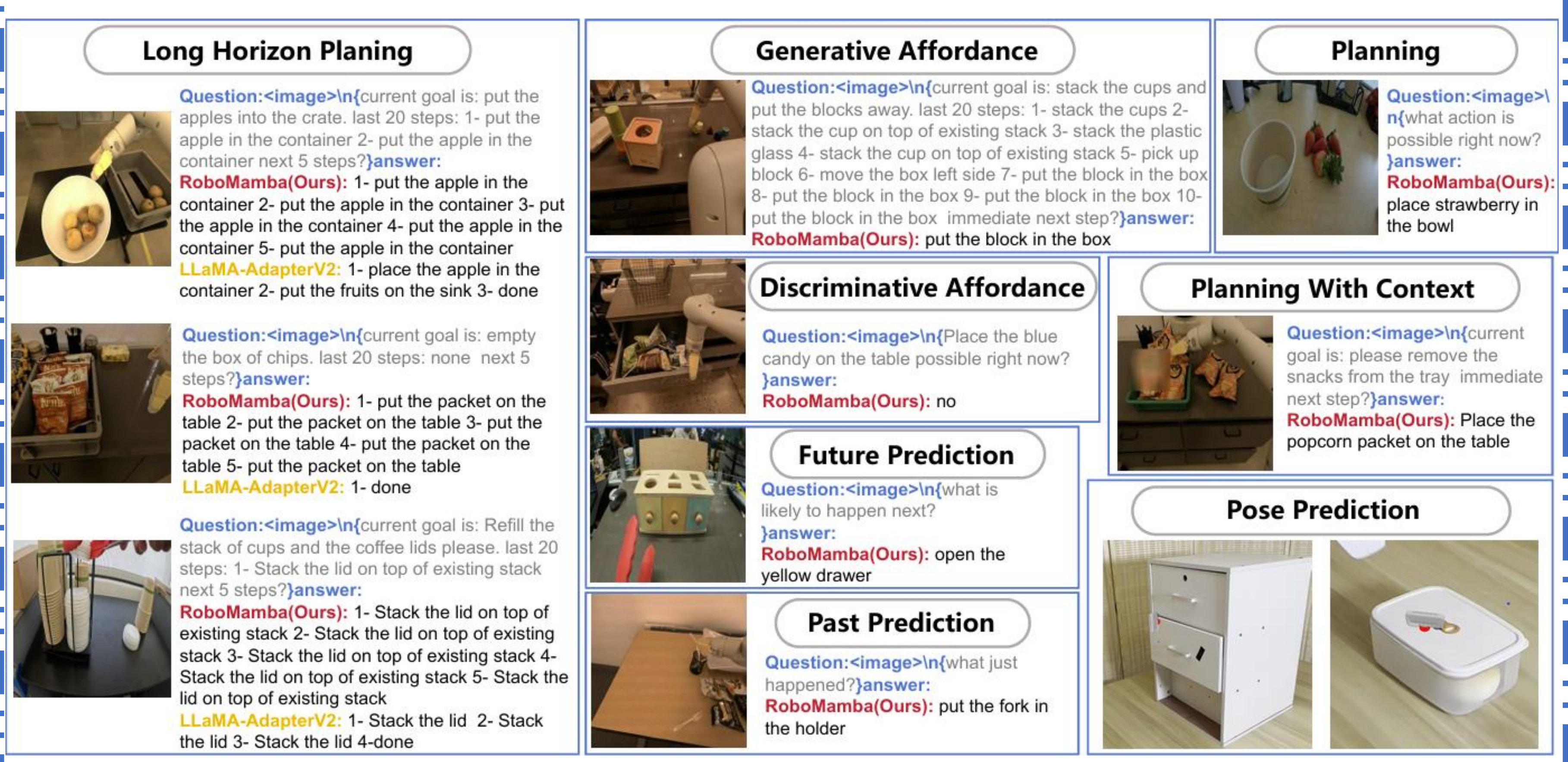
Jiaming Liu[1 *,†], Mengzhen Liu[1*], Zhenyu Wang[1], Pengju An[1], Xiaoqi Li[1], Kaichen Zhou[1], Senqiao Yang[1], Renrui Zhang[†], Yandong Guo[2], Shanghang Zhang[1,3 ✉]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University; [2]AI2Robotics; [3]Beijing Academy of Artificial Intelligence (BAAI)

1. Robomamba



RoboMamba is an efficient robotic VLA model that combines reasoning and manipulation capabilities. First, we integrate and align a vision encoder with the Mamba LLM, endowing our model with common sense and robotic-related reasoning abilities. Subsequently, we introduce an efficient fine-tuning strategy to equip RoboMamba with pose prediction abilities, requiring a few dozen minutes to fine-tune a simple policy head (3.7M parameters). In terms of inference speed, RoboMamba achieves the highest control frequency, surpassing other VLA models, running on an NVIDIA A100 GPU without any quantization or inference acceleration techniques. More real-world downstream tasks are displayed below.



2. Previous Work Limitations

First, the reasoning capabilities of pre-trained MLLMs in robotic scenarios are found to be insufficient. As shown in Figure 1 (reasoning example), this deficiency presents challenges for fine-tuned robot MLLMs when they encounter complex reasoning tasks.

Second, fine-tuning MLLMs and using them to generate robot manipulation actions incurs higher computational costs due to their expensive attention-based LLMs.

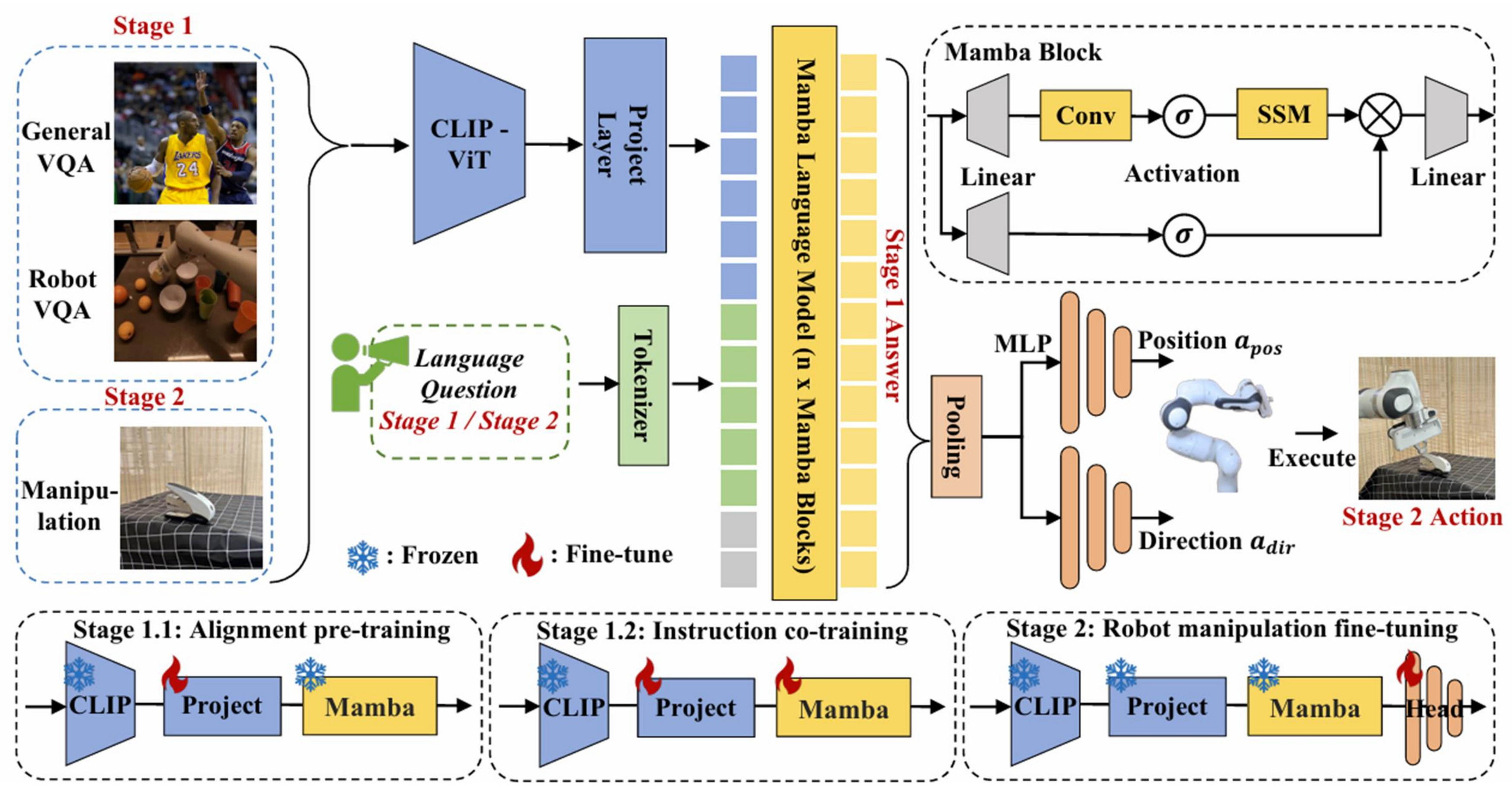
3. Main Contributions

1. We introduce RoboMamba, an efficient VLA model that integrates a vision encoder with the linear-complexity Mamba LLM, which possesses visual common sense and robotic-related reasoning abilities.
2. To equip RoboMamba with action pose prediction abilities, we explore an efficient fine-tuning strategy using a simple policy head. We find that once RoboMamba achieves sufficient reasoning capabilities, it can acquire pose prediction skills with minimal cost.
3. In our extensive experiments, RoboMamba excels in reasoning on general and robotic evaluation benchmarks, and showcases impressive pose prediction results in both simulation and real-world experiments.

4. Method

Overall framework of RoboMamba:

RoboMamba projects images onto Mamba's language embedding using a vision encoder and projection layer, which is then concatenated with text tokens and fed into the Mamba model. To predict the position and rotation of the end-effector pose, we inject simple MLP policy heads and use the global token as input, which is generated through a pooling operation from the language output tokens. Training strategy of RoboMamba. For model training, we divide our training pipeline into two stages. In Stage 1, we introduce alignment pre-training (Stage 1.1) and instruction co-training (Stage 1.2) to equip RoboMamba with both common sense and robotic-related reasoning abilities. In Stage 2, we propose robotic manipulation fine-tuning to efficiently empower RoboMamba with low-level manipulation skills.



Stage 1.1 Alignment pre-training: We freeze the parameters of the vision encoder and Mamba language model, and only update the project layer.

Stage 1.2 Instruction co-training : We freeze the parameters of the CLIP encoder and fine-tune the projection layer and Mamba on the combined instruction datasets.

Stage 2 Robot manipulation fine-tuning: We freeze all the parameters of RoboMamba and introduce a simple policy head to model Mamba's output tokens. The policy head contains two MLPs that separately learn the end-effector's position and direction

5. Experiment

Table 1: Comparison of general reasoning abilities with previous MLLMs across several benchmarks. 'Res.' indicates the resolution of the input image. RoboVQA1 to RoboVQA4 represent the BLEU-1 to BLEU-4 scores, respectively. For TinyLLaVA and LLaMA-AdapterV2, we evaluate robotic reasoning abilities after fine-tuning the pre-trained MLLMs on the RoboVQA dataset.

Method	LLM	Res.	OKVQA	VQAV2	GQA	VizWiz	POPE	MME	MMB	MM-Vet	RoboVQA$_4$	RoboVQA$_1$
BLIP-2 [43]	7B	224	45.9	-	41.0	19.6	85.3	1293.8	-	22.4	-	-
InstructBLIP [79]	7B	224	-	-	49.5	33.4	-	-	36	26.2	-	-
LLaMA-AdapterV2 [45]	7B	336	49.6	70.7	45.1	39.8	-	1328.4	-	-	8.1	27.8
MiniGPT-v2 [80]	7B	448	57.8	-	60.1	53.6	-	-	-	-	-	-
Qwen-VL [81]	7B	448	58.6	79.5	59.3	35.2	-	-	38.2	-	-	-
LLaVA1.5 [67]	7B	336	-	78.5	62.0	50.0	85.9	**1510.7**	64.3	30.5	-	-
SPHINX [64]	7B	224	62.1	78.1	62.6	39.9	80.7	1476.1	66.9	**36.0**	-	-
LLaVA-Phi [49]	2.7B	336	-	71.4	35.9	-	85.0	1335.1	59.8	28.9	-	-
MobileVLM [82]	2.7B	336	-	-	59.0	-	84.9	1288.9	59.6	-	-	-
TinyLLaVA [83]	2.7B	336	-	77.7	61.0	-	86.3	1437.3	**68.3**	31.7	29.6	43.5
RoboMamba(Ours)	2.7B	224	**63.3**	**79.6**	**64.2**	57.1	86.3	1297.2	60.9	29.4	**42.8**	**62.7**
RoboMamba(Ours)	2.7B	336	62.7	77.7	63.3	**58.1**	**87.0**	1335.5	60.7	31.4	41.8	61.9

Reasoning Capability

Table 2: Comparison of the success rates between RoboMamba and baselines across various training (seen) and test (unseen) tasks. The representation for each task icon is shown in Table 3.

Method	Seen Categories															
UMPNet [63]	0.28	0.41	0.25	0.20	0.49	0.20	0.35	0.57	0.51	0.25	0.66	0.17	0.17	0.26	0.27	0.40
FlowBot3D [57]	0.50	0.53	0.26	0.36	0.34	0.36	0.54	0.26	0.12	0.34	0.41	0.23	0.36	0.30	0.17	0.37
RoboFlamingo [14]	0.48	0.51	**0.50**	0.35	0.11	0.47	0.54	0.35	0.19	0.46	0.18	0.64	0.26	0.42	0.15	0.87
ManipLLM [15]	0.68	0.62	0.45	0.74	0.42	0.25	0.61	**0.66**	**0.56**	0.52	0.50	0.42	**0.64**	0.76	**0.63**	0.60
RoboMamba (Ours)	**0.81**	**0.73**	0.33	**0.85**	**0.86**	**0.60**	**0.81**	0.42	**0.56**	**0.54**	**0.68**	**0.81**	0.26	**0.86**	0.39	**0.91**

Method	Seen Categories				AVG	Unseen Categories										AVG
UMPNet [63]	0.27	0.37	0.19	0.60	0.34	0.32	0.36	0.18	0.37	0.21	0.12	0.04	0.53	0.28	0.13	0.26
FlowBot3D [57]	0.21	0.57	0.29	0.45	0.35	**0.36**	0.36	0.18	0.30	0.21	0.50	0.13	0.53	0.28	0.09	0.30
RoboFlamingo [14]	0.20	0.42	**0.58**	0.60	0.41	**0.36**	**0.62**	0.64	0.33	0.14	0.34	0.44	0.66	**0.41**	0.31	0.43
ManipLLM [15]	**0.41**	**0.78**	0.41	0.59	0.56	0.21	0.25	**0.79**	**0.76**	0.52	**0.76**	0.43	**0.85**	0.26	0.52	0.51
RoboMamba(Ours)	0.40	0.55	0.37	**0.80**	**0.63**	0.19	0.23	0.67	0.66	**0.57**	0.45	**0.65**	0.68	0.30	**0.93**	**0.53**

Manipulation Capability

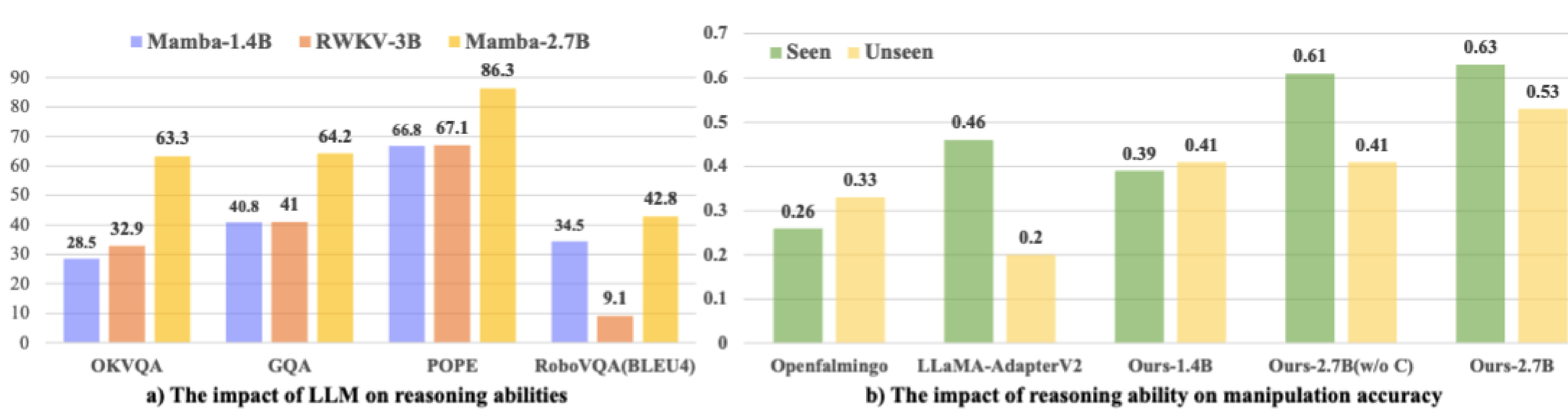


Figure 3: **Ablation study a)** The impact of LLM on reasoning abilities. **Ablation study b)** The impact of reasoning ability on manipulation accuracy.

Ablation Study

Paper Link

Web Page