NEURAL INFORMATION PROCESSING SYSTEMS

ChatVLA-2: Vision-Language-Action Model with Open-World Embodied Reasoning from Pretrained Knowledge

Zhongyi Zhou*, Yichen Zhu* †, Xiaoyu Liu, Zhibin Tang,
Junjie Wen, Yaxin Peng, Chaomin Shen†, Yi Xu
∗ Co-first author. † Corresponding author.

Website: https://chatvla-2.github.io/

Contact: X: Zhongyi Zhou @cat2045535 RedNote: 743282100 WeChat: zzy89OuO

RedNote **WeChat**

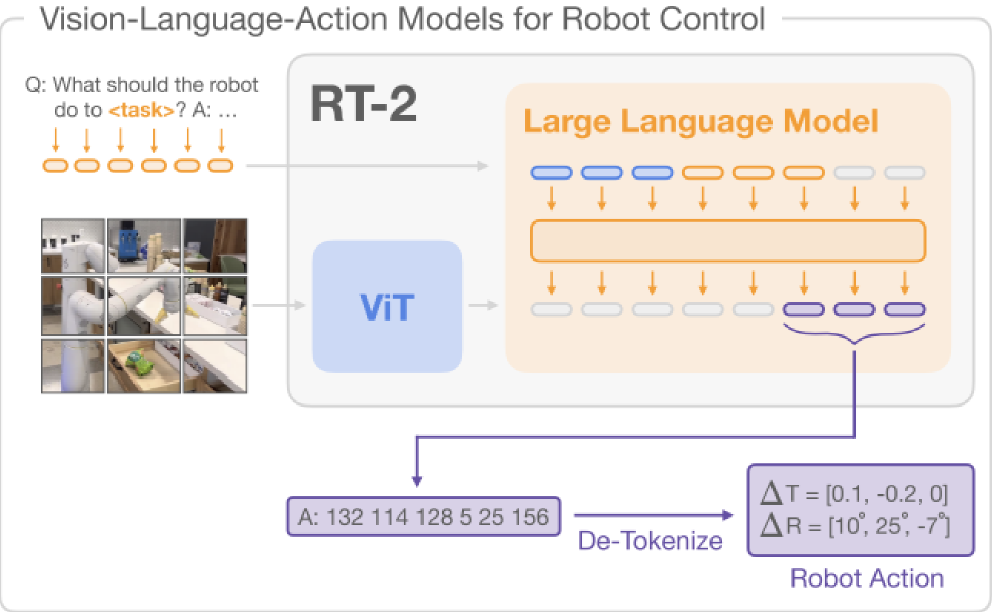
Part 1.I-Background： From VLM to VLA

- Visual-Language Model（VLM）
  - Integrating visual and language information through cross-modal alignment to achieve scene understanding

- Visual-Language-Action Model (VLA)

Vision Language Action Model (VLA)
= Vision-Language Model (VLM) + Action Model

  - Integrate vision, language, and action modalities to enable **end-to-end learning from perception to decision to execution**.
  - They typically **fine-tune the pre-trained VLMs to predict robot actions**.



Brohan A, Brown N, Carbajal J, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control[J]. arXiv preprint arXiv:2307.15818, 2023.

Part 1.II-Motivation

Question: Since modern VLA architectures **build upon pre-trained vision-language models (VLMs)**,

1. Can VLAs equipped with the comprehensive capabilities inherent from VLMs?

e.g. Recognizing everyday objects, reasoning about spatial relationships, and solving mathematical problems

User Find the red ball near the gray cat

VLM No, the red ball is not near the gray cat. The gray cat is sitting on a pillow.

2. Can these capabilities used to enhance generalization?

Part 1.II-Motivation

We argue that **a generalizable VLA model**

should retain and expand upon the VLM's core competencies:

1. Open-world Embodied Reasoning

The VLA should inherit the knowledge from VLM, i.e., recognize anything that the VLM can recognize, be capable of solving math problems, and possess visual-spatial intelligence

2. Reasoning Following

Effectively translating the open-world reasoning into actionable steps for the robot.

Part 1.II-Motivation

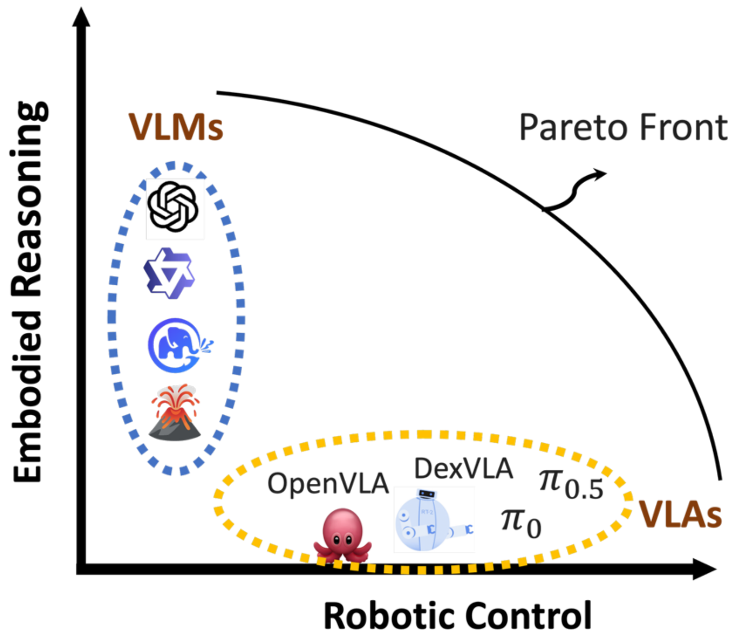
How to enable VLA with
1. Open-world Embodied Reasoning
2. Reasoning Following

considering the large gap between distributions?

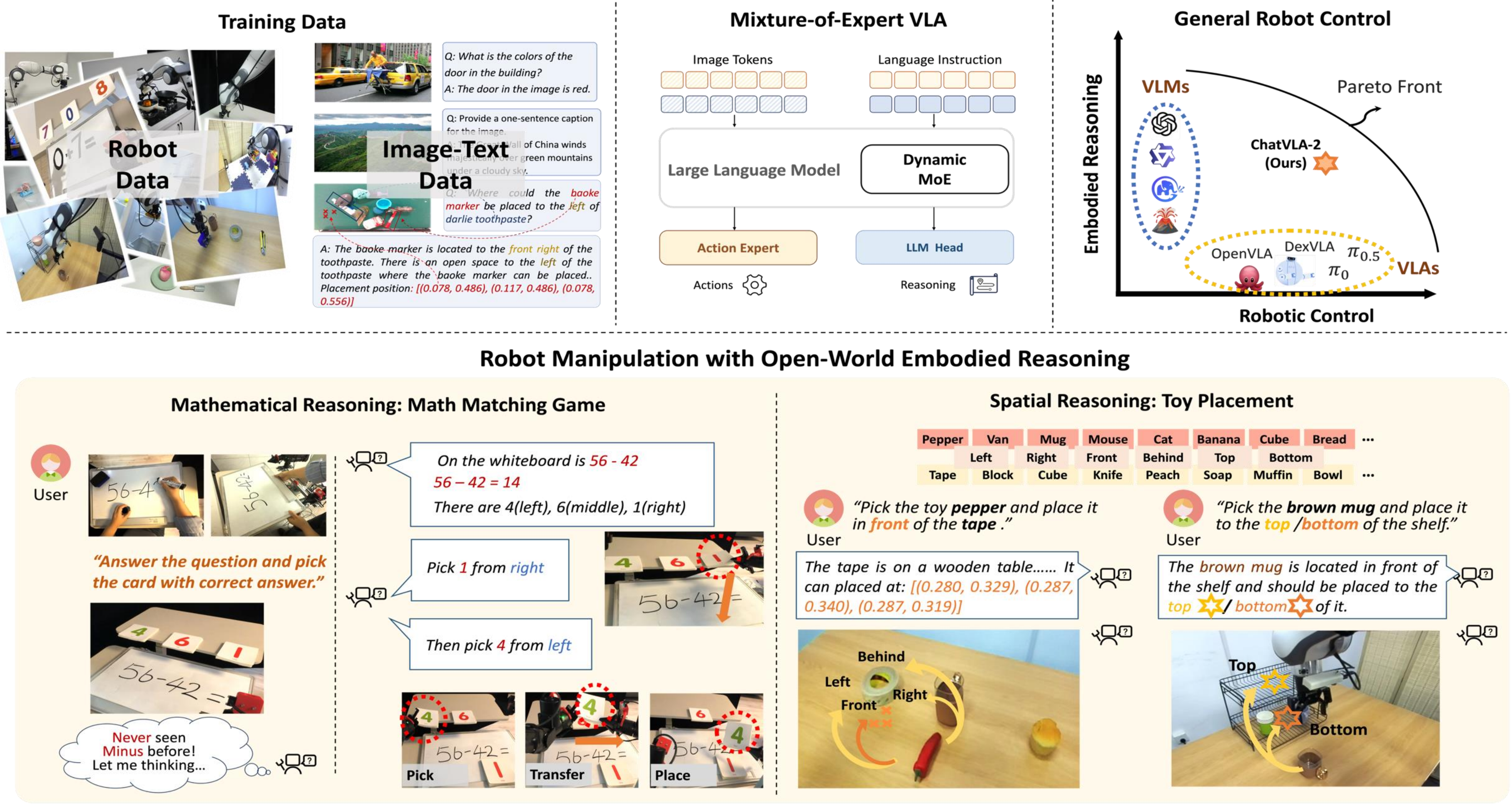
We argue that this can be achieved by adhering to two fundamental principles

1. Identifying overlapping feature spaces between multimodal understanding and robot control

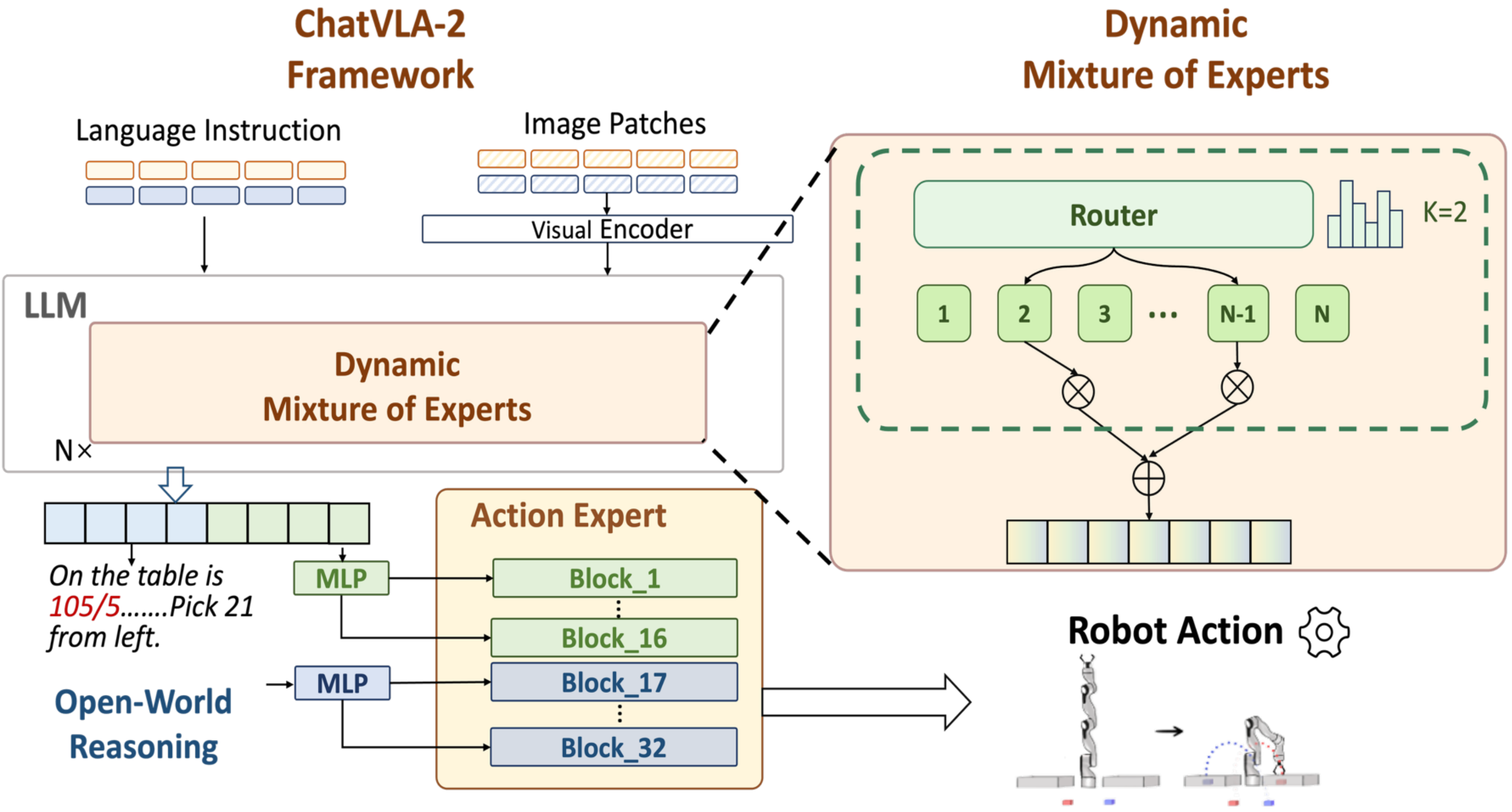
2. Ensuring VLA models act according to their internal reasoning.



Part 2. ChatVLA-2: Vision-Language-Action Model with Open-World Embodied Reasoning from Pretrained Knowledge



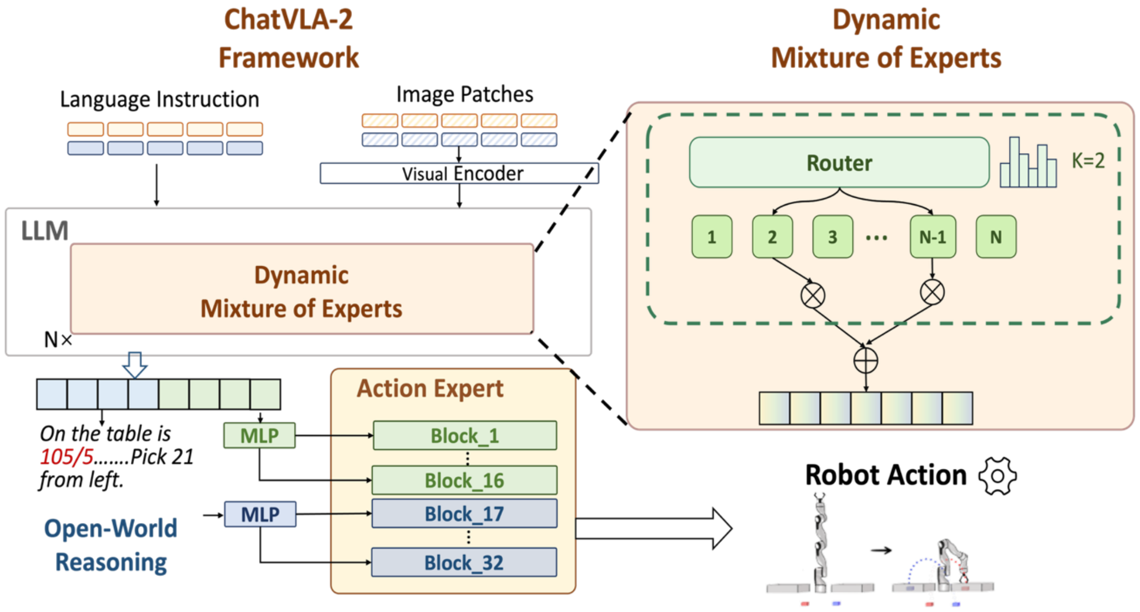
Part 2. ChatVLA-2: Vision-Language-Action Model with Open-World Embodied Reasoning from Pretrained Knowledge



Model Architecture

- A **dynamic mixture-of-experts** architecture to disentangle conflicting features between multimodal understanding and robotic control, while effectively integrating mutually beneficial features.

Part 2. ChatVLA-2: Vision-Language-Action Model with Open-World Embodied Reasoning from Pretrained Knowledge



Model Architecture

- A **reasoning-following enhancement module** is incorporated to ensure that the VLA model adheres to logical reasoning when performing actions.

Original ScaleDP AdaLN:

$$\mathrm{AdaLN}_i(x) = (\gamma_i(t, o) + 1)\, x + \beta_i(t, o), \quad 1 \le i \le N.$$

ChatVLA-2 AdaLN (piecewise over depth):

$$\mathrm{AdaLN}_i(x) = (\gamma_i(t, o) + 1)\, x + \beta_i(t, o) \quad \text{if } i < \frac{N}{2}$$

$$\mathrm{AdaLN}_i(x) = (\gamma_i(t, r) + 1)\, x + \beta_i(t, r) \quad \text{if } i \ge \frac{N}{2}$$

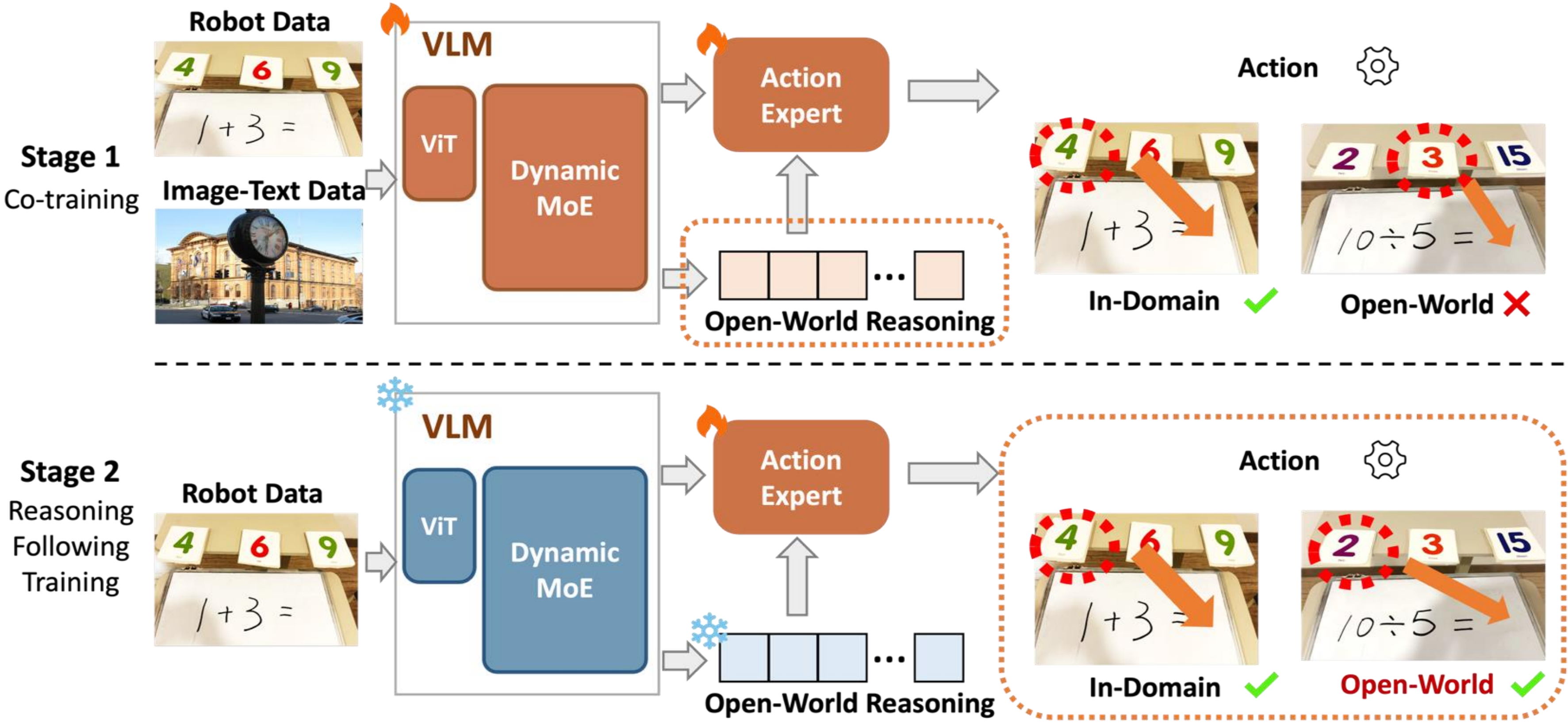
where the original observation o **is replaced to reasoning r** in the latter half layers.

Part 2. ChatVLA-2: Vision-Language-Action Model with Open-World Embodied Reasoning from Pretrained Knowledge

Two-Stage Training Strategy

- Stage 1

Co-training on image-text and robot data is essential for enabling the robot foundation model to reason and understand scenes in the wild. During this stage, we train the model on both tasks.

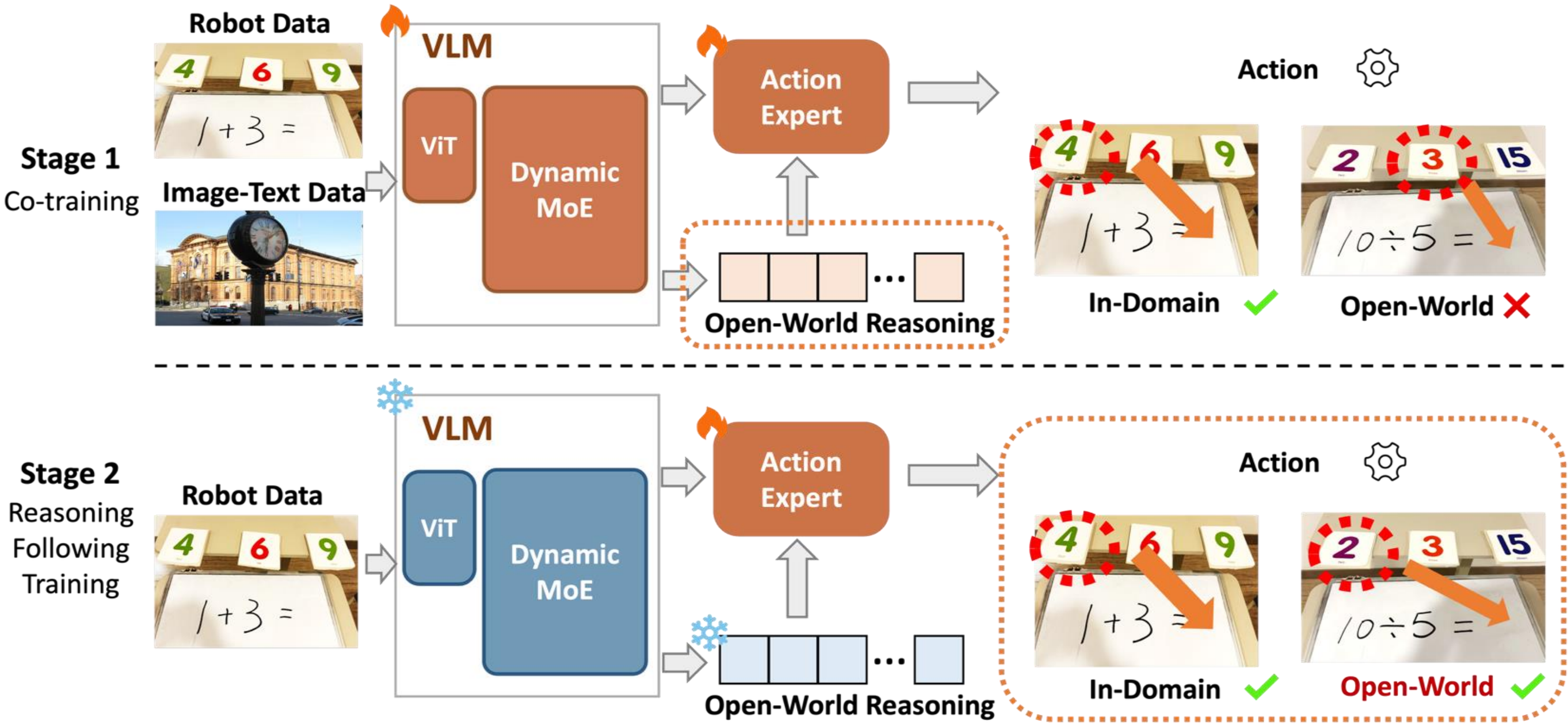


Part 2. ChatVLA-2: Vision-Language-Action Model with Open-World Embodied Reasoning from Pretrained Knowledge

Two-Stage Training Strategy

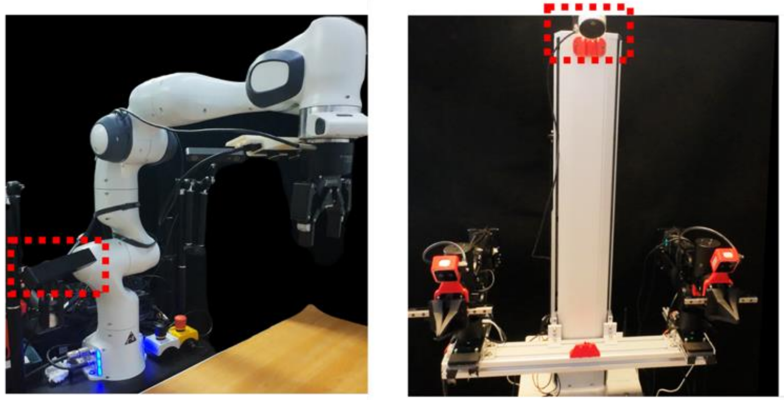
- Stage 2

We freeze the entire VLM and train only the action expert, thereby **preserving open-world reasoning** while **enhancing instruction-following abilities** in VLA.



Part 3.I- Experimental setup

Franka Setup | ARX Setup

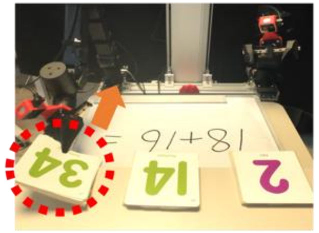


Camera

Math Matching Game

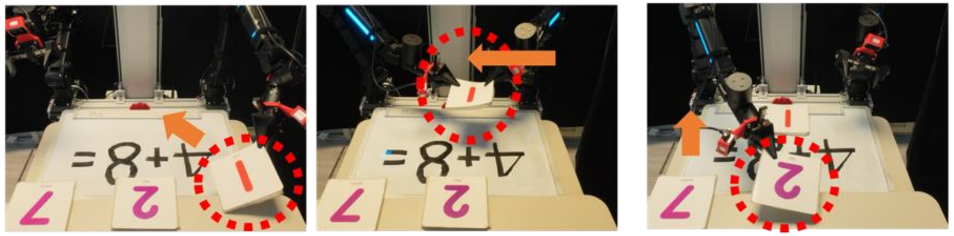
Instruction: *Answer the question and pick the card with correct answer.*

1 Step



On the board is 18+16=34...
Pick 34 from right

2 Steps



On the board is 4+8, 4+8=12......
There are 1(left), 2(middle), 7(right)
First, pick 1 from left

Then pick 2 from middle

Toy Placement

Instruction: *Pick the [obj] and place it to [place] of the [target]*

Pick the avocado | Pick the corn | Pick the mug

...

To

Behind the tape | Left of pink block | Right of orange bus | Front of bowl | Top/Bottom of shelf

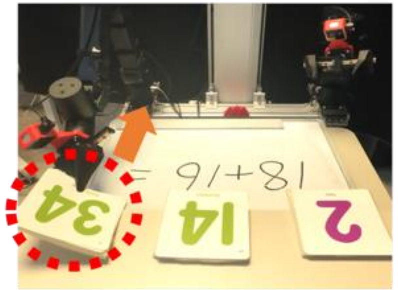
...

Part 3.I- Results on Mathematical Reasoning

Mathematical Reasoning: Math Matching Game

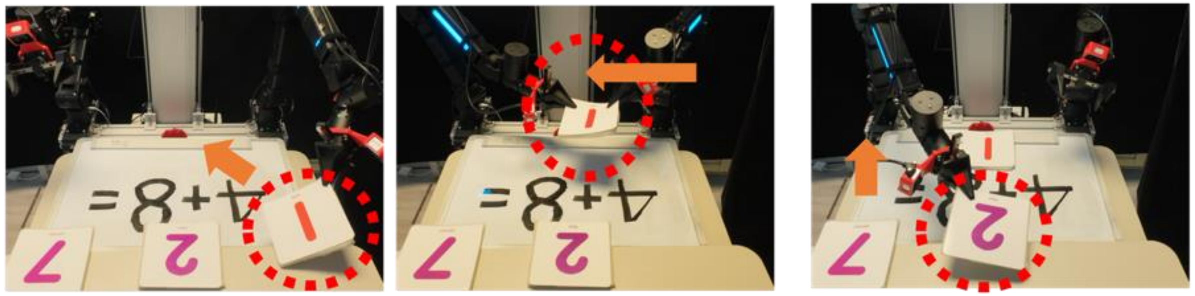
Instruction: *Answer the question and pick the card with correct answer.*

1 Step



On the board is 18+16=34…
Pick 34 from right

2 Steps



On the board is 4+8, 4+8=12……
There are 1(left), 2(middle), 7(right)
First, pick 1 from left

Then pick 2
from middle

Evaluation metrics

1) Manipulation success rate
2) OCR
   - recognizing hand-written numbers: 1’
   - identifying card values and their positions: 1’
   - correctly recognizing the sign: 2’
3) Mathematical reasoning
   - correct answer: 1’
   - correctly selecting the card: 1’

Part 3.I- Results on Mathematical Reasoning

Table 1: **Results on the math matching game.** We evaluate multiple models on both in-domain settings, where the data is presented in the training data, and open-world setups. We evaluate average score of **OCR (4 scores in total)** and **mathematical reasoning (2 scores in total)**, and average success rate of task execution at both setups.

Method	In Domain		Open-World		
	Reasoning Score	Success Rate	OCR Score	Math Reasoning Score	Success Rate
Octo [70]	/	2/13	/	/	0/52
Diffusion Policy [32]	/	7/13	/	/	3/52
OpenVLA [31]	/	2/13	/	/	0/52
GR00T N1 [66]	/	4/13	/	/	3/52
DexVLA [2]	5.2/6	**12/13**	0.21/4	0.06/2	10/52
ChatVLA [7]	5.8/6	10/13	1.08/4	0.42/2	4/52
π_0 [1]	/	**12/13**	/	/	8/52
ChatVLA-2 (Ours)	**6.0/6**	11/13	**3.58/4**	**1.73/2**	**43/52**

Part 3.II- Experimental setup

Spatial Reasoning: Toy Placement

Instruction: *Pick the [obj] and place it to [place] of the [target]*

Pick the avocado | Pick the corn | Pick the mug ...

To

Behind the tape | Left of pink block | Right of orange bus | Front of bowl | Top/Bottom of shelf ...

Evaluation metrics

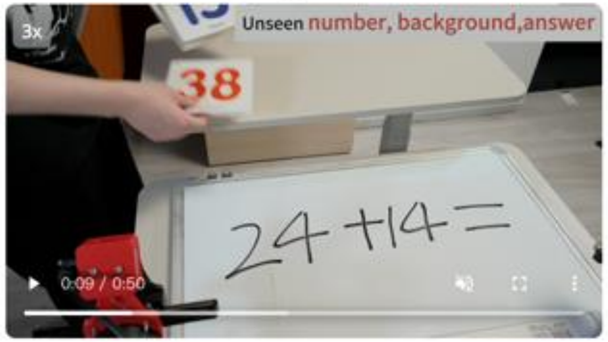
1) Manipulation success rate
2) Object recognition score
3) Spatial affordance score

Table 2: **Results on the toy placement task.** We evaluate multiple models on both in-domain settings, where the data is presented in the training data, and open-world setups. We evaluate average object recognition score, spatial affordance score and task success rate at both setups.

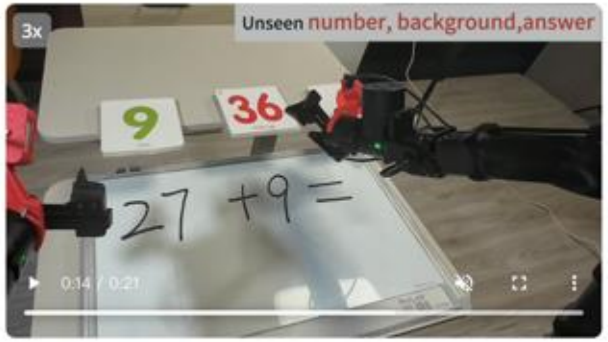
Method Manipulation	In Domain Object recognition	In Domain Spatial Affordance	In Domain Avg. Success Rate	Open-World Object recognition	Open-World Spatial Affordance	Open-World Avg. Success Rate
Octo [70]	/	/	19/67	/	/	13/156
Diffusion Policy [32]	/	/	52/67	/	/	17/156
OpenVLA [10]	/	/	23/67	/	/	10/156
GR00T N1 [66]	/	/	31/67	/	/	12/156
DexVLA [2]	**1**	0.97	**63/67**	0.23	0.12	36/156
ChatVLA [7]	**1**	0.97	60/67	0.71	0.35	22/156
π_0 [1]	/	/	61/67	/	/	25/156
ChatVLA-2 (Ours)	**1**	**0.99**	61/67	**0.94**	**0.88**	**127/156**

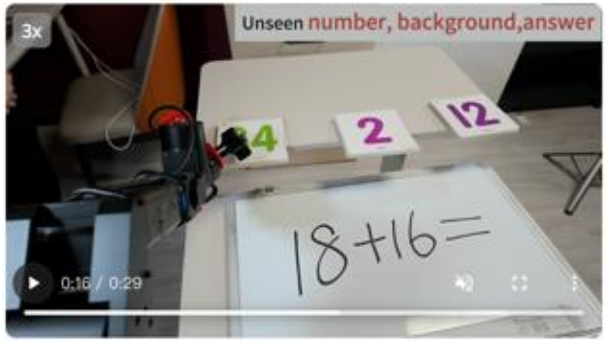
Part 4 - Demos

Solving Sequential Tasks



Unseen number, answer, background, and even symbol!

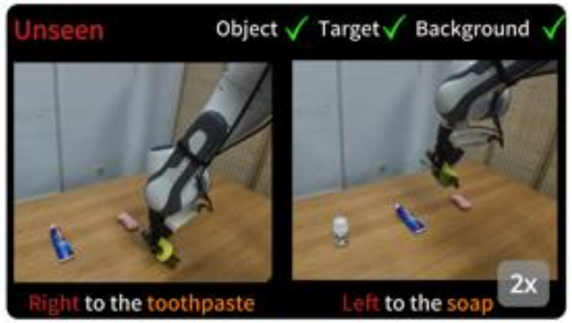


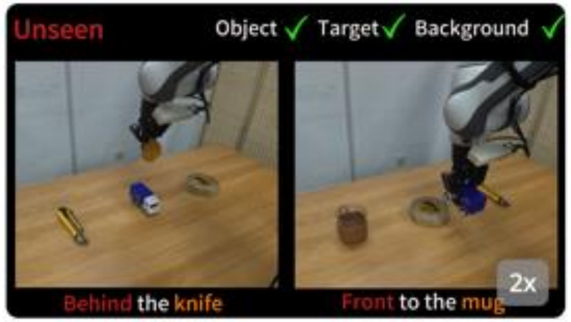


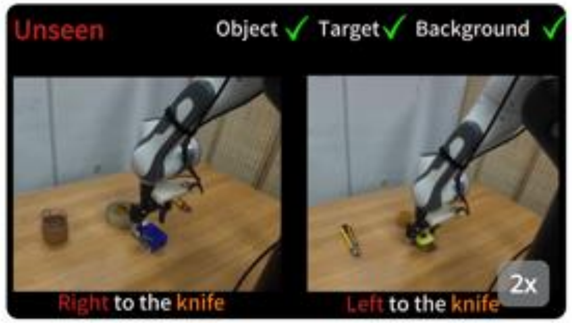
More on website!

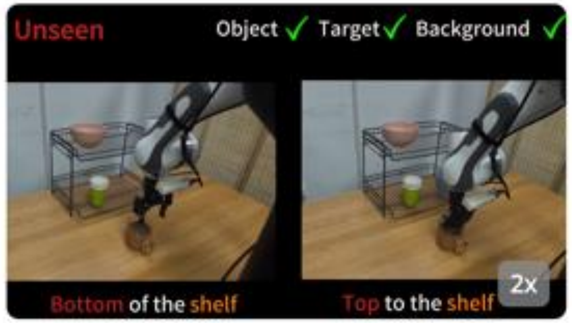
https://chatvla-2.github.io/

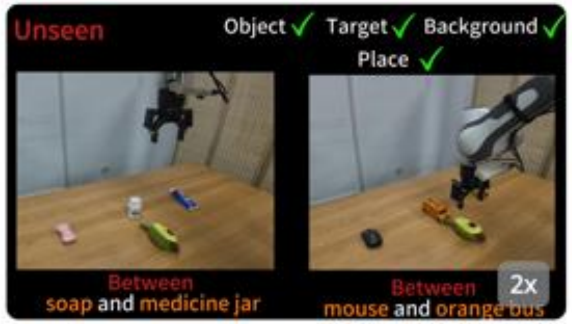
Part 4 - Demos

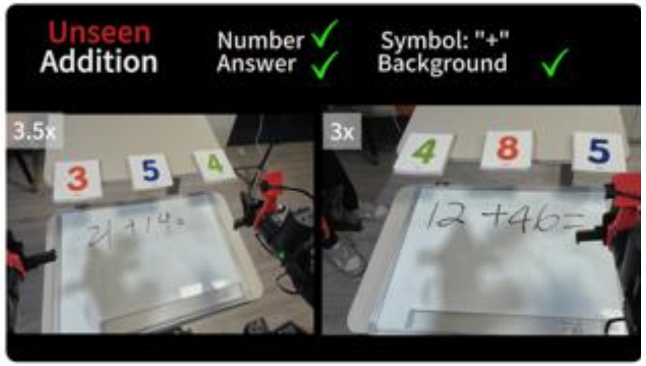


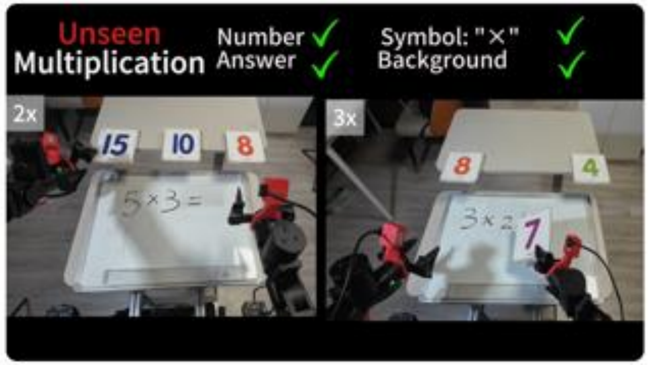


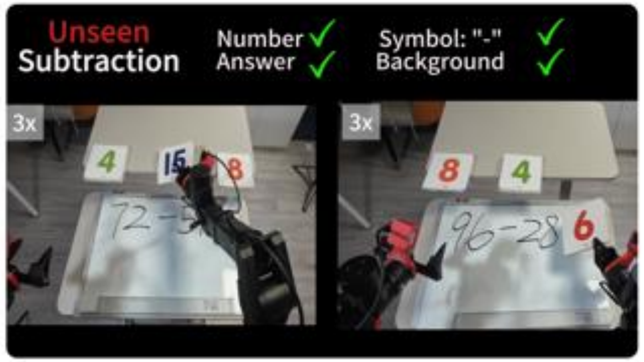


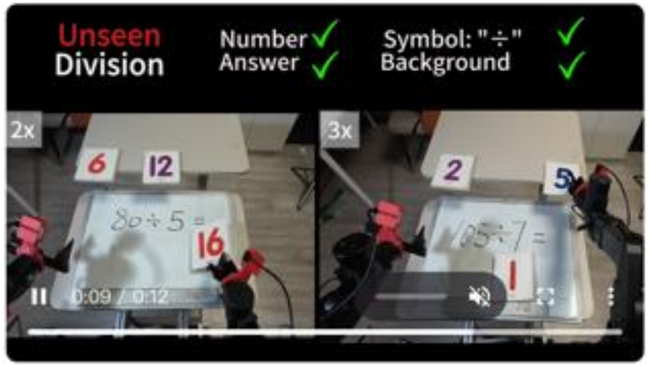












More on website!

https://chatvla-2.github.io/

Thanks

Website: https://chatvla-2.github.io/

Contact: X: Zhongyi Zhou @cat2045535 RedNote: 743282100 WeChat: zzy89OuO

RedNote **WeChat**