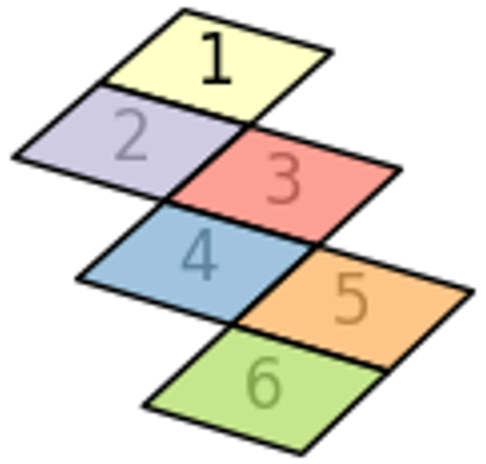
See. Think. Act.

Training Multimodal Agents with Reinforcement Learning

Linjie Li

06/12/2025

Question

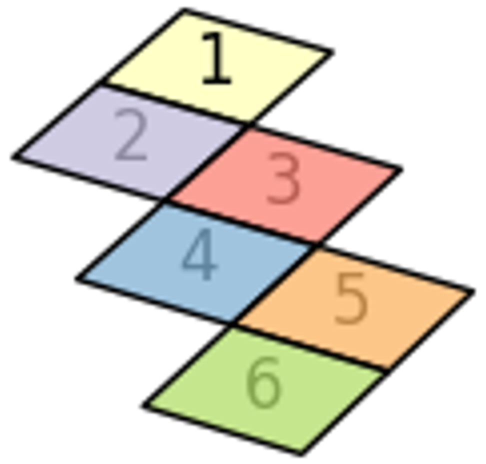


Can the net be folded to form a cube, yes or no?

Yes!

Image credit: “Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations”

Question



Can the net be folded to form a cube, yes or no?

Step-by-Step Human Mental Simulation

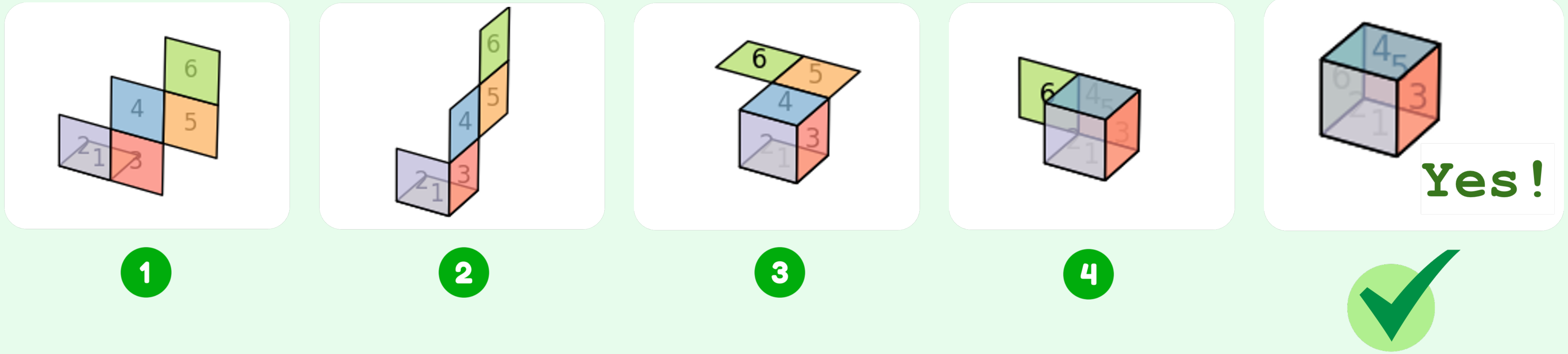


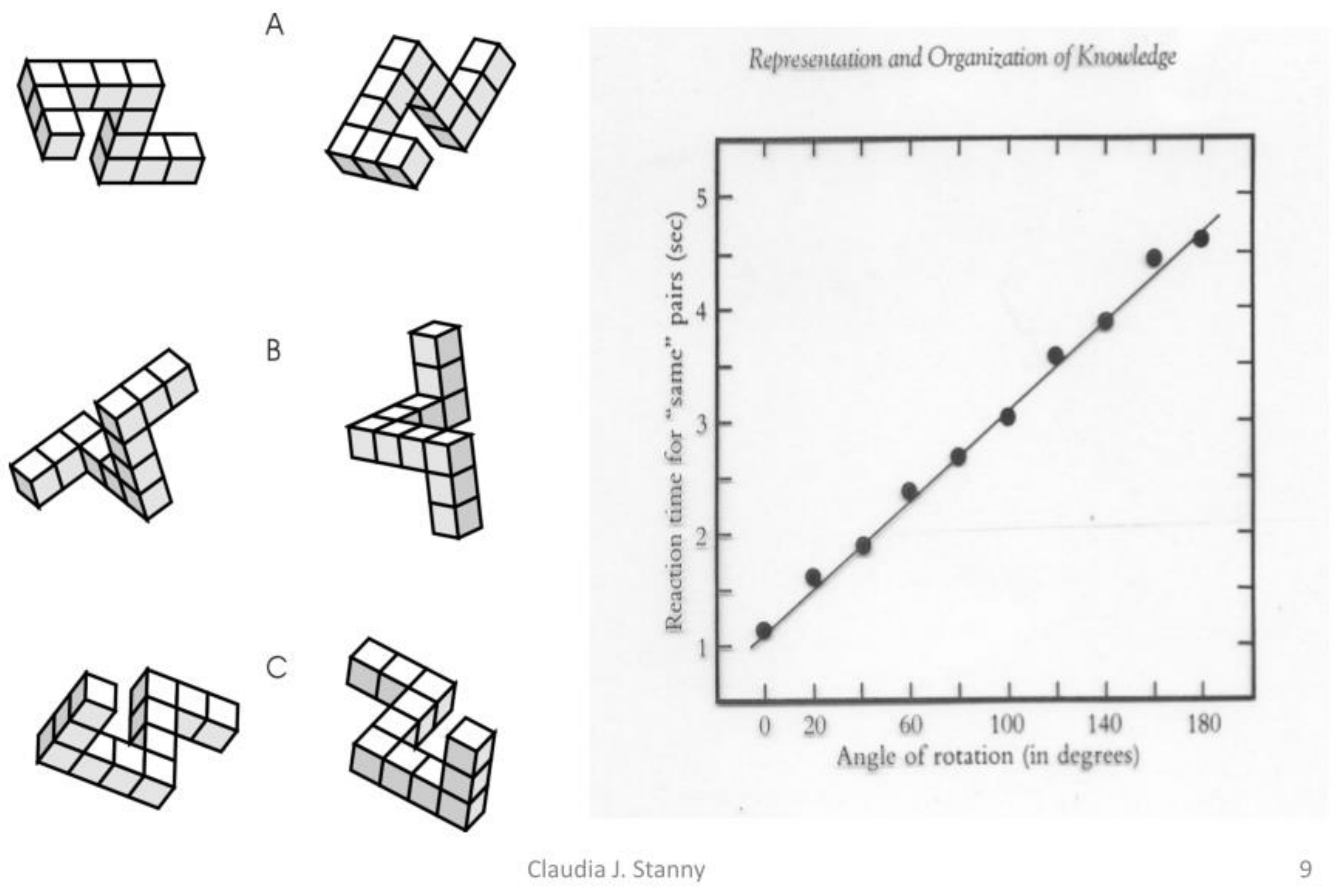
Image credit: “Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations”

STARE

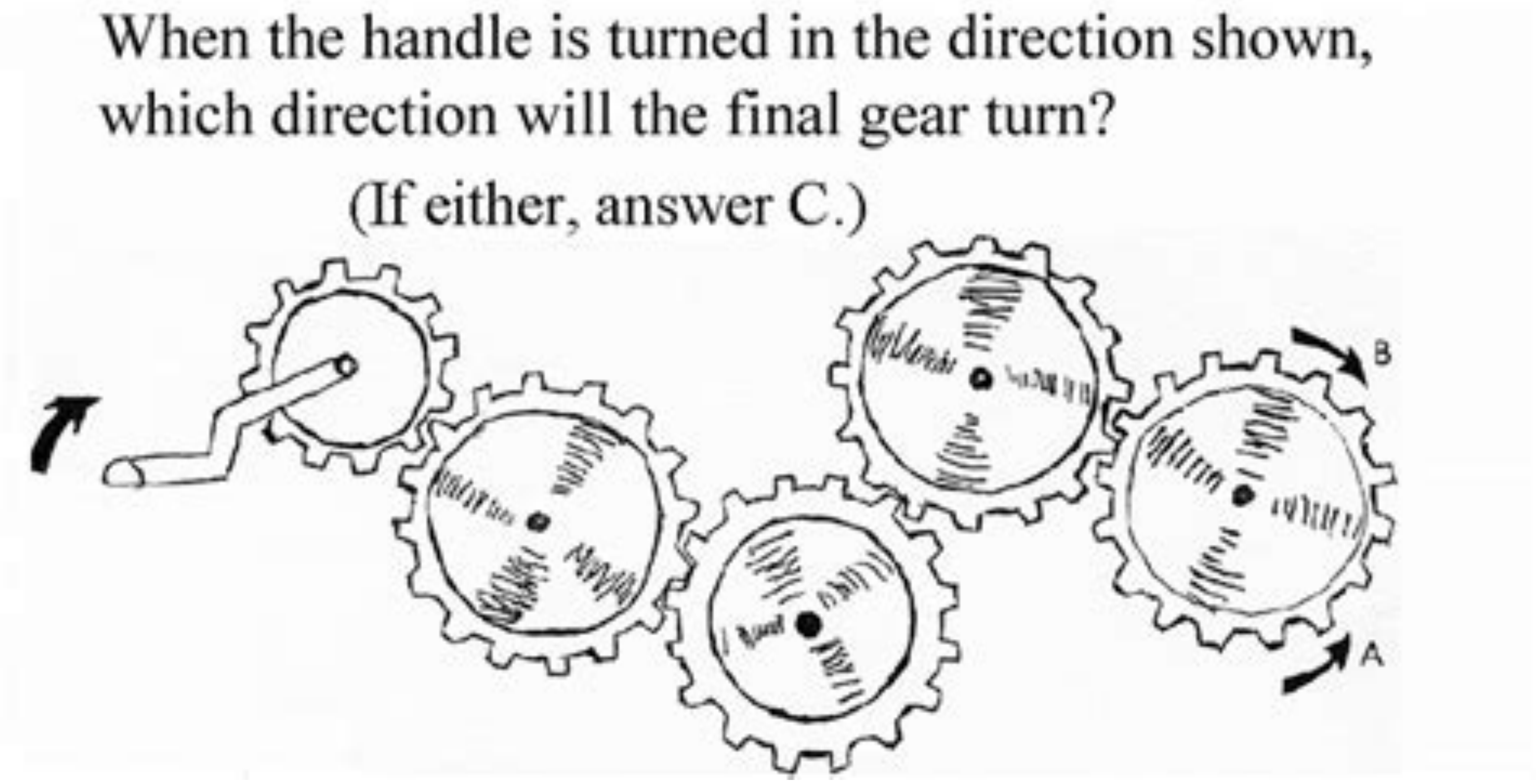
Image credit: "Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations"

Visual Simulation is Critical to Human (non-verbal) Reasoning

Mental Rotation
(Shepard & Metzler, 1971)



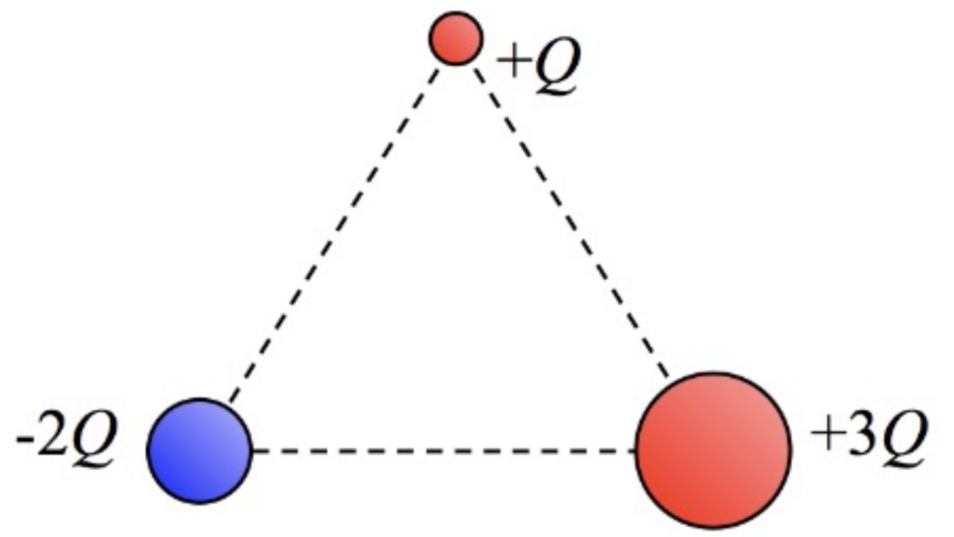
Mechanical reasoning by mental simulation
(Hegarty, 2004)



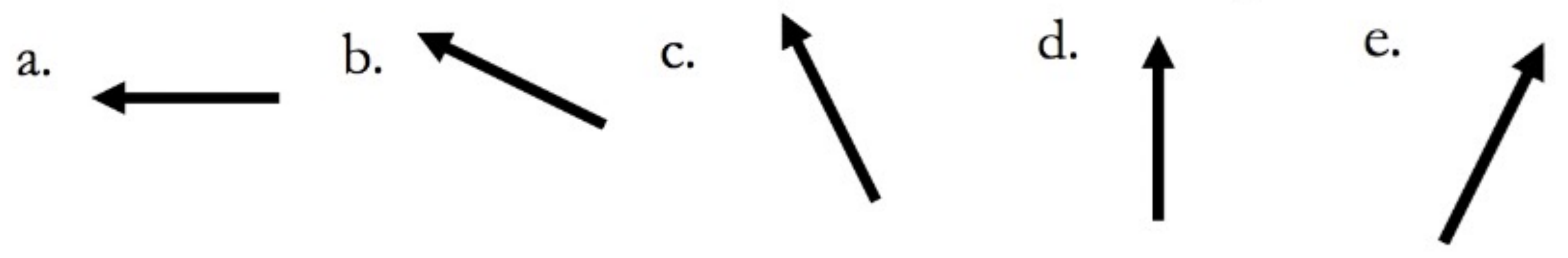
Visual Simulation is Critical to Human (non-verbal) Reasoning

"Spatial ability predicts performance in mathematics and eventual expertise in science, technology and engineering." (Tosto, M. G. *et al. 2014*)

Multimodal reasoning question in EMMA



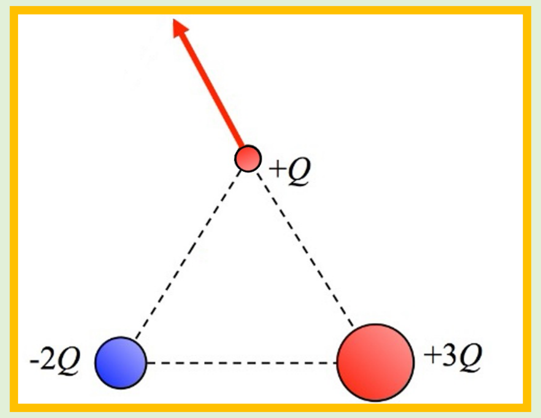
Question: Three point charges, of charge $+Q$, $-2Q$, and $+3Q$, are placed equidistant as shown. Which vector best describes the net direction of the electric force acting on the $+Q$ charge?



Human reasoning

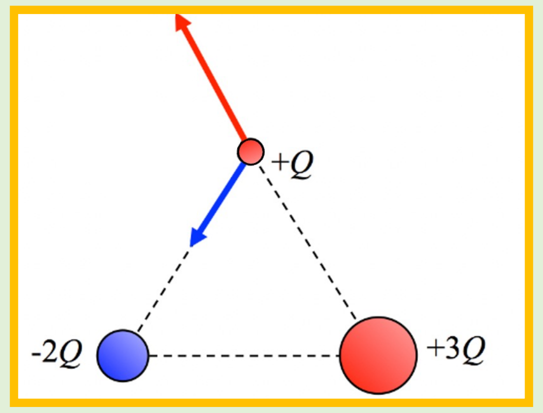
The direction of the electric force due to $-2Q$ and $+3Q$ can be visualized with a quick free-body sketch of the electric forces:

1



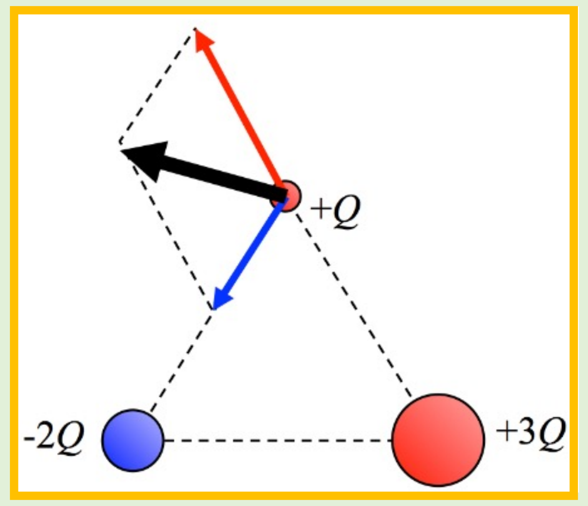
repel

2



attract

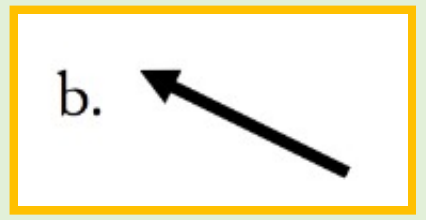
3



Combining the forces:

∵ |+3Q|>|-2Q|

∴ repel > attract



Model reasoning (GPT-4o)

1. **Forces Due to the Charges**:
   - The $-2Q$ charge attracts the $+Q$ charge. (Force points downward-left).
   - The $+3Q$ charge repels the $+Q$ charge. (Force points downward-right).
2. **Force Magnitudes (Coulomb's Law)**:
   - Force due to $-2Q$: $F_{-2Q} = \frac{2kQ^2}{r^2}$.
   - Force due to $+3Q$: $F_{+3Q} = \frac{3kQ^2}{r^2}$.

   ……

Error: The issue arose from not correctly identifying the force direction from $+3Q$ to $+Q$, which causes the net force to point ↖, not downward-right↘.

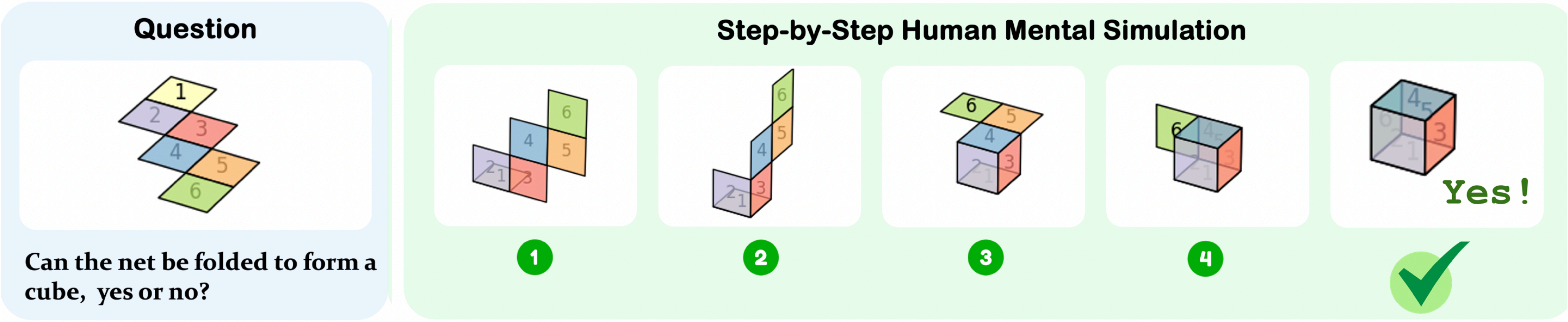
Image credit: "Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark"

See. Visual Think. Act.

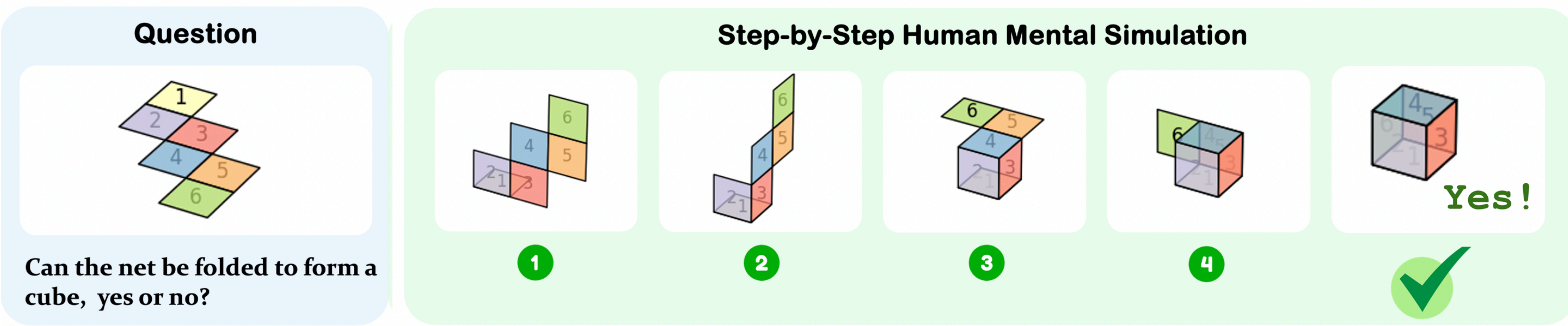
Training Multimodal Agents with Reinforcement Learning

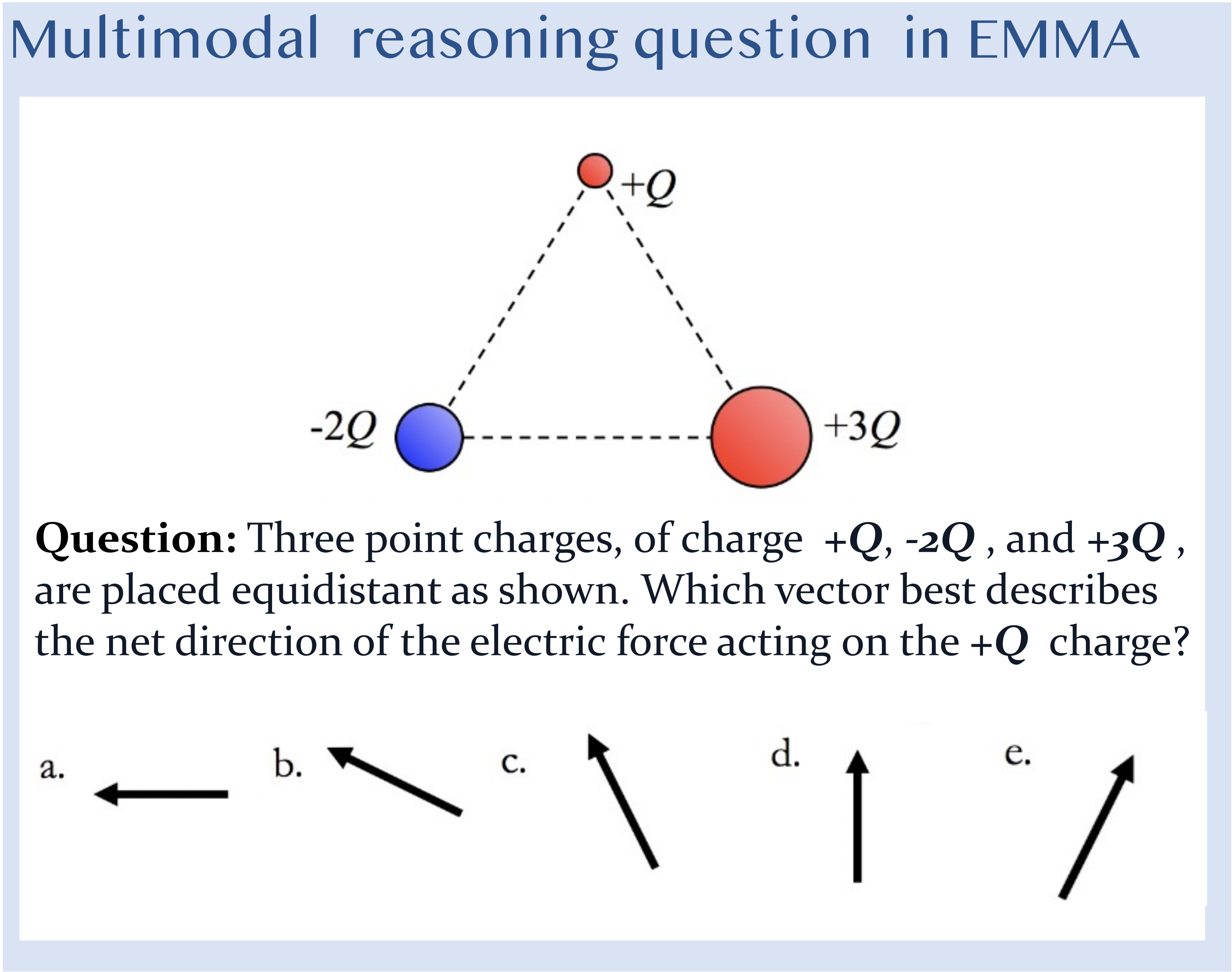
Visual Simulation is Critical to Human (non-verbal) Reasoning

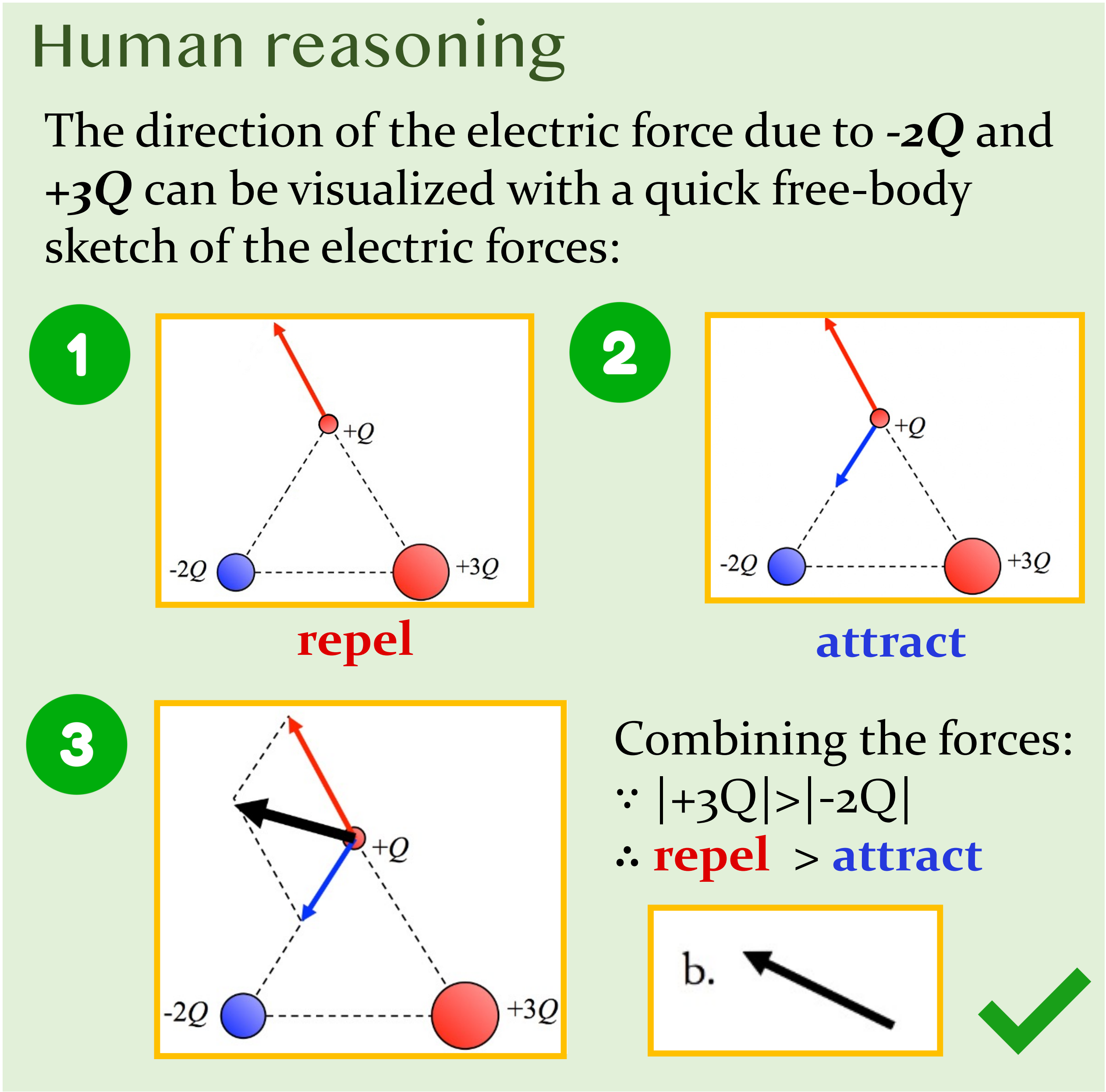
People with spatial intelligence - "skillfully use the ability to create images, spatial relationships, and visualizations in the mind." (Pawlak-Jakubowska & Terczyńska 2023)



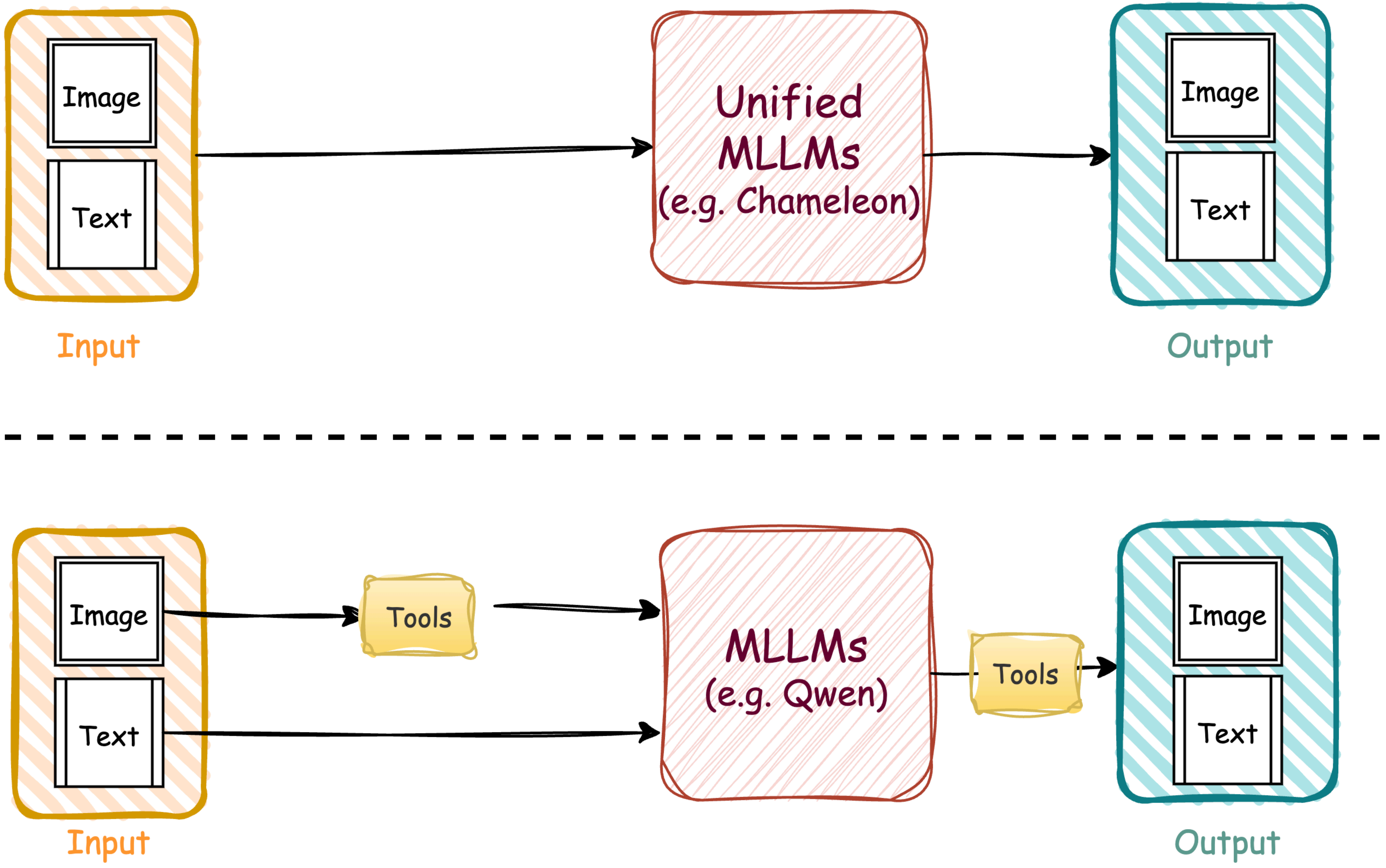
Visual Simulation is Critical to Human (non-verbal) Reasoning



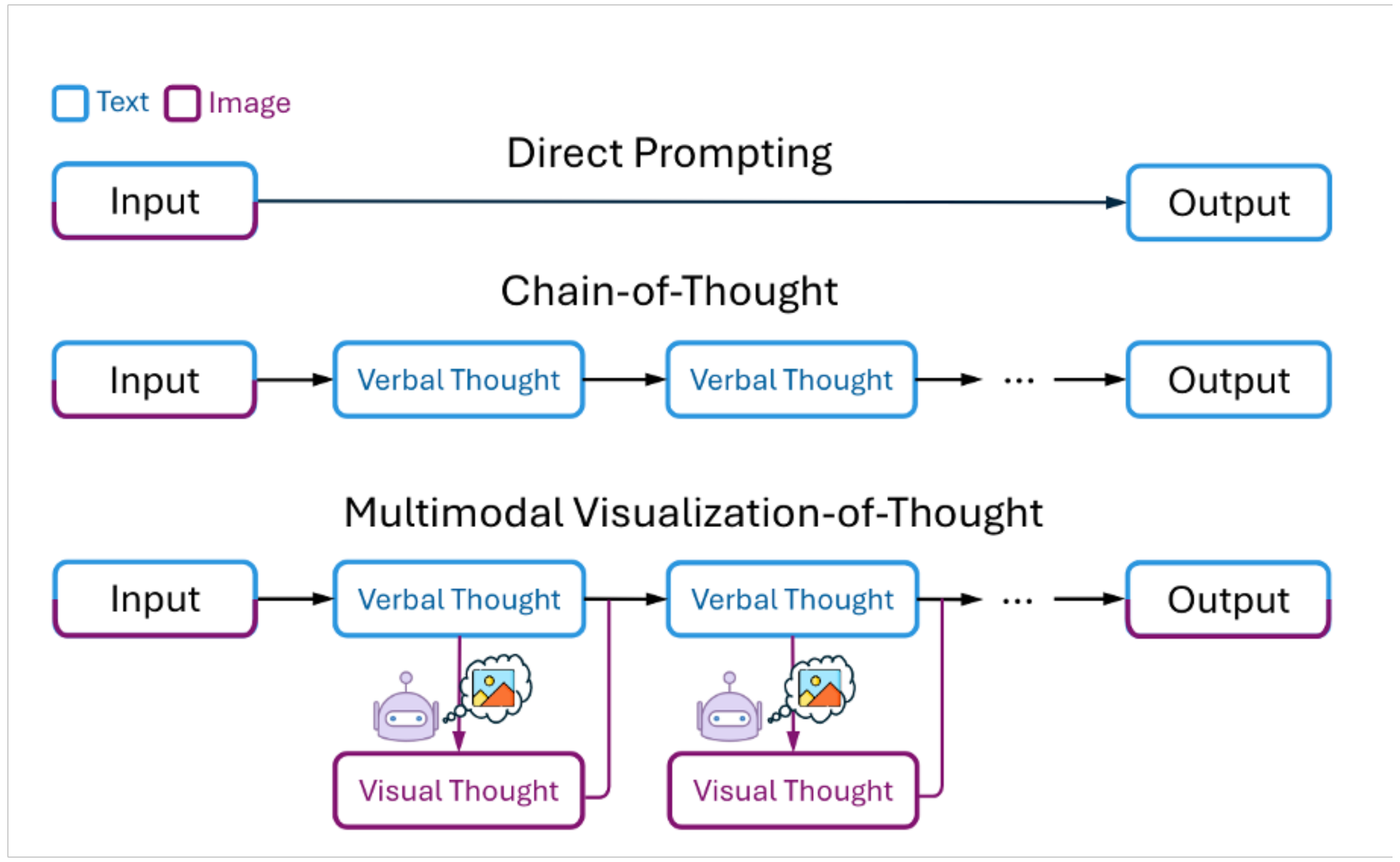


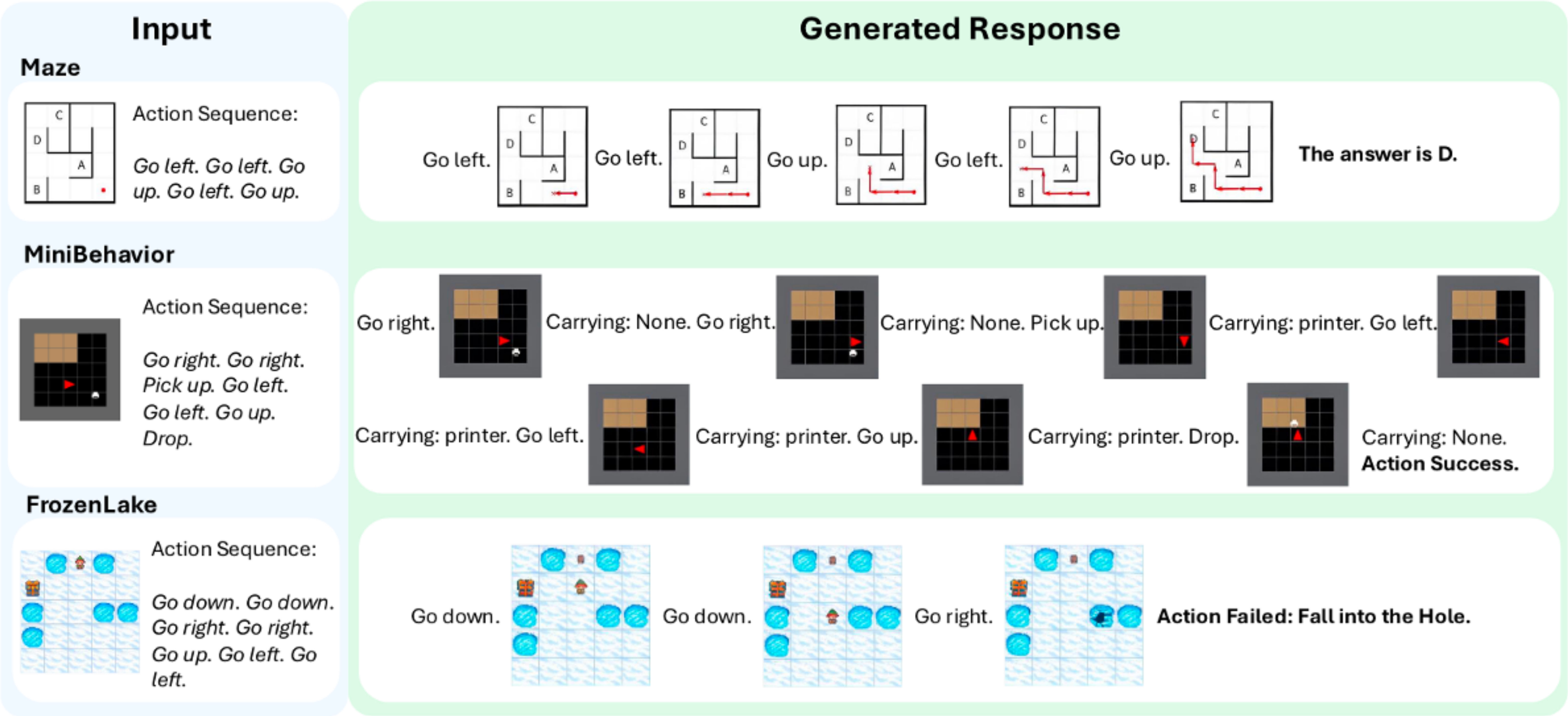


Enabling Models to Think Visually



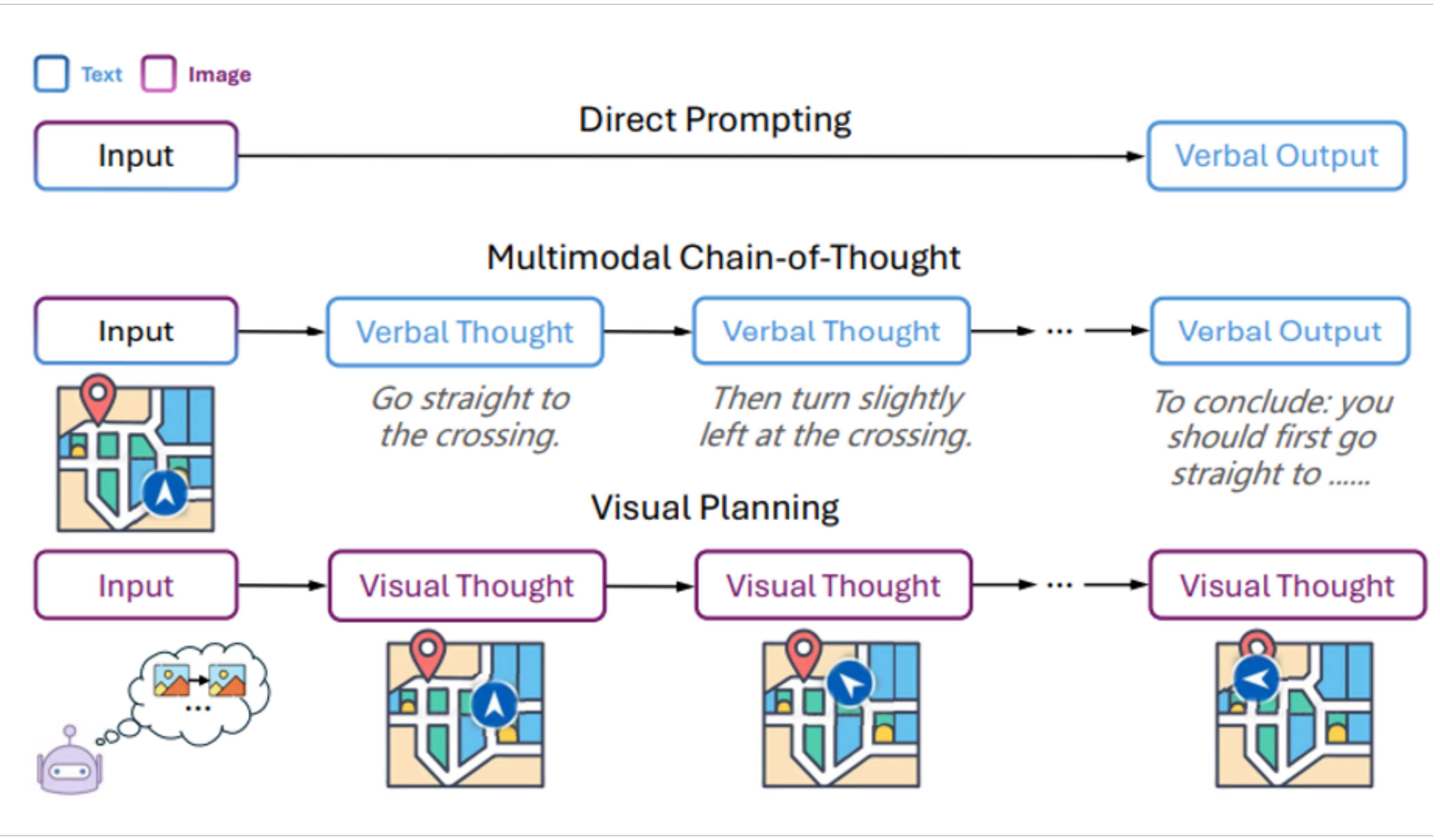
Enabling Models to Think Visually via Image Generation

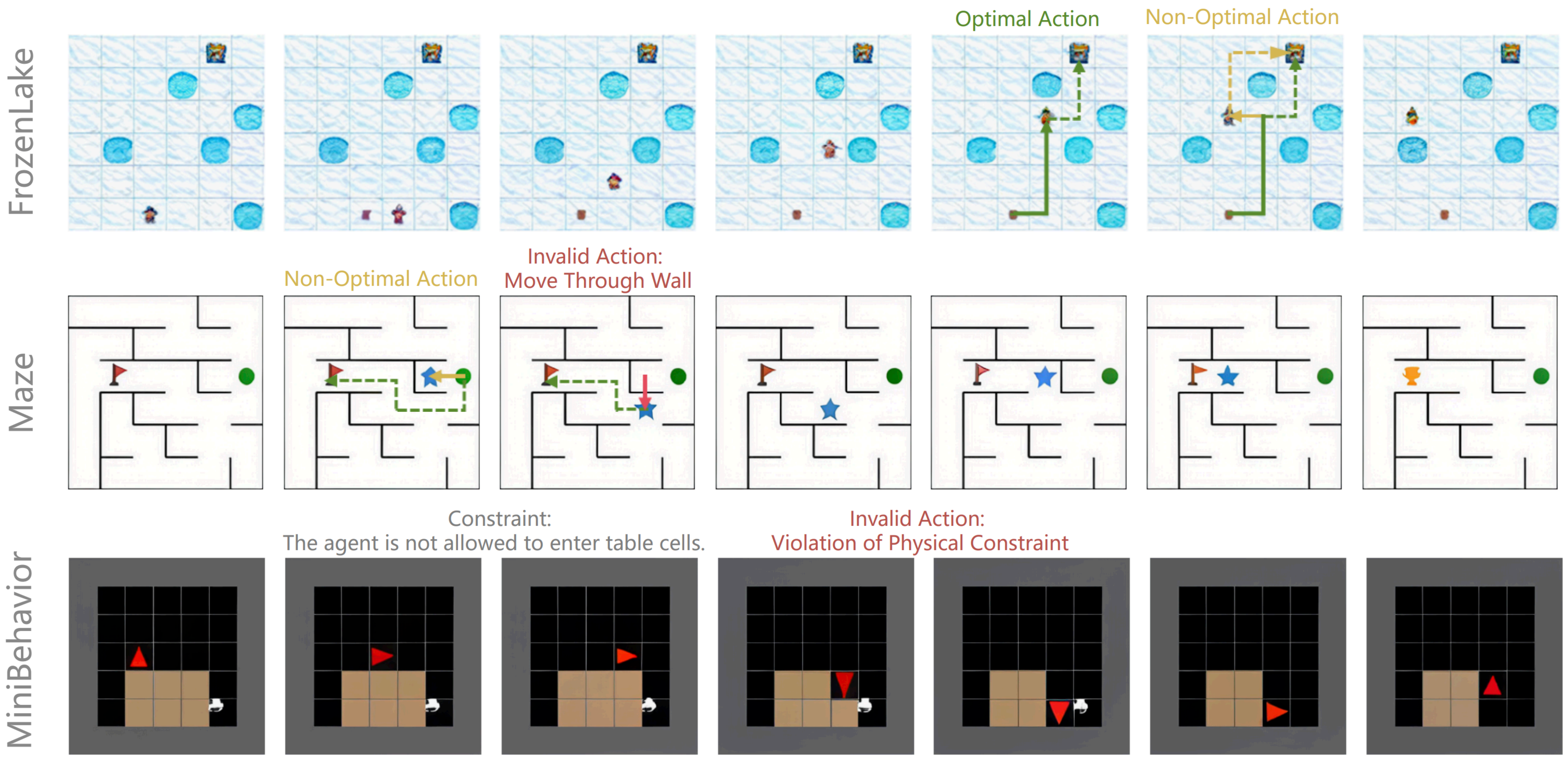




Credit: "Imagine while Reasoning in Space: Multimodal Visualization-of-Thought"

Enabling Models to Think Visually via Image Generation



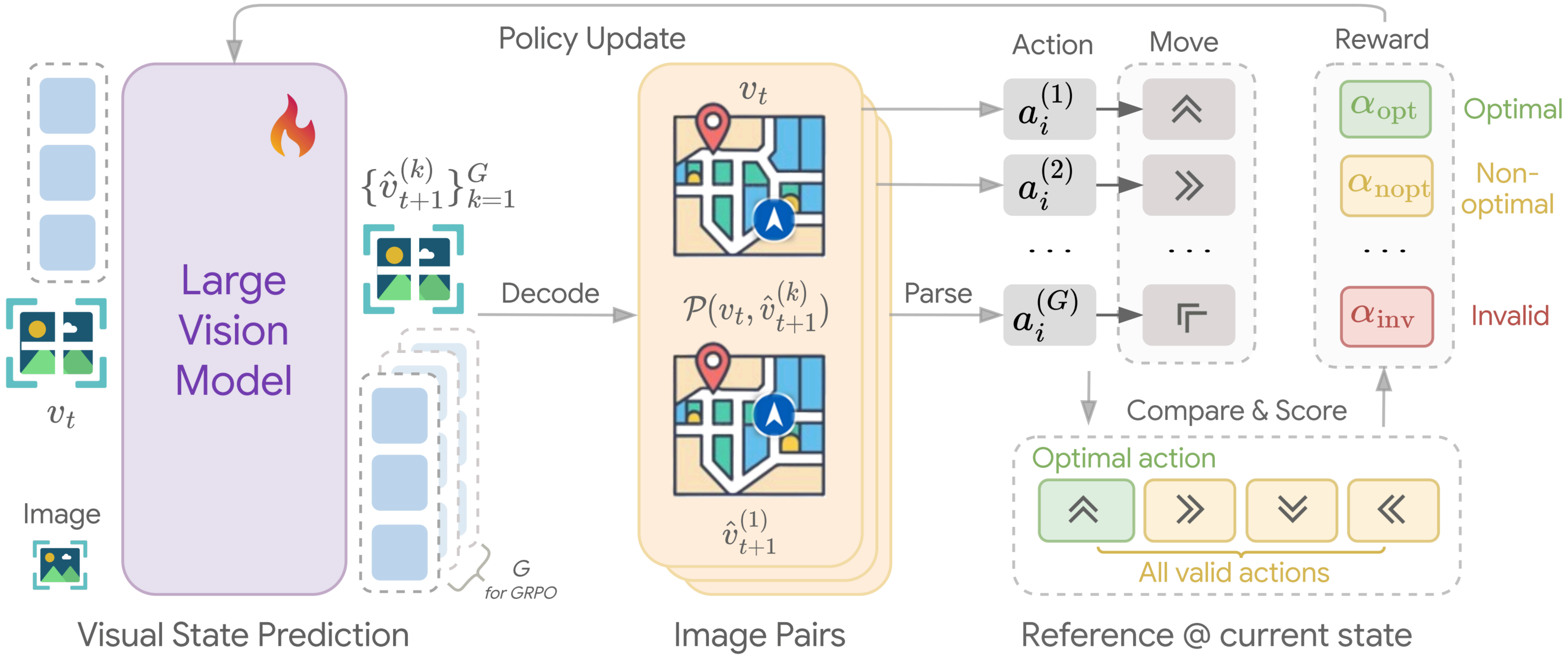


Credit: "Visual Planning: Let's Think Only with Images"

Enabling Models to Think Visually via Image Generation

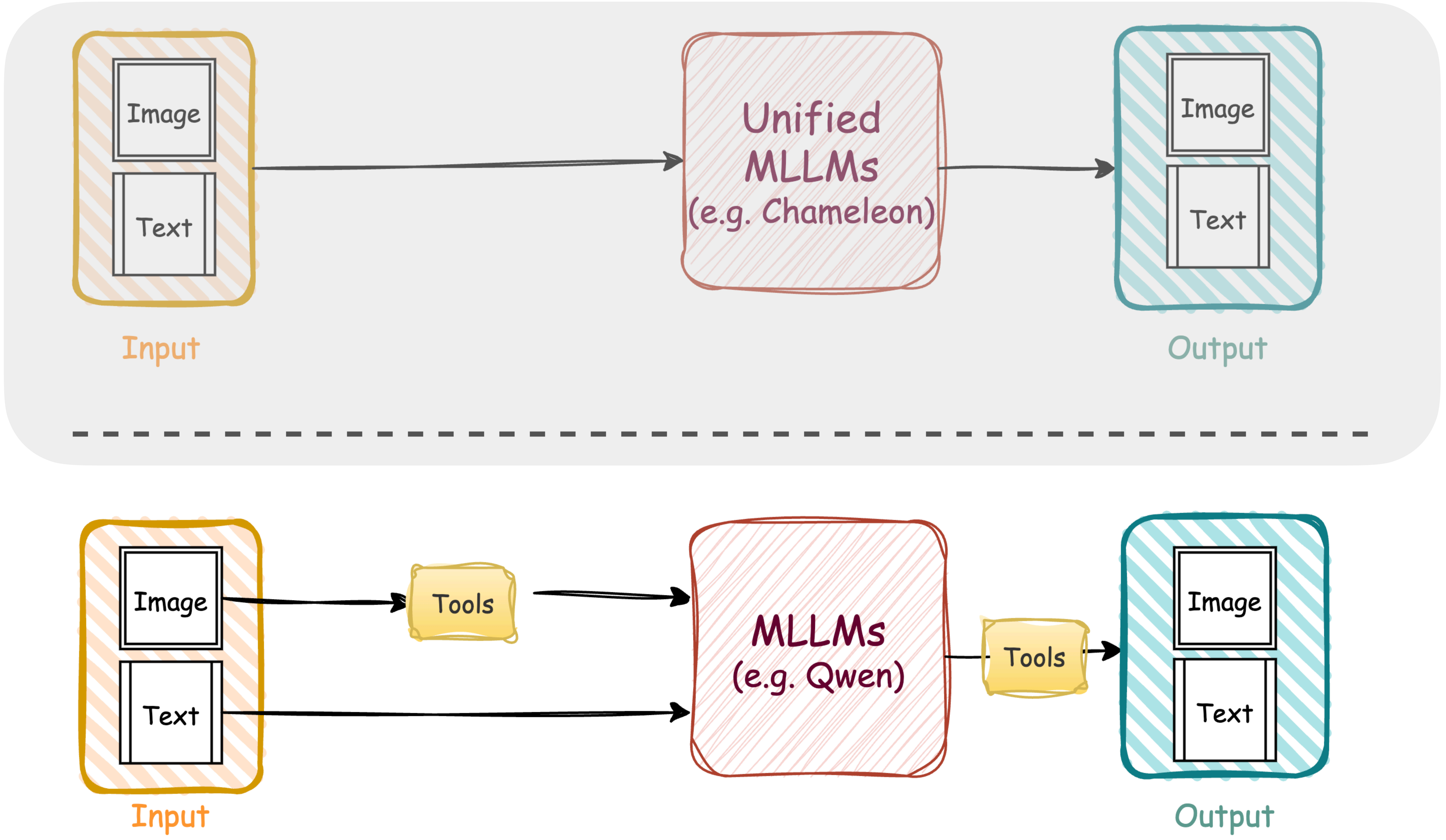
How to reward pure image outputs?

Reward actions - a rule-based parser that turns image-to-image transitions into discrete moves



Credit: "Visual Planning: Let's Think Only with Images"

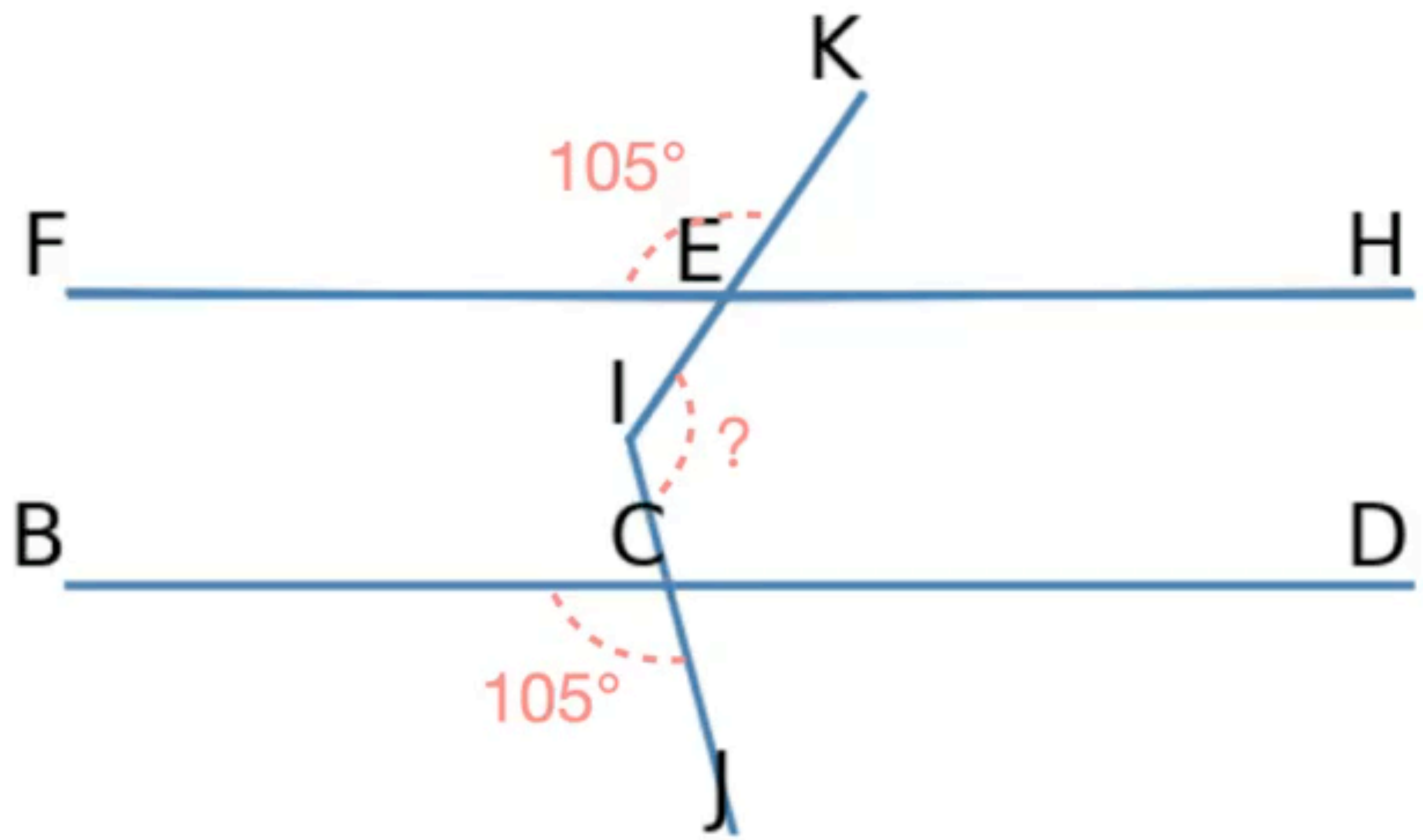
Enabling Models to Think Visually via Tool Using



Enabling Models to Think Visually via Tool Using

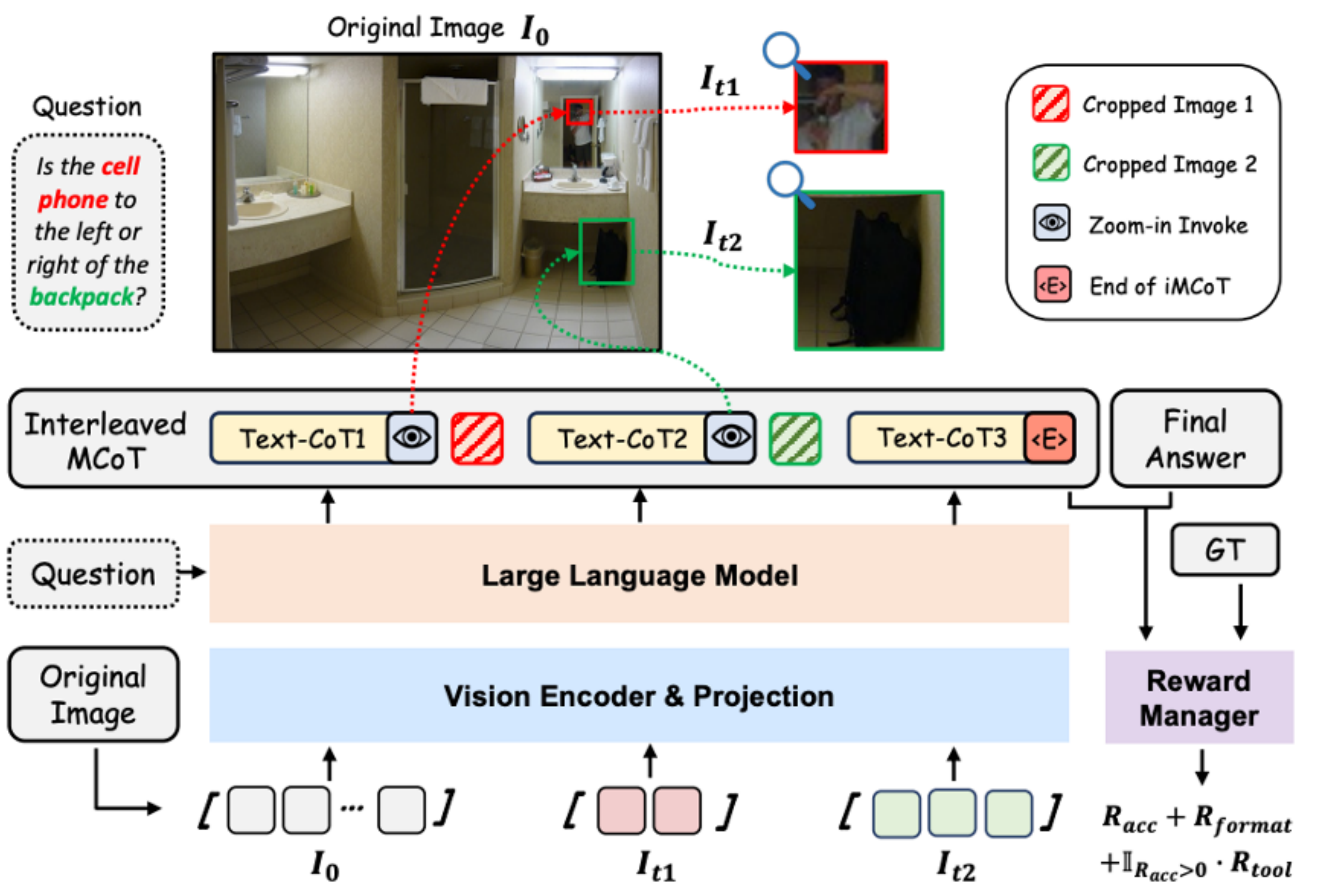
Query: Given ∠BCJ=105°, ∠KEF=105°. Find ∠EIC

Input Image:



Credit: "Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models"

Enabling Models to Think Visually via Tool Using



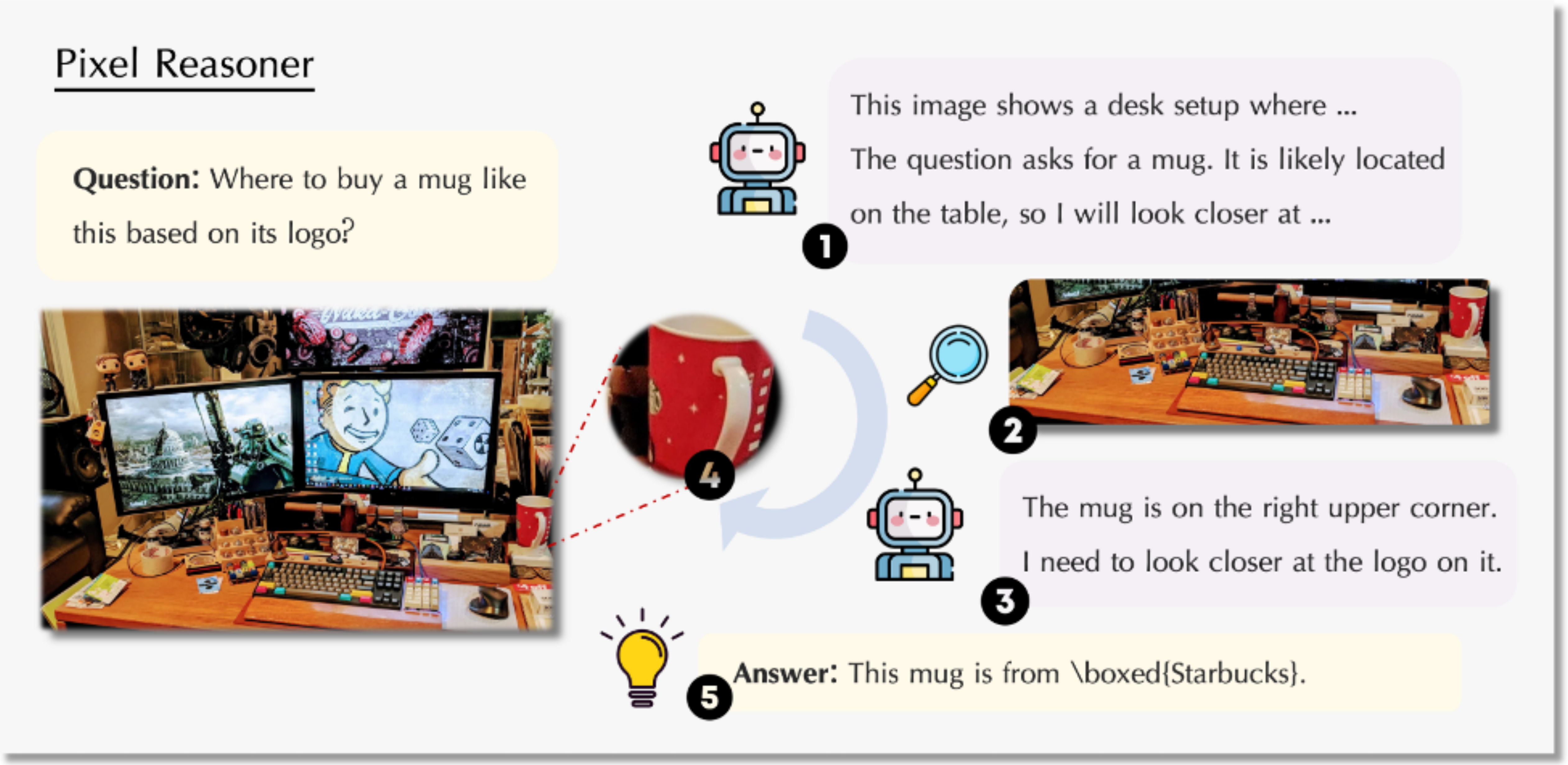


Illustration of Pixel-Space Reasoning

Credit: "DeepEyes: Incentivizing "Thinking with images" via Reinforcement Learning"

Credit: "Pixel Reasoner"

Enabling Models to Think Visually via Tool Using

S1 (Initial Exploration) → S2 (High-Frequency Tool Usage) → S3 (Efficient Exploitation)

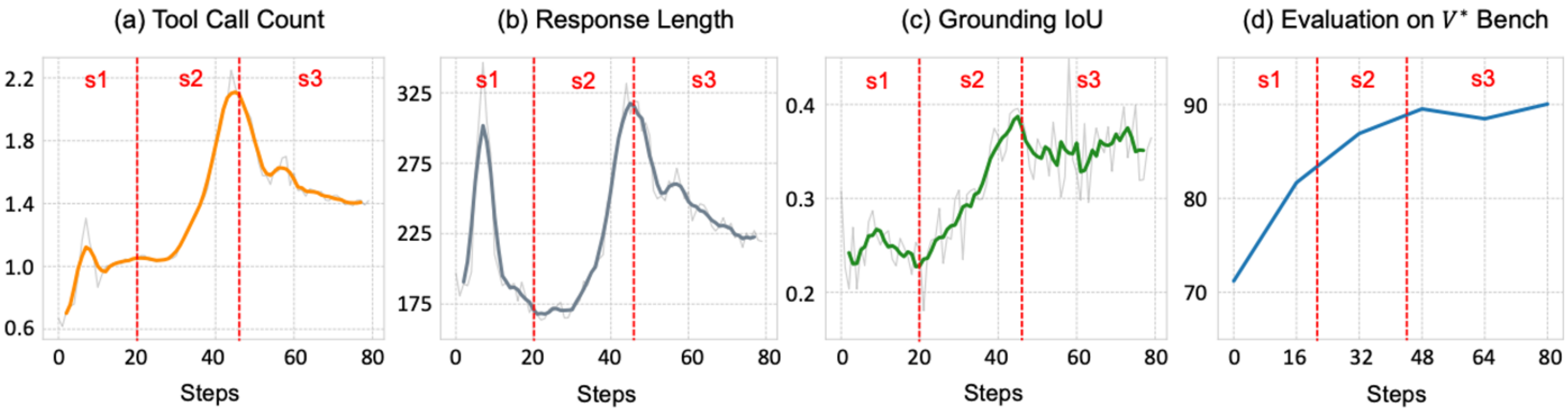


Figure 3: **Training dynamics of *DeepEyes*.** s1/2/3 represent different stages.

Credit: "DeepEyes: Incentivizing "Thinking with images" via Reinforcement Learning"

Enabling Models to Think Visually via Tool Using

S1 (Initial Exploration) → S2 (High-Frequency Tool Usage) → S3 (Efficient Exploitation)

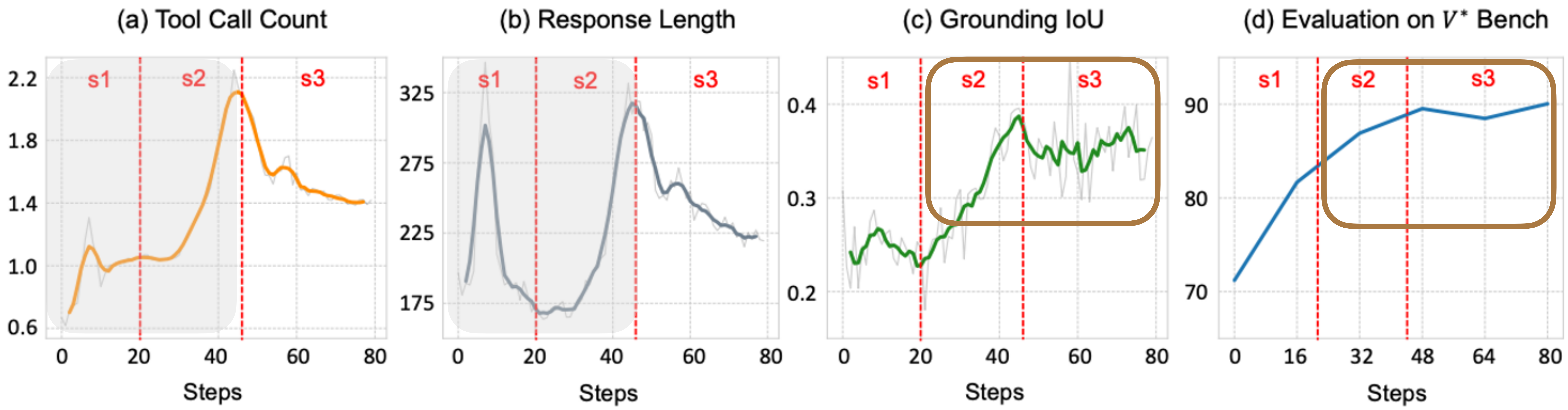


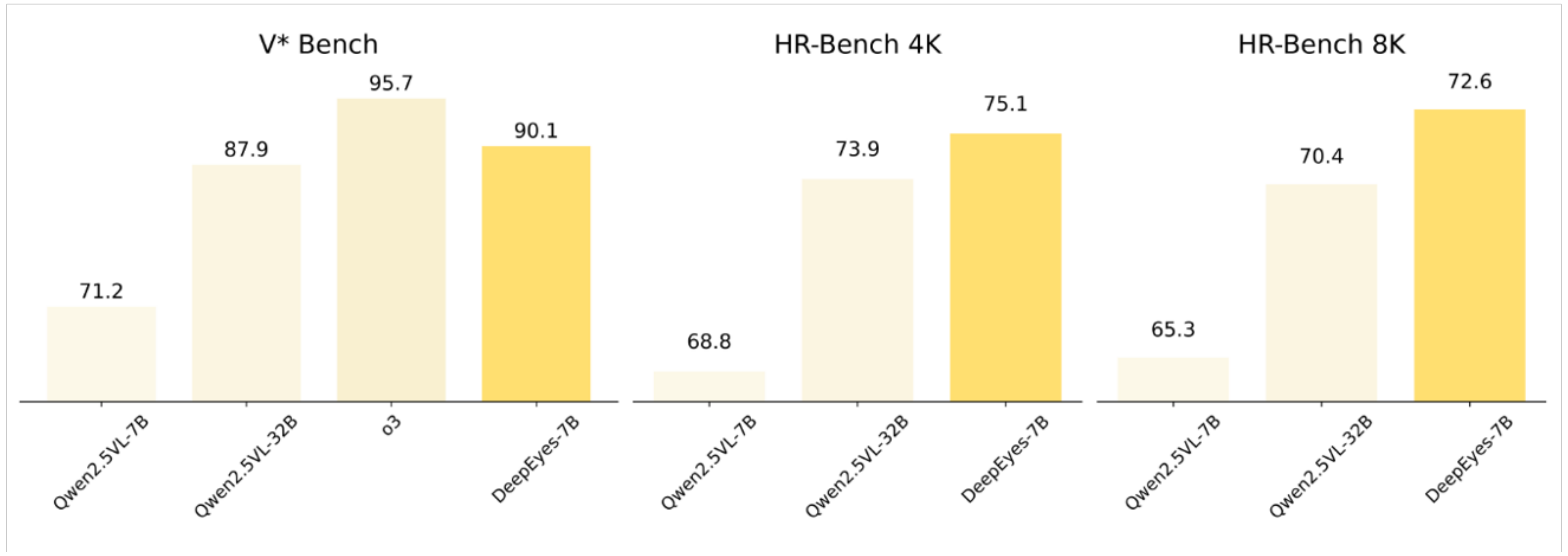
Figure 3: **Training dynamics of *DeepEyes*.** s1/2/3 represent different stages.

Model learns to deliver better results with fewer, more precise tool interactions.

Credit: “DeepEyes: Incentivizing “Thinking with images” via Reinforcement Learning”

Enabling Models to Think Visually via Tool Using

Zoom-in/Crop as fundamental image manipulation tool *=> Not only helps visual search performances but also improves on grounding, math reasoning and reduces hallucinations.*



Model	Param Size	refCOCO	refCOCO+	refCOCOg	ReasonSeg	POPE Adversarial	POPE Popular	POPE Random	POPE Overall
LLaVA-OneVision [62]	7B	-	-	-	-	-	-	-	88.4
Qwen2.5-VL [58]	7B	90.0	84.2	87.2	-	-	-	-	-
Qwen2.5-VL* [58]	7B	89.1	82.6	86.1	68.3	85.9	86.5	87.2	85.9
DeepEyes	7B	89.8	83.6	86.7	68.6	84.0	87.5	91.8	87.7
Δ (*vs* Qwen2.5-VL 7B)	-	+0.7	+1.0	+0.6	+0.3	-1.9	+1.0	+4.6	+1.8

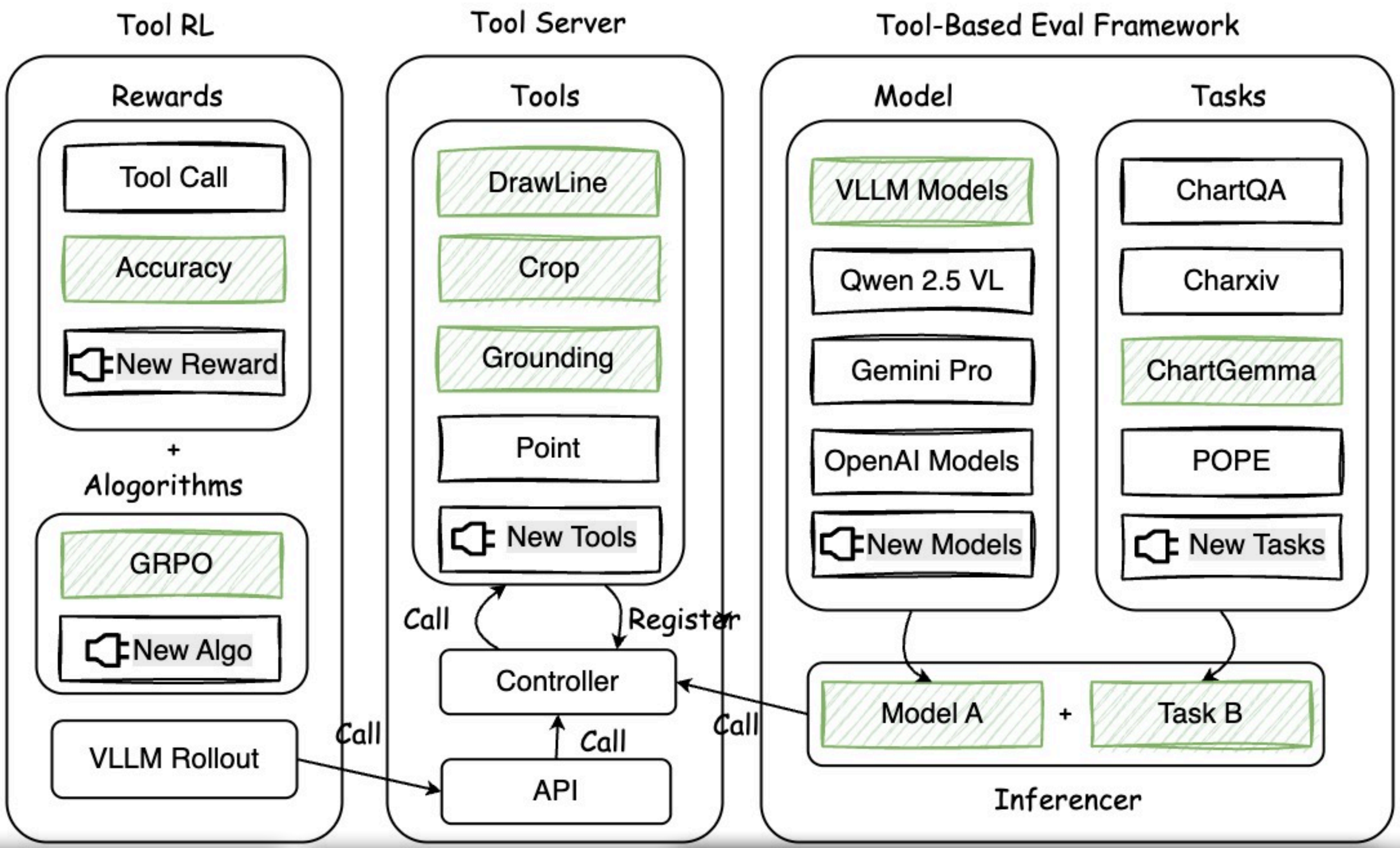
Model	Param Size	Math Vista [64]	Math Verse [65]	Math Vision [66]	We Math [67]	Dyna Math [68]	Logic Vista [69]
LLaVA-OneVision [62]	7B	58.6†	19.3†	18.3†	20.9†	-	33.3†
Qwen2.5-VL [58]	7B	68.2	49.2	25.1	35.2†	-	44.1†
Qwen2.5-VL* [58]	7B	68.3	45.6	25.6	34.6	53.3	45.9
DeepEyes	7B	70.1	47.3	26.6	38.9	55.0	47.7
Δ (*vs* Qwen2.5-VL 7B)	-	+1.9	+1.7	+1.0	+4.3	+1.7	+1.8

Credit: “DeepEyes: Incentivizing “Thinking with images” via Reinforcement Learning”

Scaling RL with Vision Tools

OpenThinkIMG

A modular RL framework for "Thinking with images" that allows easy extension by the community



- *Flexible reward and algorithm designs*

- *Provides core tool functionalities*
- *Offers a clean API for RL rollouts and inferences*

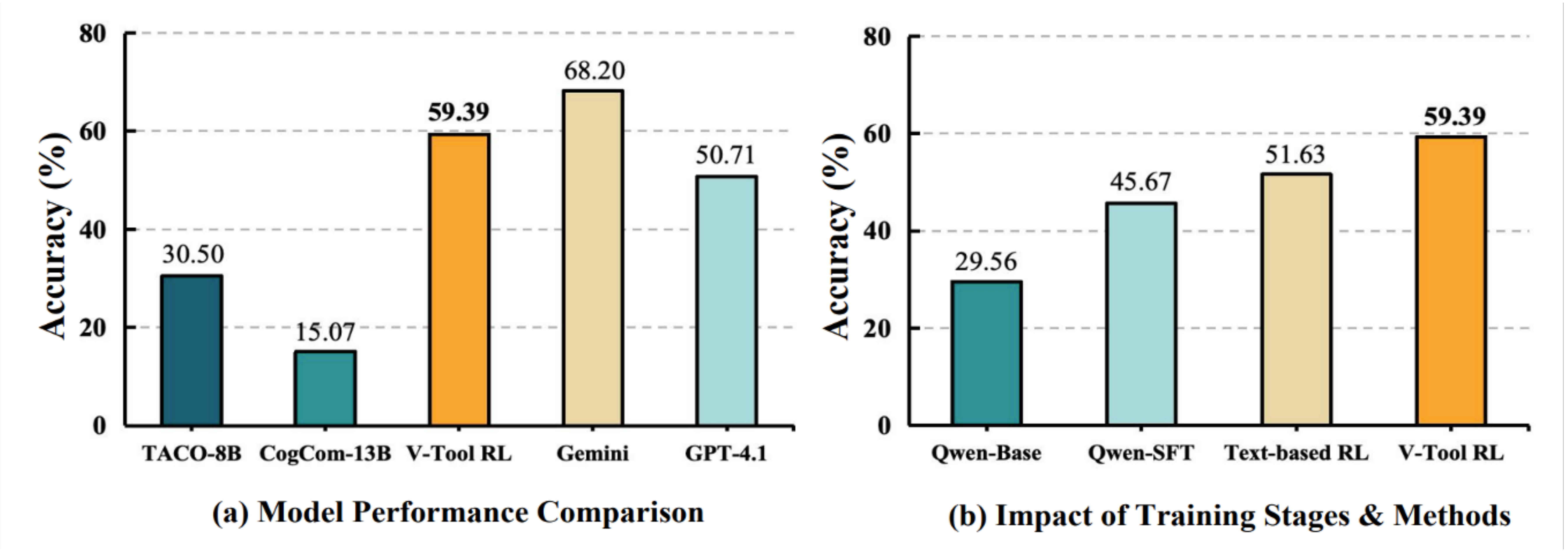
ss task, model

https://github.com/OpenThinkIMG/OpenThinkIMG

Scaling RL with Vision Tools

OpenThinkIMG

Model trained with the OpenThinkIMG framework (V-Tool RL) reaching comparable performance to closed-source models

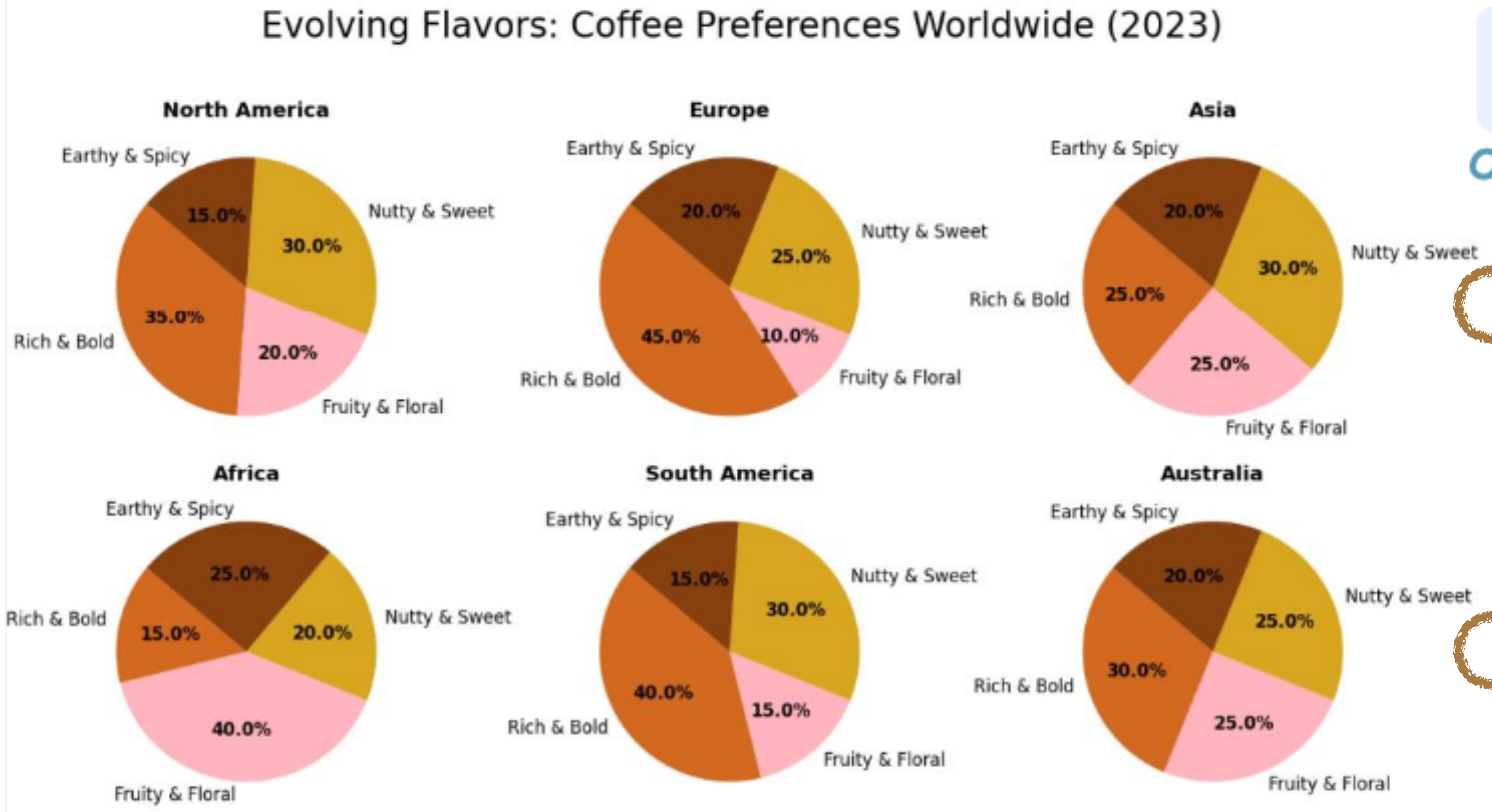


Performance Comparison on ChartGemma test set.

https://github.com/OpenThinkIMG/OpenThinkIMG

Scaling RL with Vision Tools

Case study on precise visual manipulation with more tools: ZoomIn + OCR



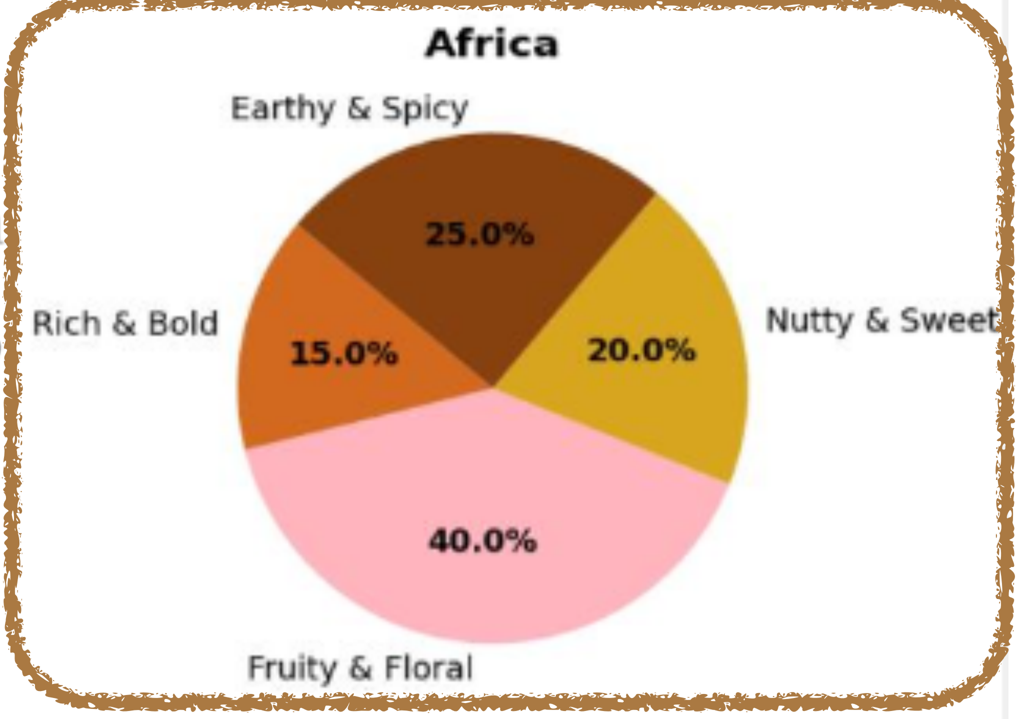
Q: What is the percentage difference between "Fruity & Floral"and "Rich & Bold" preferences in Africa?

Our Model

t_0: To determine the percentage difference between "Fruity & Floral" and "Rich & Bold" preferences in Africa, I will zoom in on the relevant section of the chart and then use OCR to extract the values.

a_0: **ZoomInSubfigure**(image-0)

o_0: ZoomInSubfigure model outputs: Select subplot done.
Please summarize the model outputs and answer my first question



t_1: Now that I have zoomed in on the relevant section of the chart, I will use OCR to extract the values for category names and their percentages.

a_1: **OCR**(image-1)

o_1: OCR model outputs:
[Earthy & Spicy: 25.0%, Nutty & Sweet: 20.0%, Fruity & Floral: 40.0%, Rich & Bold: 15.0%],
Please summarize the model outputs and answer my first question.

t_2: The OCR returns "Fruity & Floral : 25.0%" and "Rich & Bold : 40.0%." To quantify the gap, I compute 40.0% – 25.0%, yielding a 15.0% difference. Therefore, the final answer is 15.0%.

a_2: **Terminate**(15%) ✓

GPT 4.1

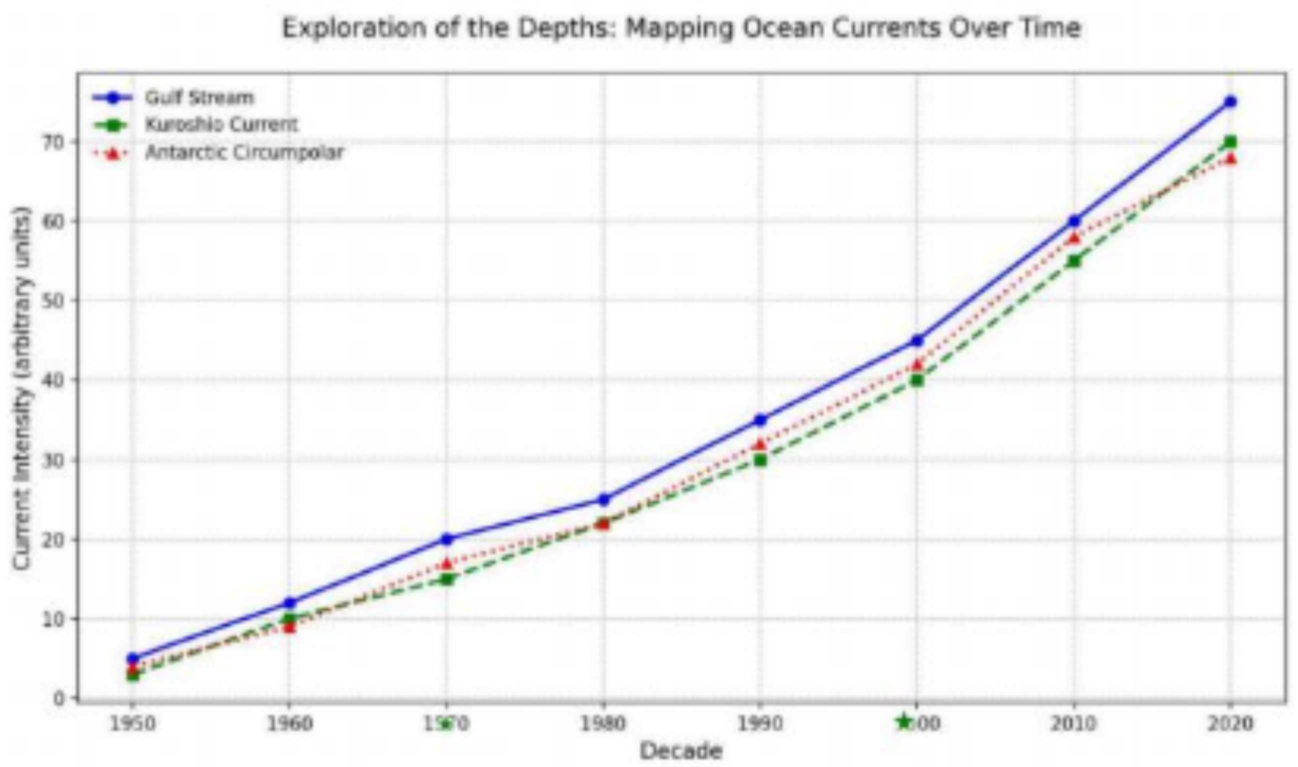
I need to extract the percentages for 'Fruity & Floral' and 'Rich & Bold' from the Africa pie chart. Then I can calculate the percentage difference, so I will visually read the values from the chart. In Africa: 'Fruity & Floral' is 40.0% and 'Rich & Bold' is 18.0%. The percentage difference is 40.0% - 18.0% = 22.0%.
Therefore, the final answer is 22.0%. ✗

https://github.com/OpenThinkIMG/OpenThinkIMG

Scaling RL with Vision Tools

OpenThinkIMG

Case study on precise visual manipulation with more tools: Point + DrawLines



Q: Between the decades 1970 and 2000, which ocean current exhibited the largest increase in intensity?

Our Model

t_0: To determine which current rose the most between 1970 and 2000, I began by pinpointing the exact x-axis positions for those two decades on the original plot and used the Point tool to mark each curve's value at 1970 and 2000.

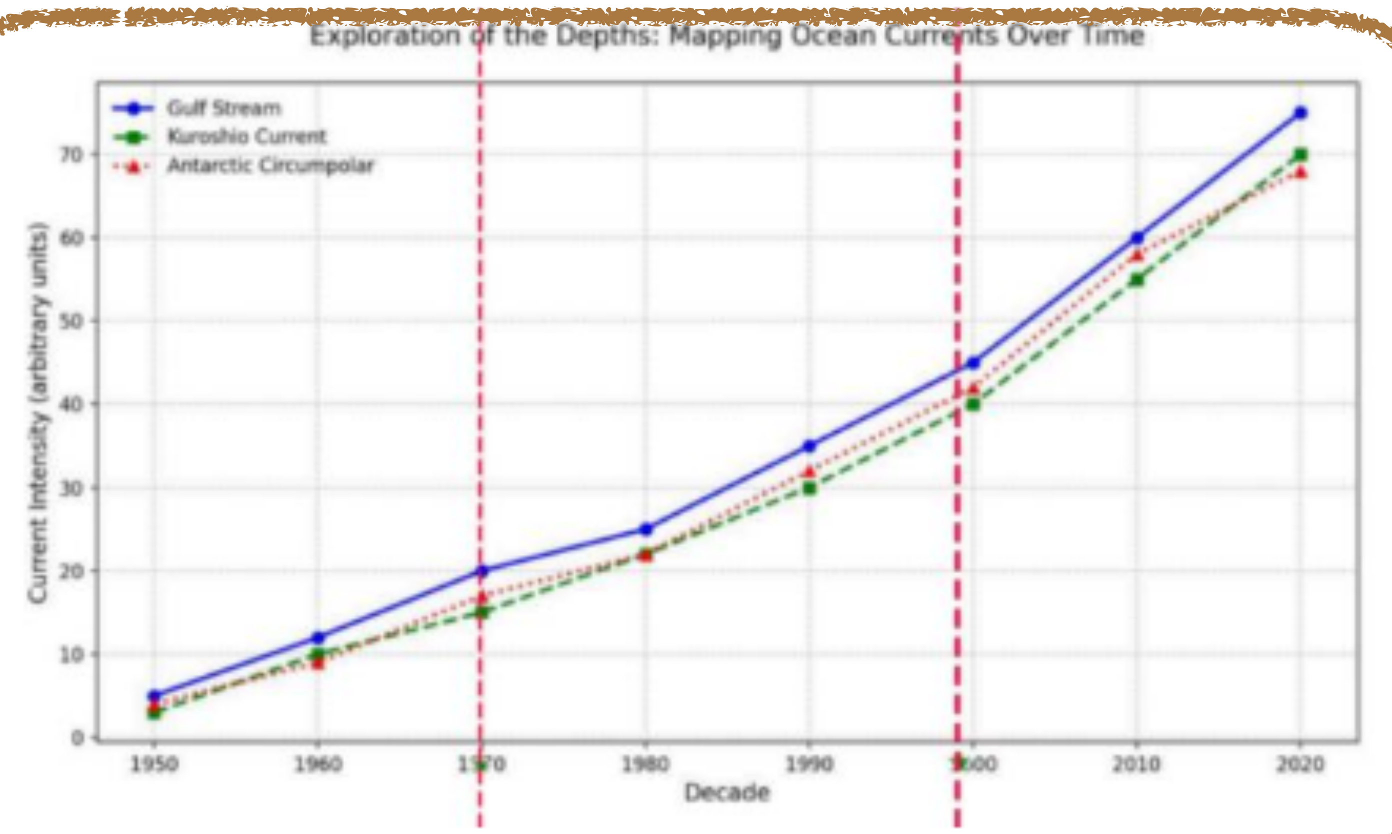
a_0: **Point**(image-0, 1970, 2000)

o_0: Point model outputs: point x1=4.3 y1=1.3 alt=1700. point x2=8.2 y2=1.3 alt=2000. Please summarize the model outputs and answer my first question.

t_1: Based on the marked-points image, I then overlaid red dashed vertical lines at x = 1970 and x = 2000 to create clear visual guides.

a_1: **DrawVerticalLineByX**(image-0, point x1=4.3 y1=1.3, point x2=8.2 y2=1.3)

o_1: DrawVerticalLineByX model outputs: Line drawn successfully. Generated image: image-1. Please summarize the model outputs and answer my first question.



t_2: From the resulting annotated plot, I read off the intensities—Gulf Stream rising from 20 to 45, Kuroshio Current from 15 to 40, and Antarctic Circumpolar from 17 to 42—and calculated that each current increased by 25 units over that period, leading to a three-way tie for the largest gain.

a_2: **Terminate**(three-way tie) ✓

GPT 4.1

Let's calculate the difference in y-coordinates for each current between 1970 and 2000. The largest vertical change (from a higher y to a lower y) corresponds to the greatest increase in intensity. Gulf Stream: 67.3 to 43.3 (difference 24.0), Kuroshio: 70.0 to 47.9 (difference 22.1), Antarctic Circumpolar: 70.4 to 47.3 (difference 23.1). The Gulf Stream exhibits the largest increase. ✗

https://github.com/OpenThinkIMG/OpenThinkIMG

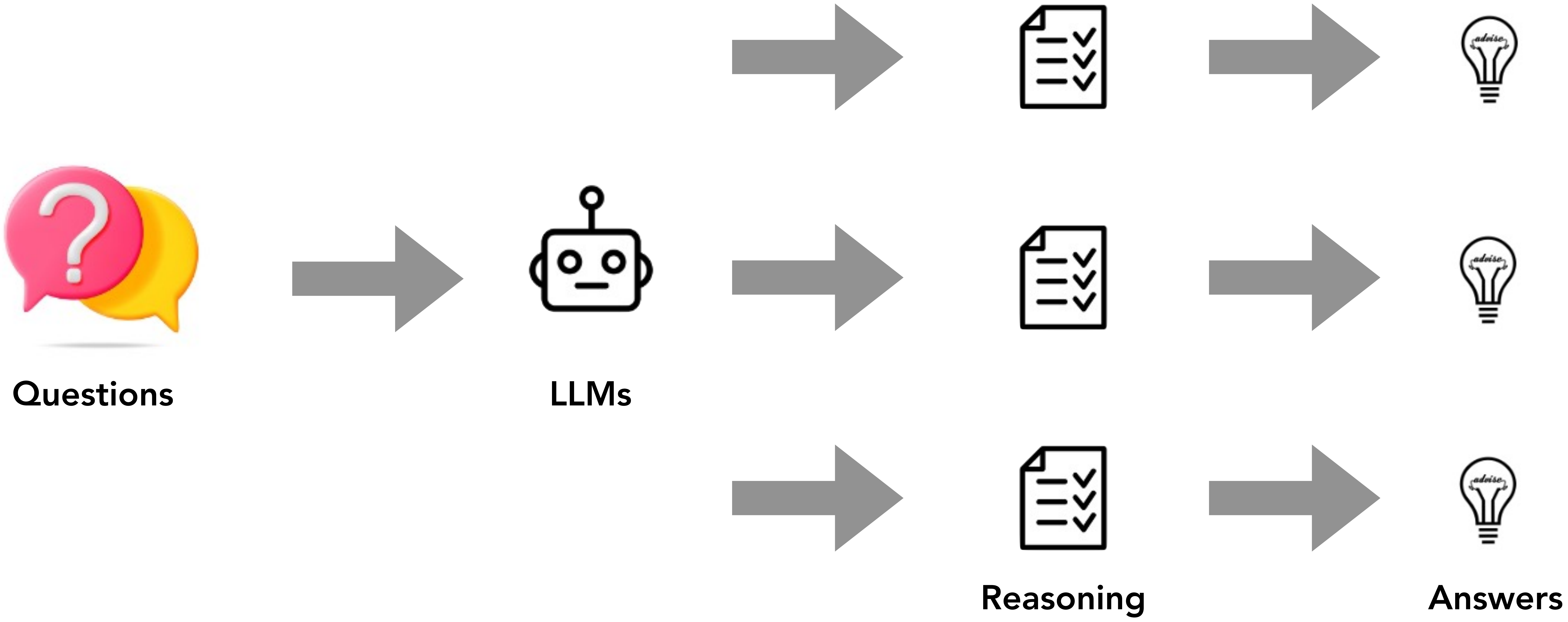
See. Visual Think. Act.

Training Multimodal Agents with Reinforcement Learning

"See". Think. Act.

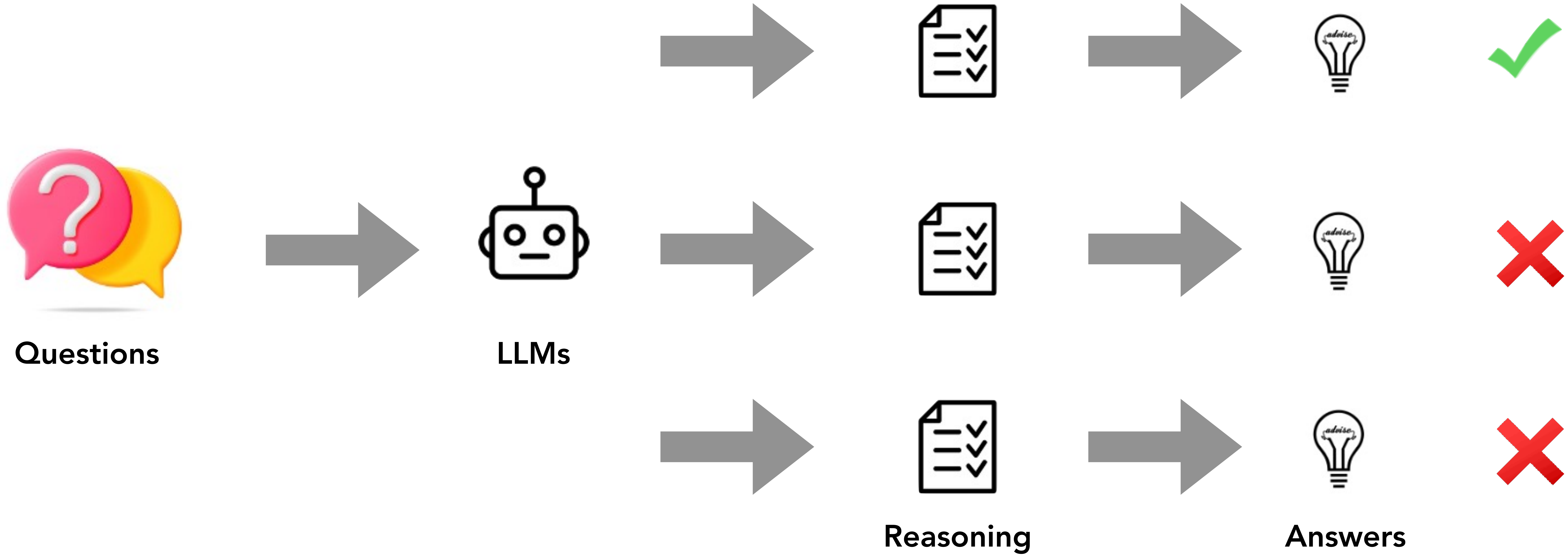
Training Language Agents with Reinforcement Learning

Revisiting RL with Verifiable Reward for LLMs



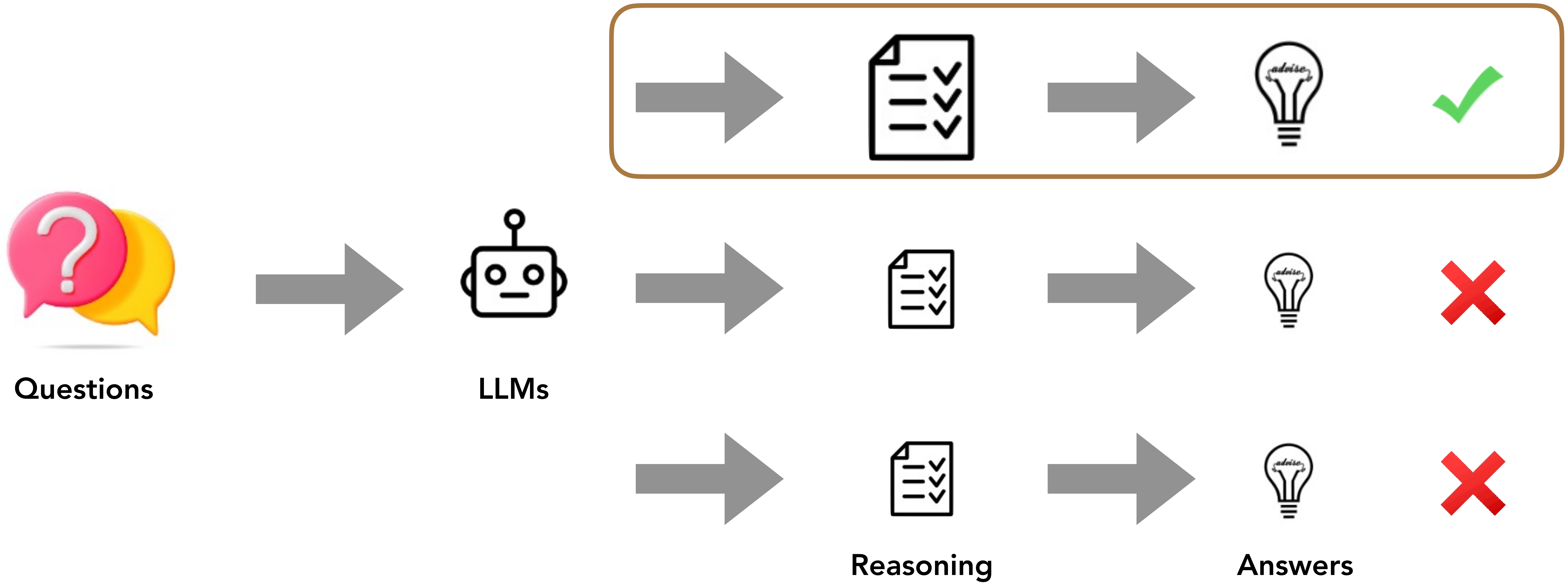
Step 1 - Rollout

Revisiting RL with Verifiable Reward for LLMs



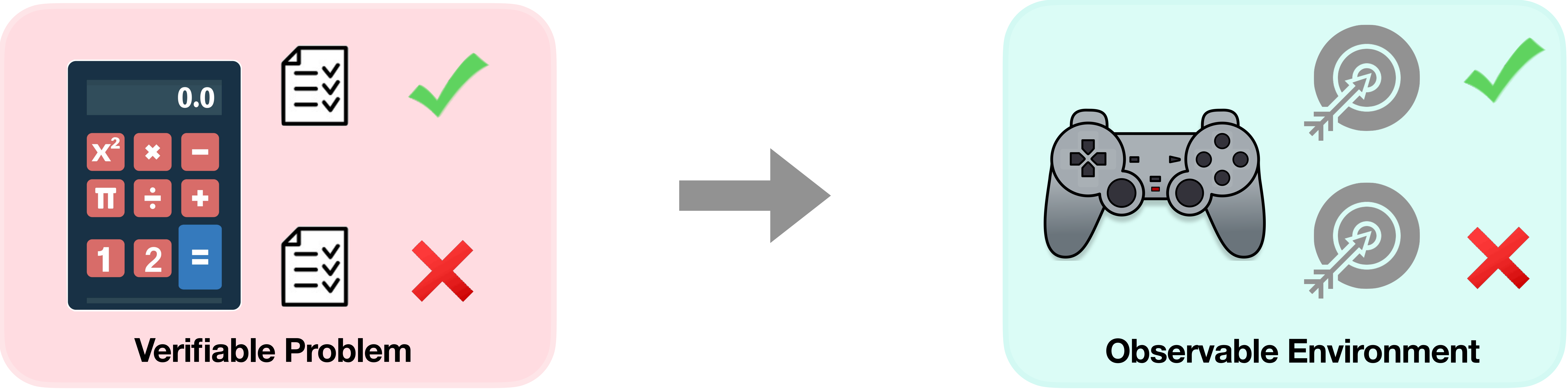
Step 2 - Verification

Revisiting RL with Verifiable Reward for LLMs



Step 3 - Reinforce Reasoning

Extending to More Real-World Setting



Verifiable Problem	Observable Environment
Single-turn	Multi-turn feedback
Finite problem set	“Infinite” state combos
One answer	Many trajectories
Full observability	Partial observability

Key Challenges of RL in Observable Environment

Multiturn Interaction

Statefulness

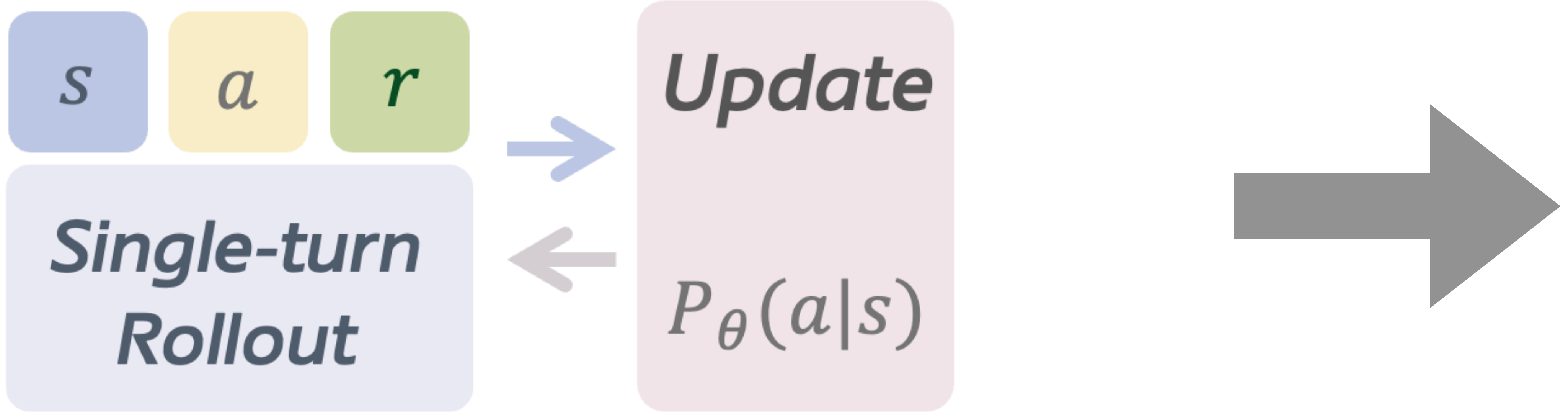
Markov Decision Processes!

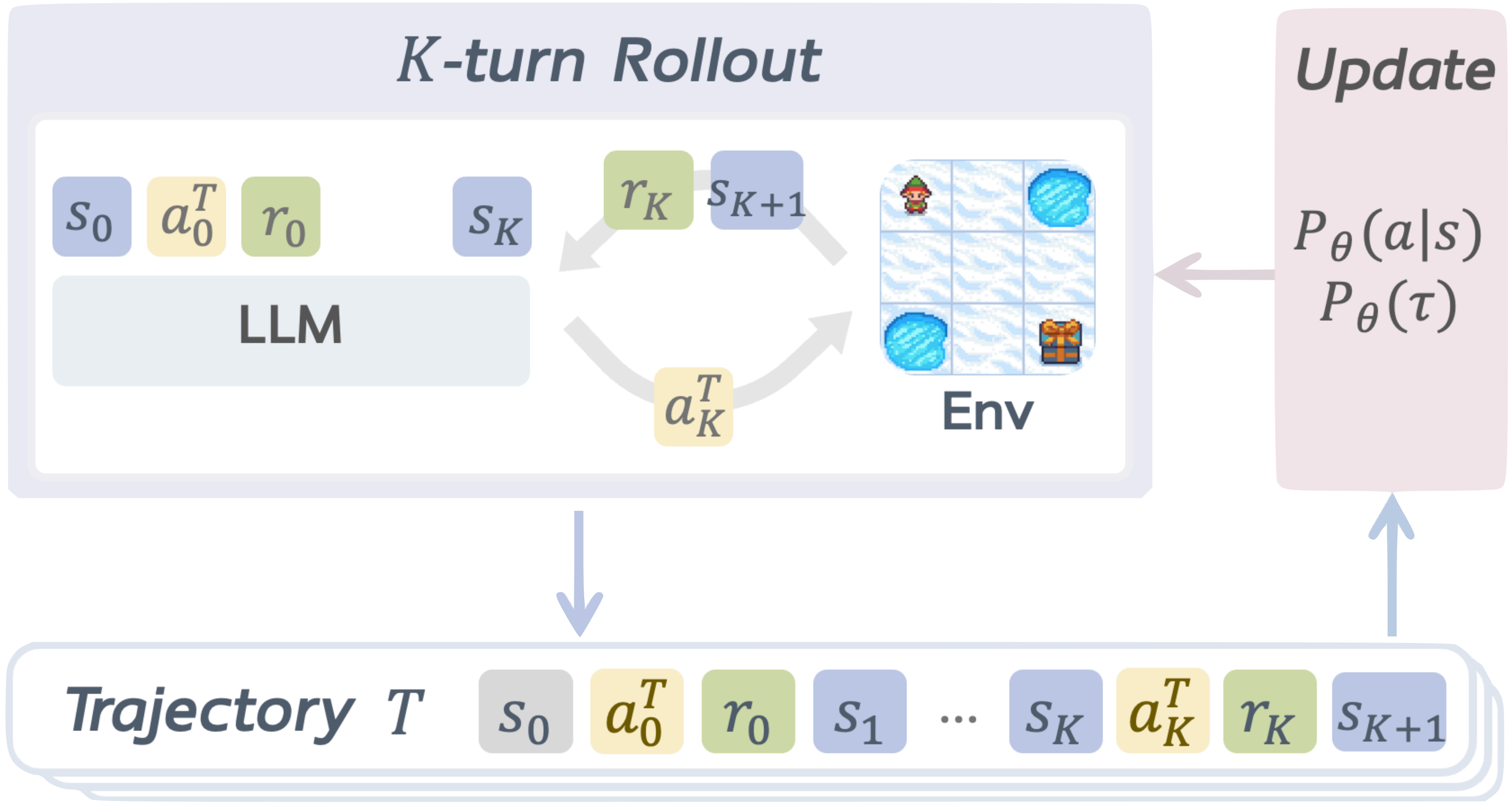
StarPO: State-Thinking-Action-Reward Policy Optimization

MDP as a sequence prediction.

s_0	a_0^T	r_0	s_1	...	s_K	a_K^T	r_K	s_{K+1}
Initial State	Action	Reward	State		State	Action	Reward	State

Reinforcing the entire multi-turn interaction trajectory





https://ragen-ai.github.io

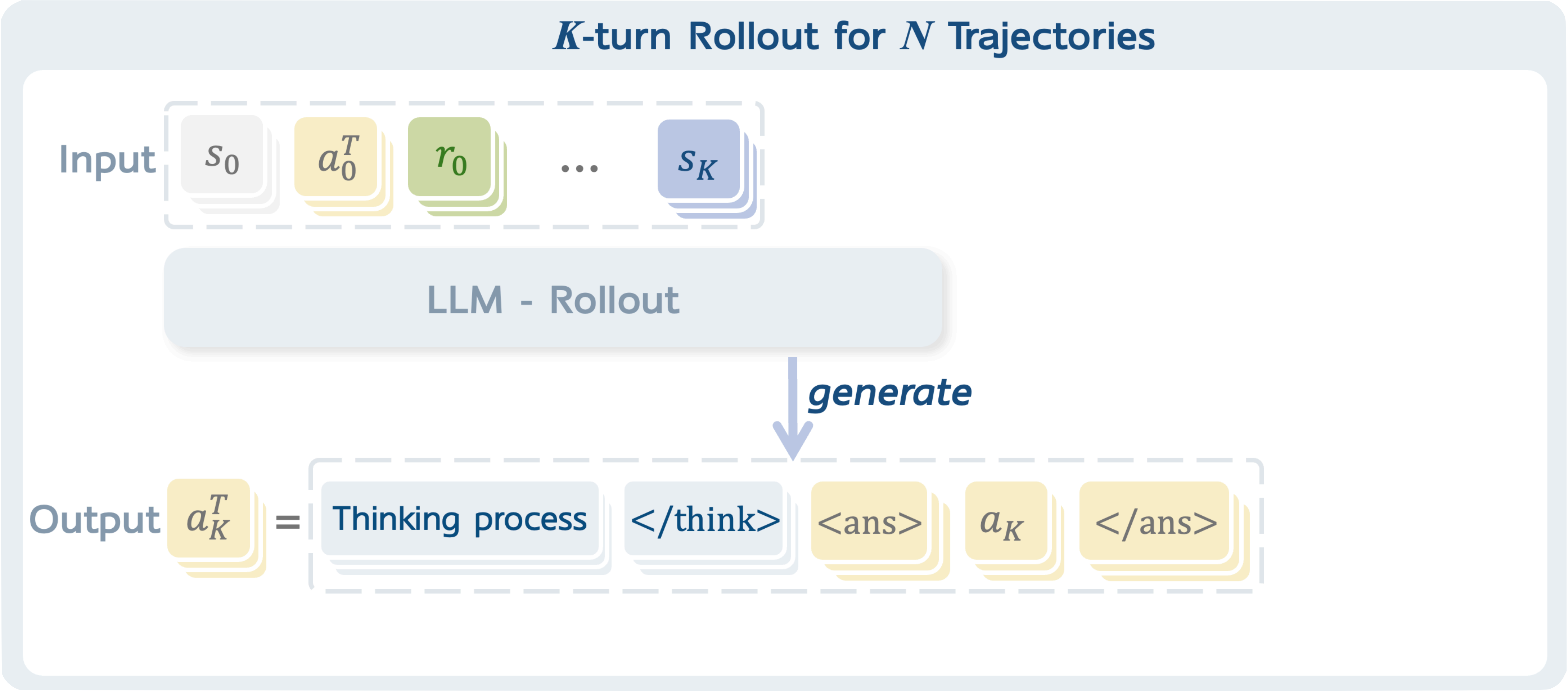
StarPO: State-Thinking-Action-Reward Policy Optimization

At each turn, the model takes in the trajectory history staring from the initial state to the current state.

Step 1 - Trajectory Rollout

https://ragen-ai.github.io

StarPO: State-Thinking-Action-Reward Policy Optimization

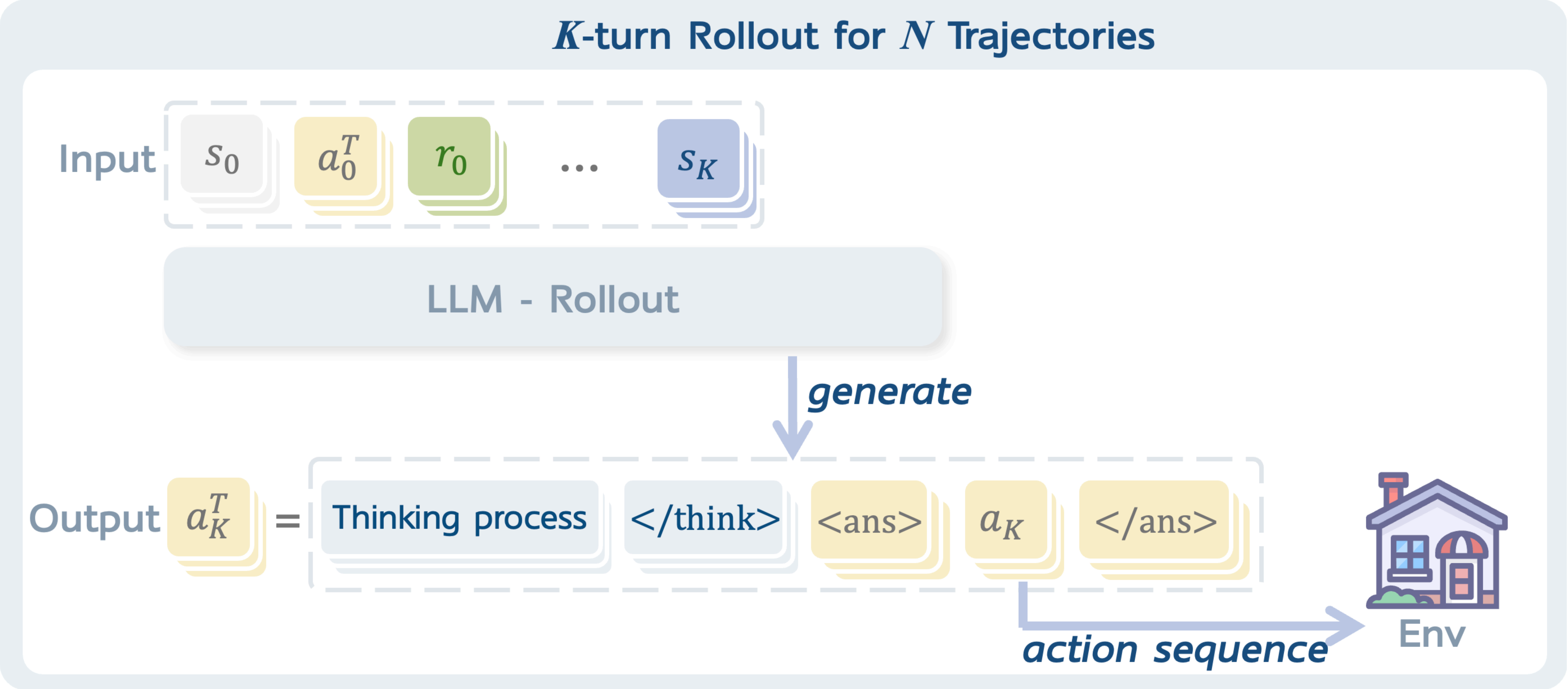


At each turn, the model generates a structured output containing reasoning and action(s)

Step 1 - Trajectory Rollout

https://ragen-ai.github.io

StarPO: State-Thinking-Action-Reward Policy Optimization

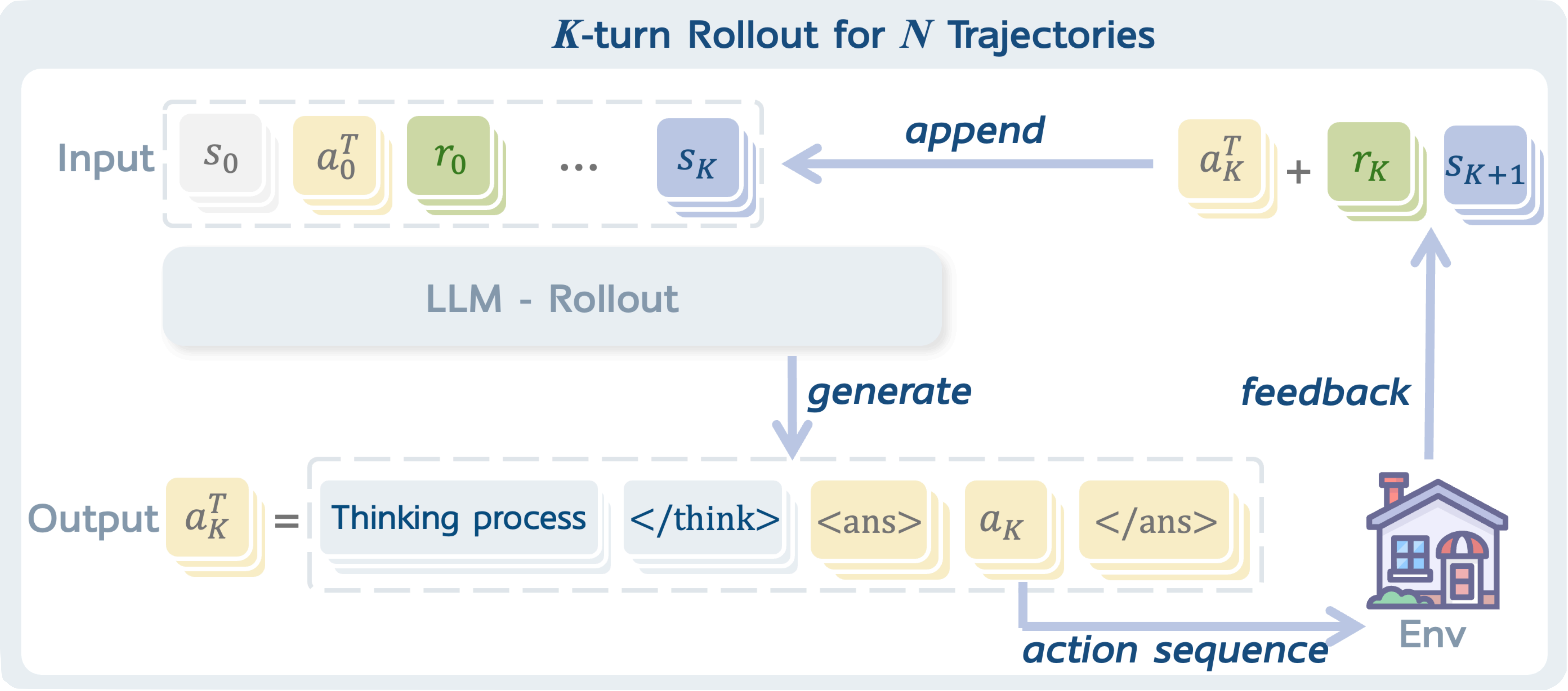


At each turn, action sequences are sent to the environment to be excused step-by-step.

Step 1 - Trajectory Rollout

https://ragen-ai.github.io

StarPO: State-Thinking-Action-Reward Policy Optimization

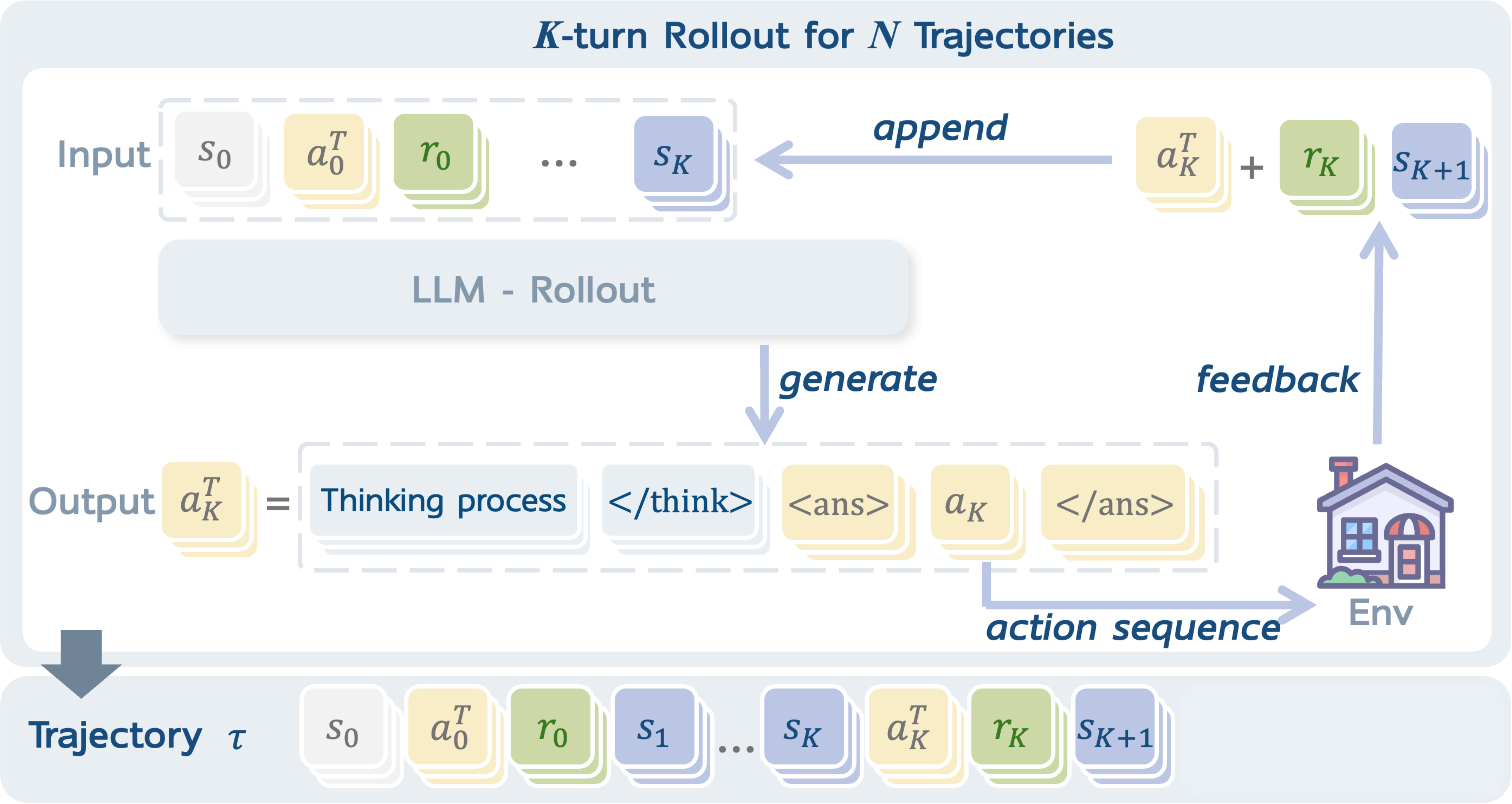


Collect the turn-level reward and the new state to append to the input sequence.

Step 1 - Trajectory Rollout

https://ragen-ai.github.io

StarPO: State-Thinking-Action-Reward Policy Optimization

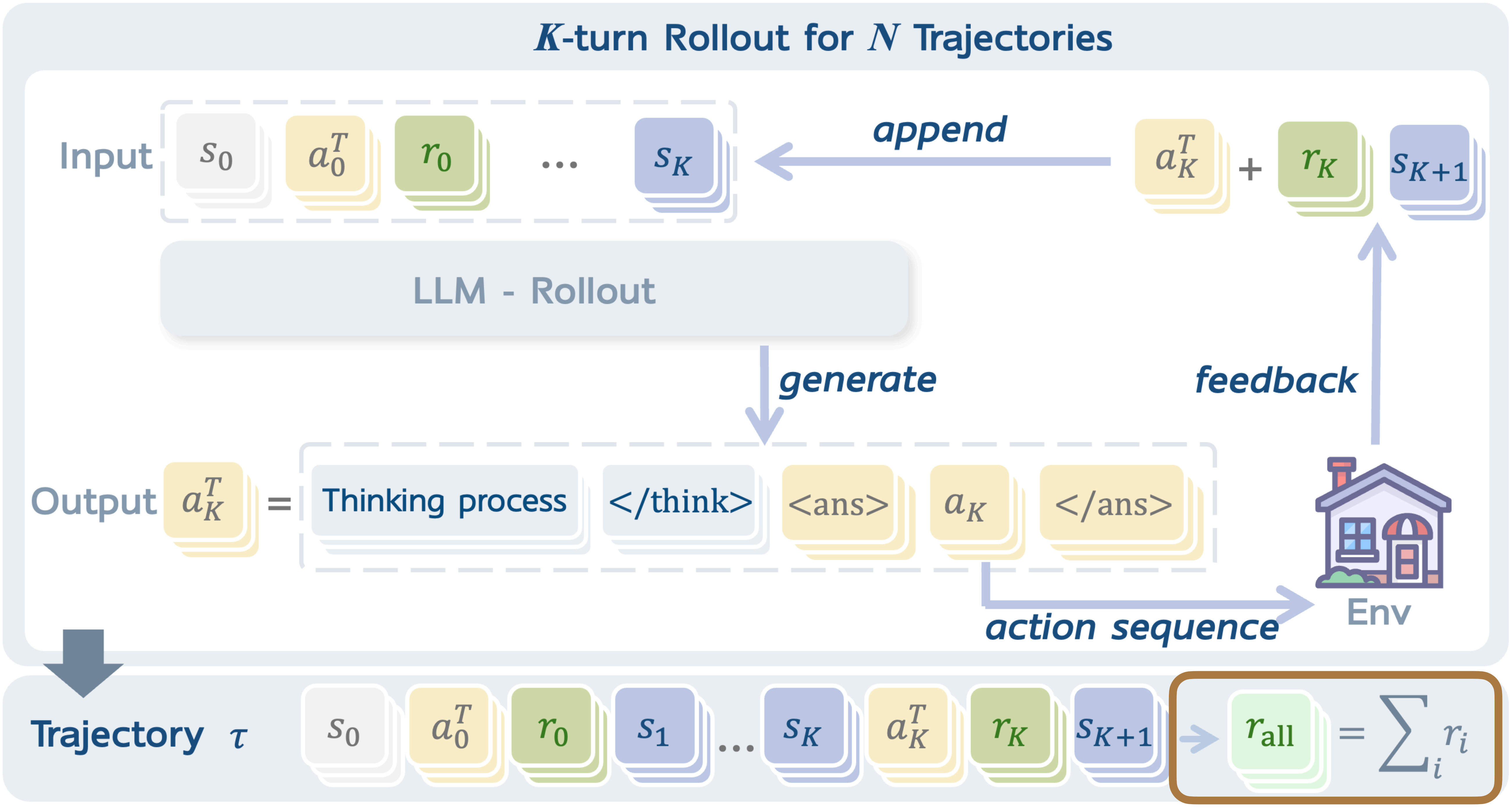


Repeat for K turns to collect N Trajectories.

Step 1 - Trajectory Rollout

https://ragen-ai.github.io

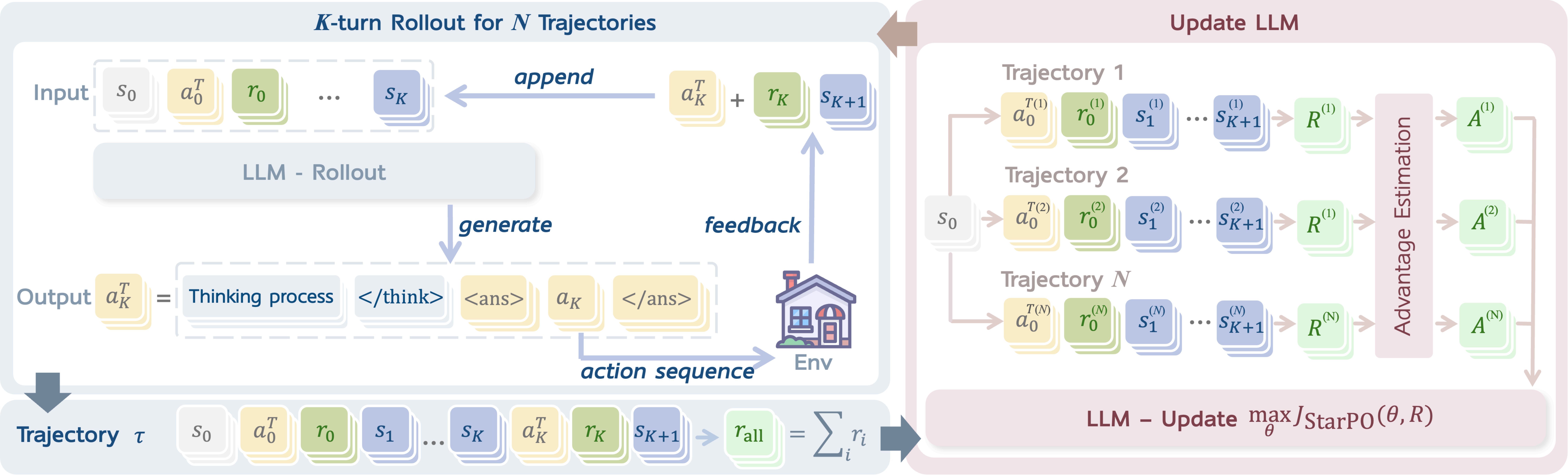
StarPO: State-Thinking-Action-Reward Policy Optimization



Compute trajectory-level rewards.

Step 2 - Trajectory Verification

StarPO: State-Thinking-Action-Reward Policy Optimization

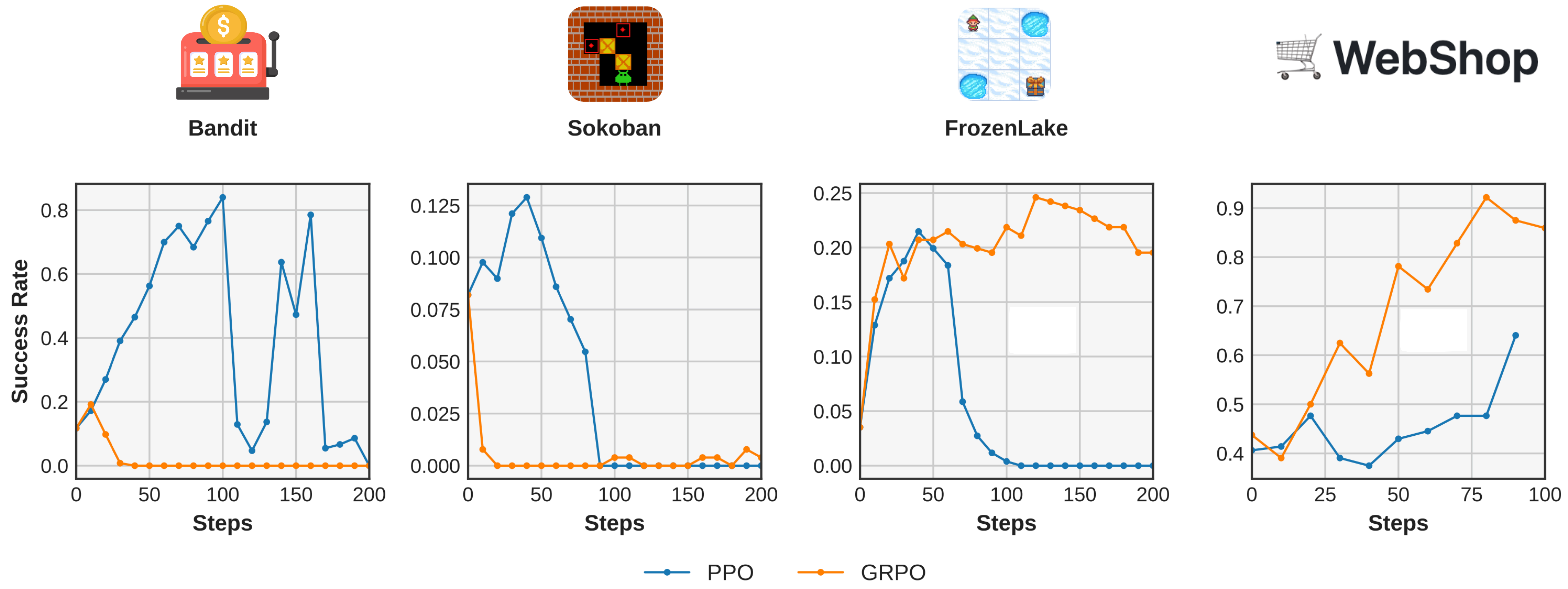


Step 3 - Reinforce Multi-turn Trajectory

https://ragen-ai.github.io

RL in Observable Environment is Challenging

Single-turn RL may not be directly adaptable



https://ragen-ai.github.io

RL in Observable Environment is Challenging

RAGEN

"Echo Trap" - Agents overfit to locally rewarded reasoning patterns, suppressing exploration

Step	Reasoning traces on Bandit
0	Dragon is often associated with patience, strength, and reliability, while Phoenix is associated with speed, agility, and creativity.
0	The symbolic meaning of Dragon and Phoenix's names suggests that they may have different reward distributions... Therefore, I think Phoenix is more likely to give higher rewards on average.
150	My initial intention is to pick arm Dragon. My first action is to go for arm Dragon.
150	My initial intention is to pick arm Dragon, so my first action is to go for arm Dragon.
150	My initial intention is to choose Arm Dragon. After analyzing both arms, my intention was to choose Arm Dragon.

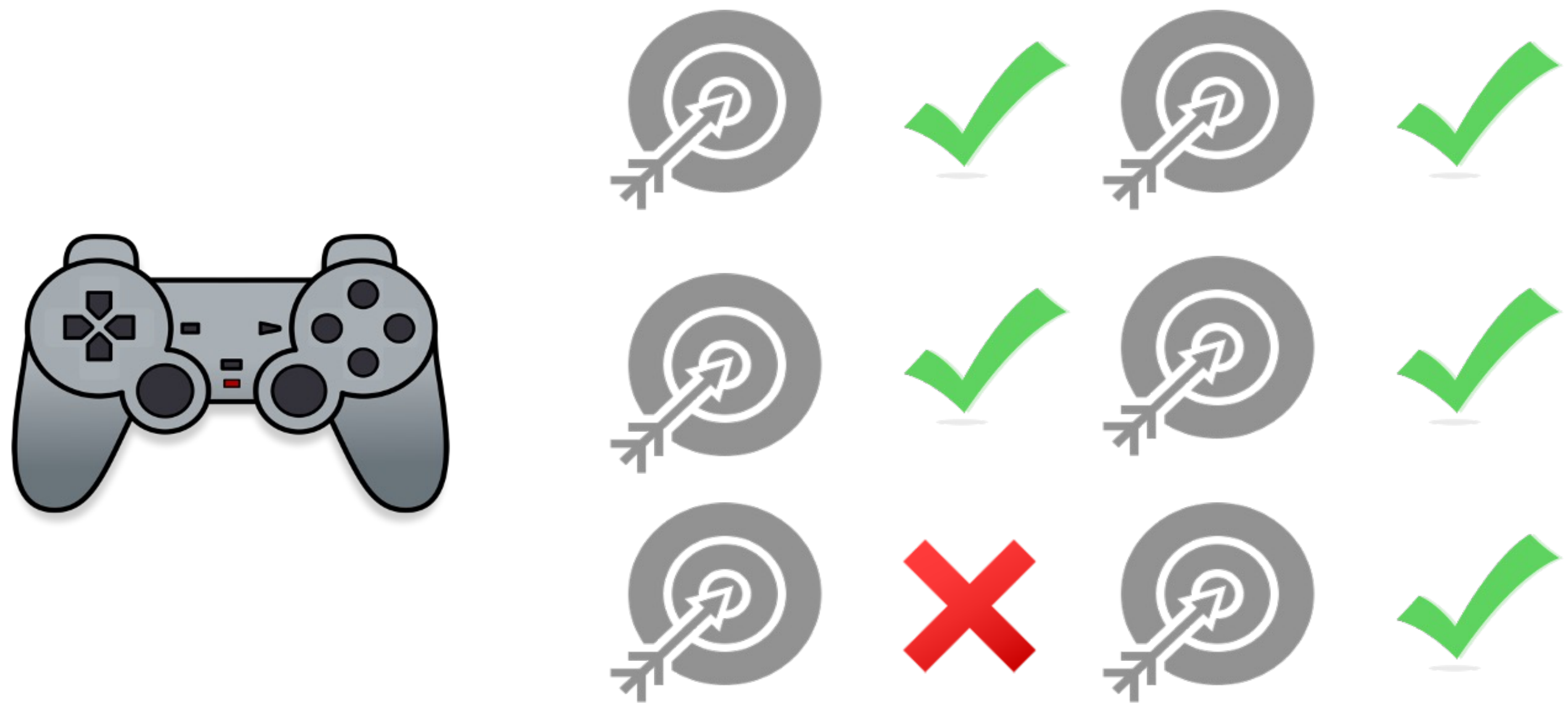
Reward Variances

Gradient Norm Spikes

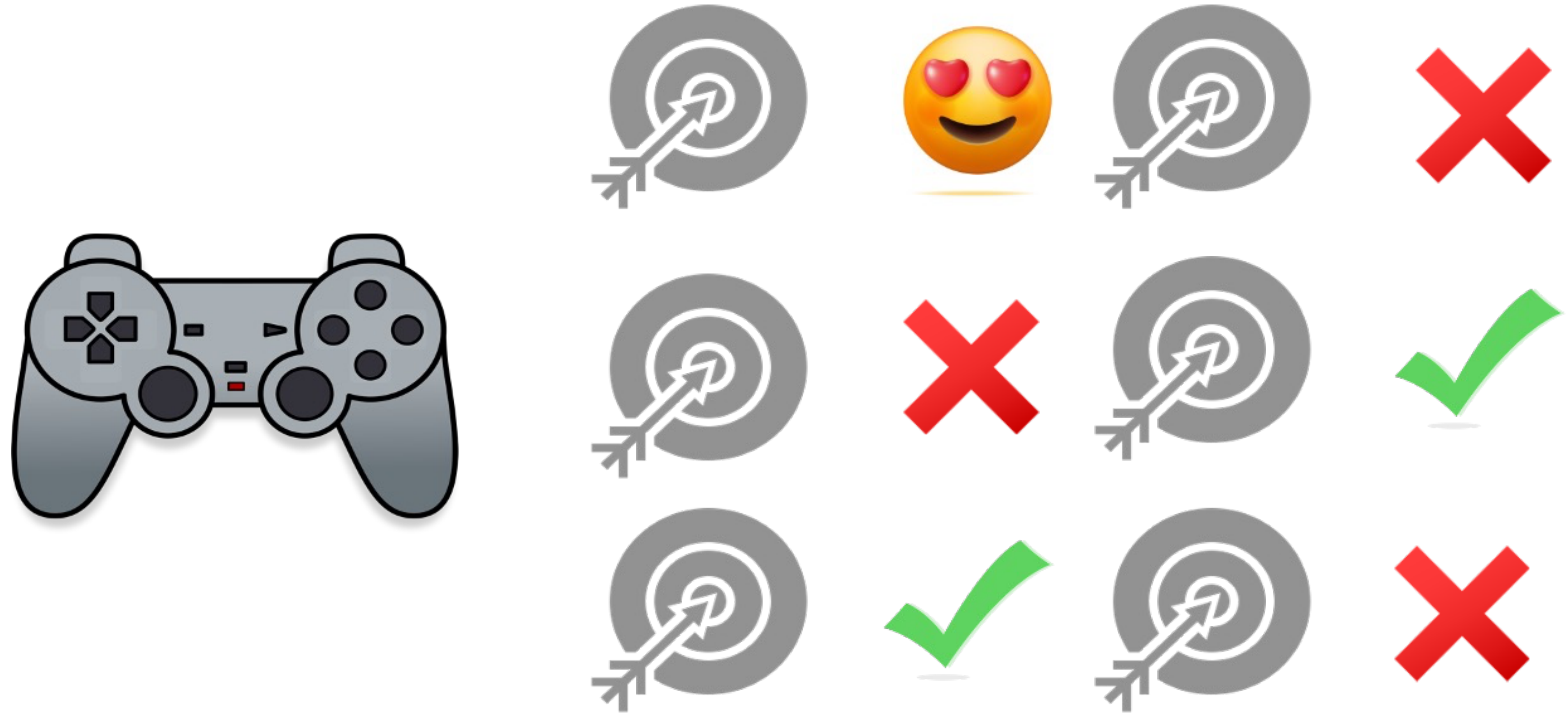
Output Entropy

How to Avoid “Echo Trap”?

Using reward variance as a proxy to measure reasoning diversity



Low reward variance

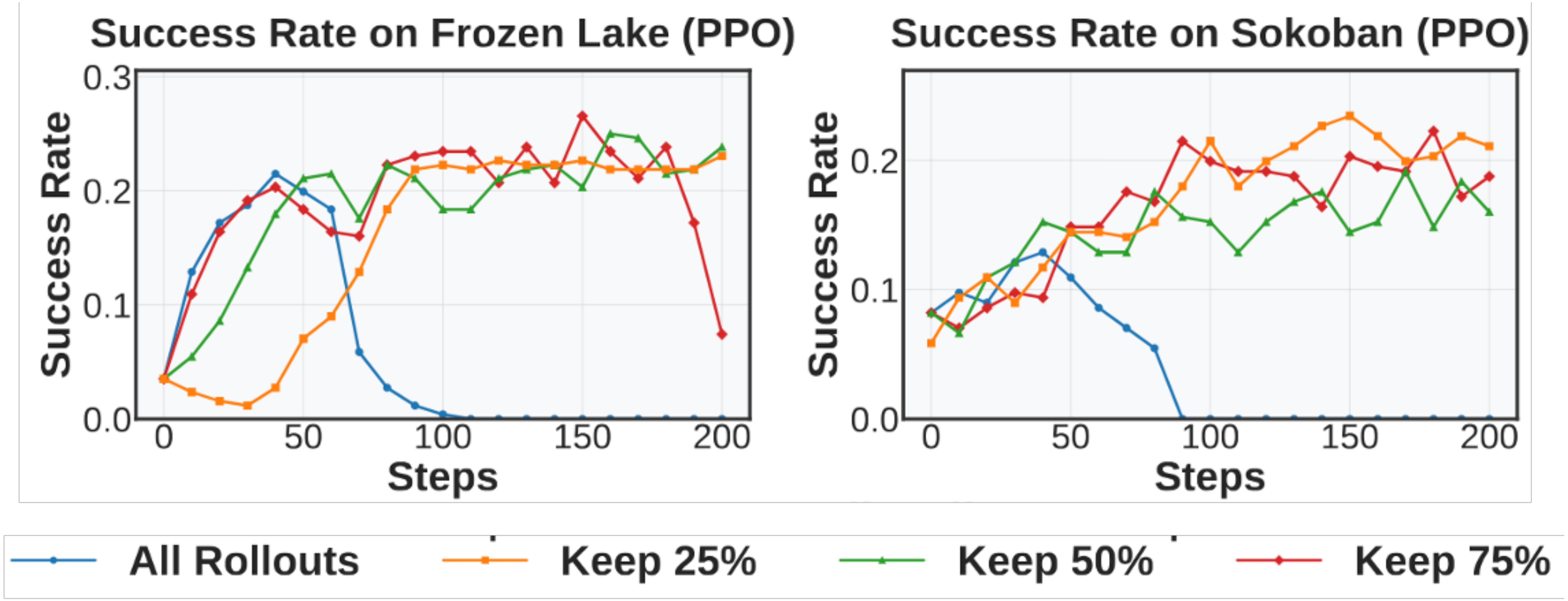


High reward variance

https://ragen-ai.github.io

How to Avoid "Echo Trap"?

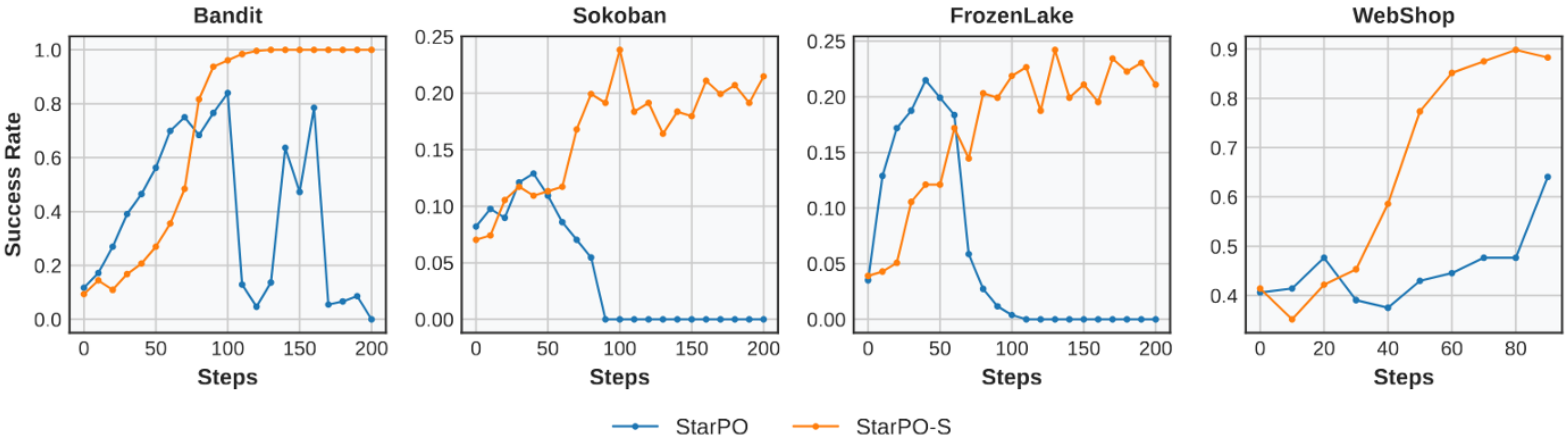
Model learns better from fewer but more diverse trajectories.



https://ragen-ai.github.io

StarPO-S: Stabilizing multi-turn RL training with LLM Agents

StarPO-s = StarPO + Filter by reward variance + Clipping + Removing KL constraint



https://ragen-ai.github.io

See. Think. Act.

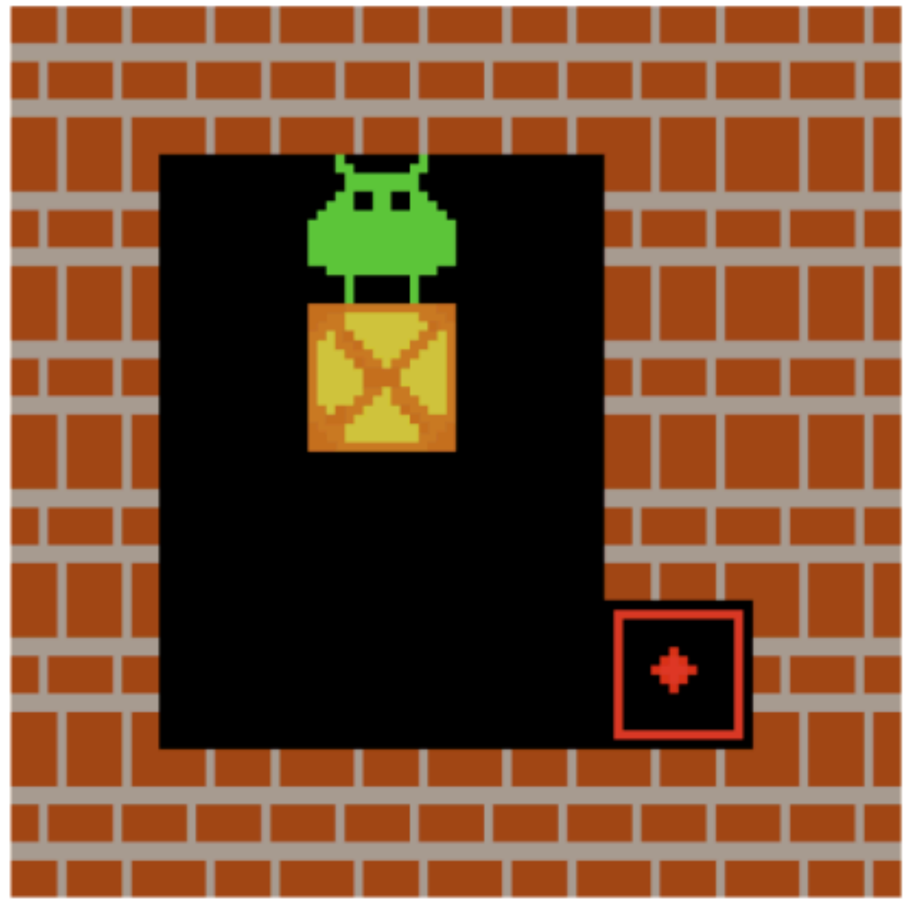
Training Multimodal Agents with Reinforcement Learning

State Representation for LLM and VLM Agents

VAGEN

Symbolic representation make it easy for LLM agents while VLM agents must first solve vision just to play.

```
# # # # # #
# _ P _ # #
# _ X _ # #
# _ _ _ # #
# _ _ _ O #
# # # # # #
```



Sokoban

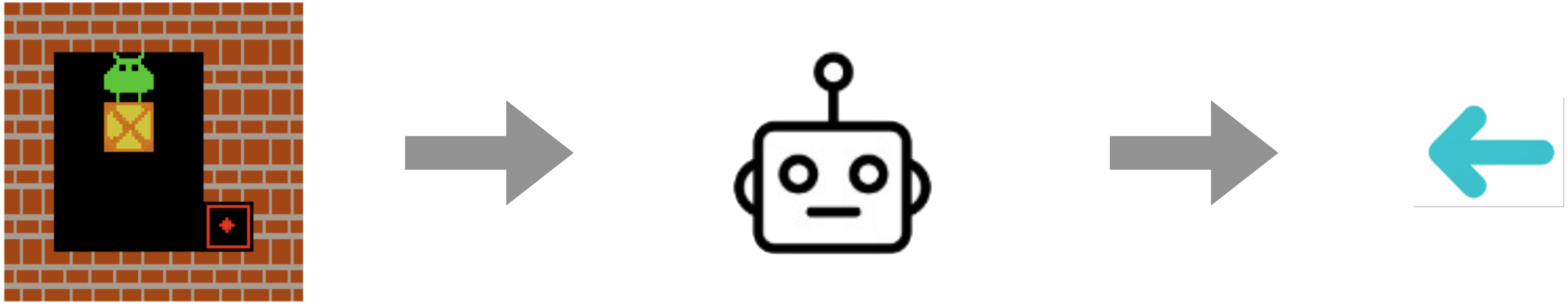
```
_ _ O O
_ _ _ P
G O O _
_ _ O _
```

FrozenLake

https://github.com/RAGEN-AI/VAGEN

How should VLMs reason about visual states?

VAGEN



Visual State Input

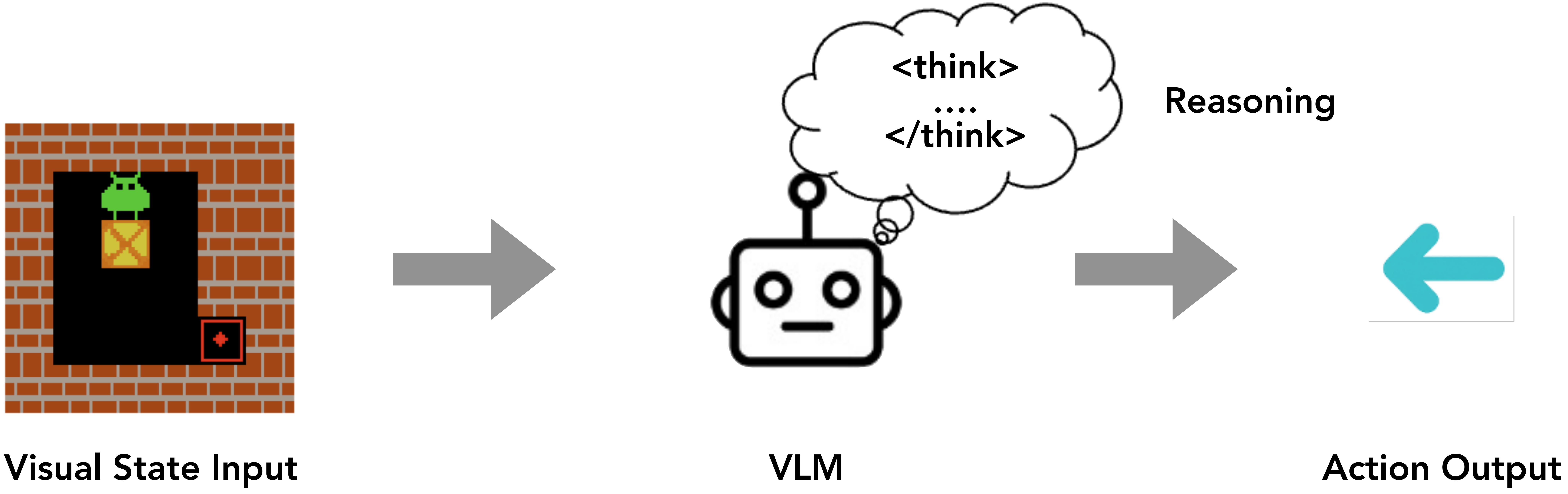
VLM

Action Output

https://github.com/RAGEN-AI/VAGEN

How should VLMs reason about visual states?

VAGEN



Option 1 - Free-Think

<think>The box looks like it needs to go over there. Maybe push it?</think>

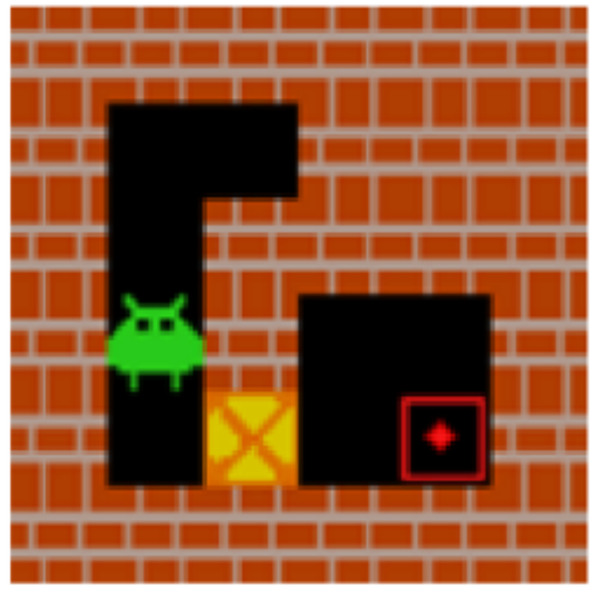
Option 2 - Explicit State Verbalization

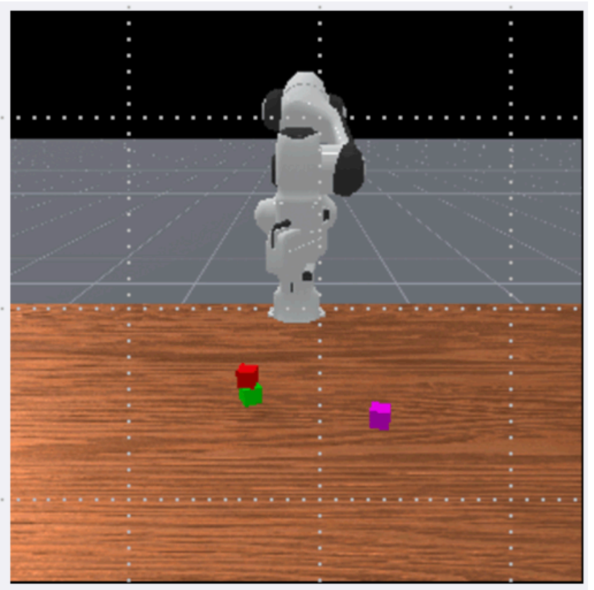
<think><observation>...</observation>...<prediction>...</prediction></think>

https://github.com/RAGEN-AI/VAGEN

How should VLMs reason about visual states?

VAGEN



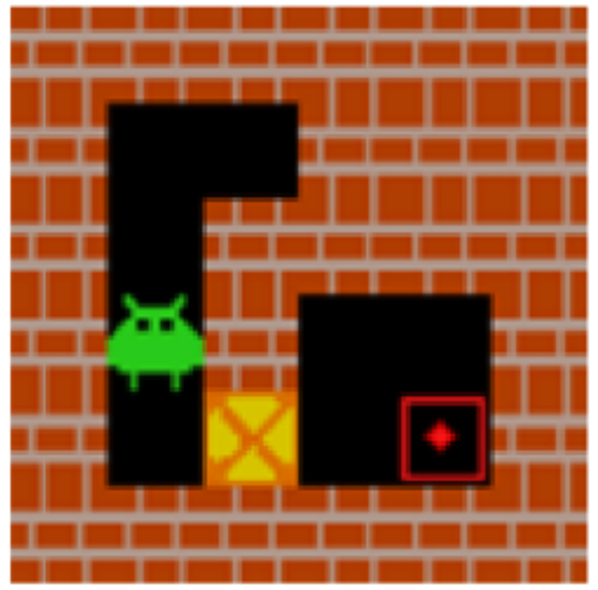


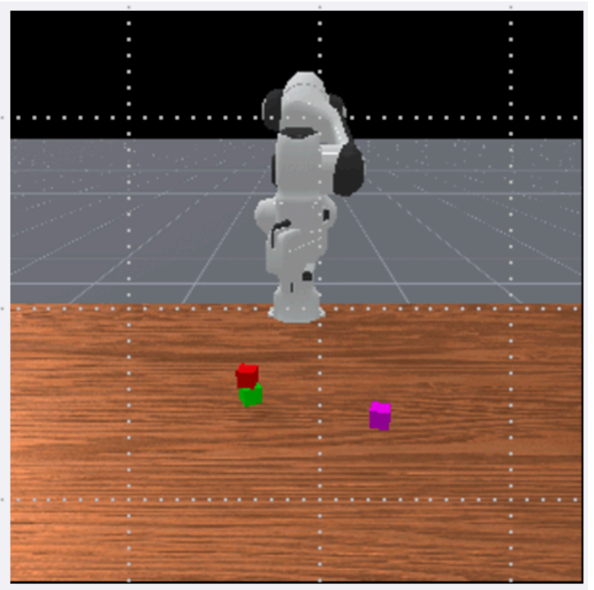
Model/Method	FrozenLake	Sokoban	Navigation			PrimitiveSkill					SVG			Overall
			Base	Common	**Average**	Place	Stack	Drawer	Align	**Average**	Dino	DreamSim	**Average**	
VAGEN: Multi-Turn RL with Visual State Reasoning (Backbone: Qwen2.5-VL-3B)														
`Free-Think`	0.39	0.43	0.63	0.63	0.63	1.00	0.63	0.00	1.00	0.66	0.90	0.64	0.77	0.58
`No-Think`	0.34	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.60	0.74	0.27
`Grounding`	0.35	0.15	0.78	0.75	0.77	0.00	0.00	0.00	0.00	0.00	0.92	0.67	**0.80**	0.41
`WorldModeling`	0.53	0.44	0.67	0.59	0.63	1.00	0.63	0.88	1.00	**0.88**	0.89	0.63	0.76	0.65
`Grounding-WorldModeling`	**0.55**	**0.44**	0.78	0.80	**0.79**	0.63	0.63	0.88	1.00	0.79	0.90	0.65	0.78	**0.67**

https://github.com/RAGEN-AI/VAGEN

How should VLMs reason about visual states?

VAGEN





Model/Method	FrozenLake	Sokoban	Navigation			PrimitiveSkill					SVG			Overall
			Base	Common	**Average**	Place	Stack	Drawer	Align	**Average**	Dino	DreamSim	**Average**	
VAGEN: Multi-Turn RL with Visual State Reasoning (Backbone: Qwen2.5-VL-3B)														
`Free-Think`	0.39	0.43	0.63	0.63	0.63	1.00	0.63	0.00	1.00	0.66	0.90	0.64	0.77	0.58
`No-Think`	0.34	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.60	0.74	0.27
`Grounding`	0.35	0.15	0.78	0.75	0.77	0.00	0.00	0.00	0.00	0.00	0.92	0.67	**0.80**	0.41
`WorldModeling`	0.53	0.44	0.67	0.59	0.63	1.00	0.63	0.88	1.00	**0.88**	0.89	0.63	0.76	0.65
`Grounding-WorldModeling`	**0.55**	**0.44**	0.78	0.80	**0.79**	0.63	0.63	0.88	1.00	0.79	0.90	0.65	0.78	**0.67**

Free-Think

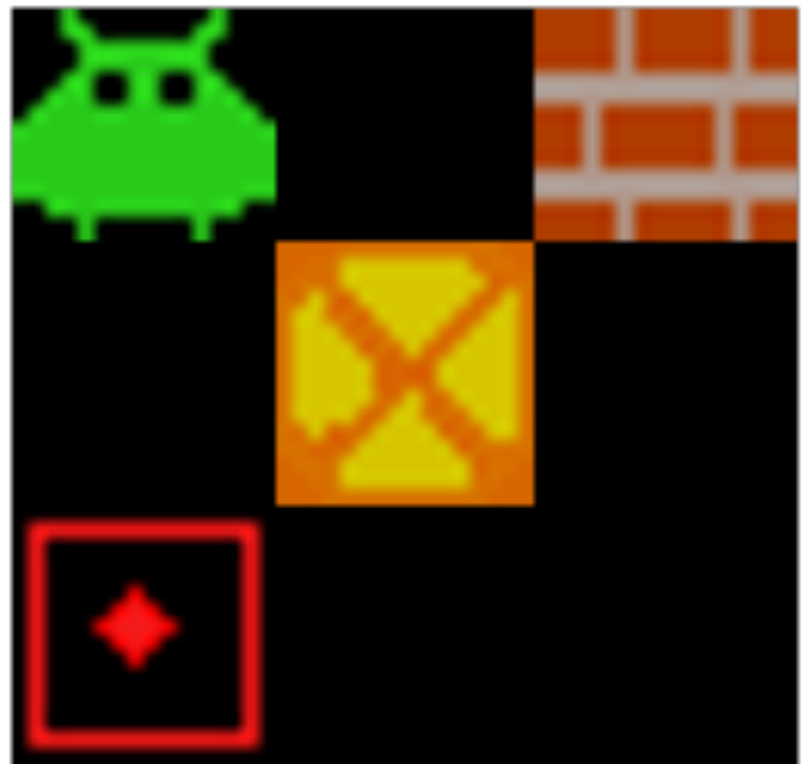
<think>I see the player and the box. The target is somewhere else. I should probably move the box towards it.</think>

Explicit State Verbalization - Grounding+World Modeling

<think><observation>Player at (2,1), Box at (2,2), Target at (4,2)</observation>...<prediction>If Action=Push_Down, Box will be at (3,2), Player at (3,1)</prediction></think>

How should VLMs verbalize visual states?

VAGEN



Visual State

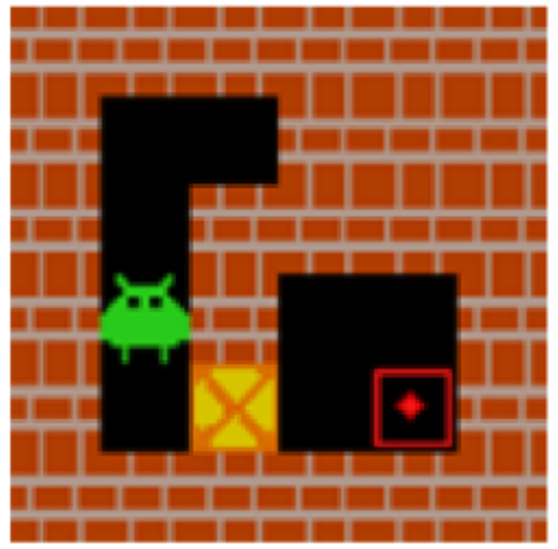
Natural Language	Symbolic	Structured

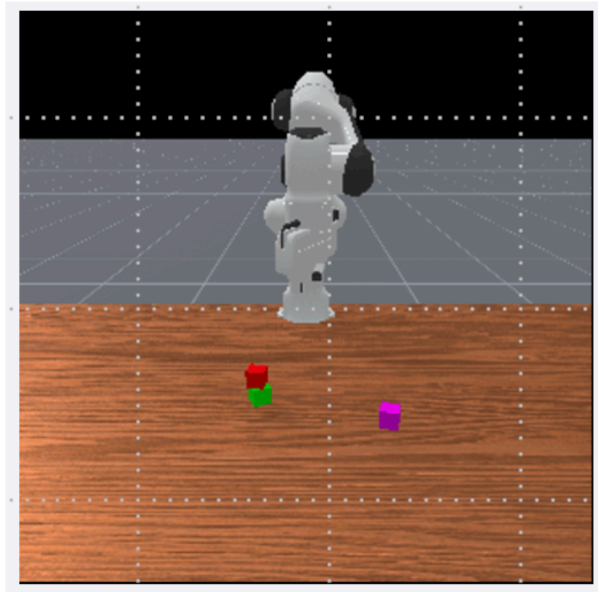
"The player is at the upper-left, the box is to the right of the player, the target is below the player"	P _ O
	_ X _
	# _ _
		{
		'player': [0,0],
		'box': [1,1],
		'target': [0,2],
		}

https://github.com/RAGEN-AI/VAGEN

How should VLMs verbalize visual states?

VAGEN





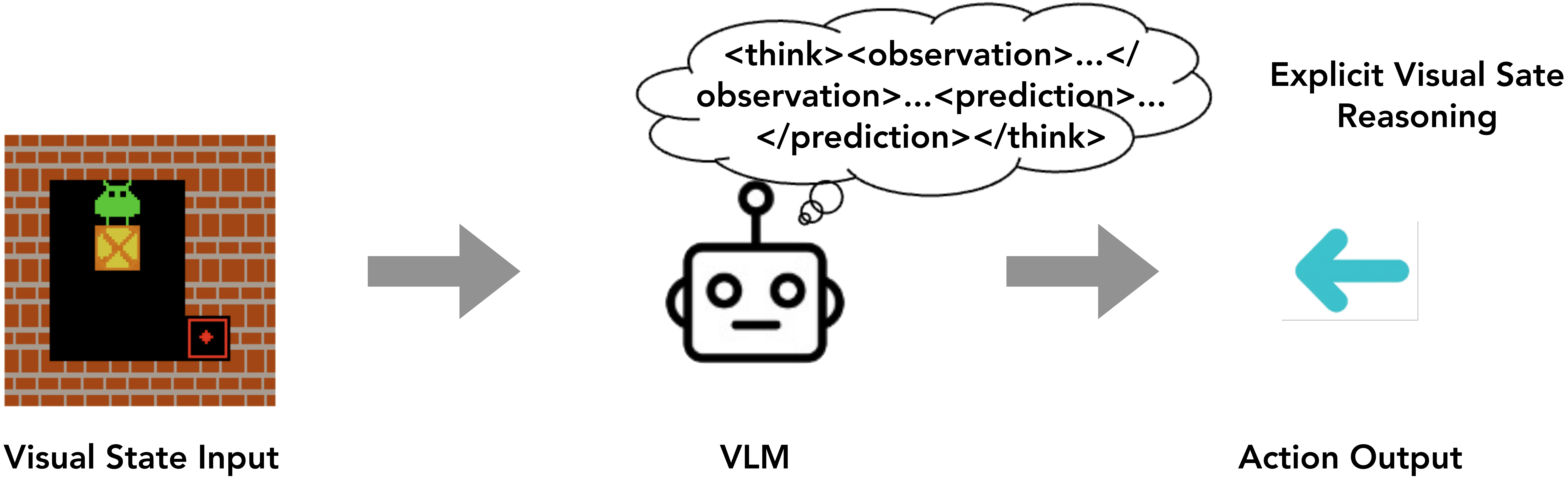
Visual State Representation	FrozenLake	Sokoban	PrimitiveSkill				
			Place	Stack	Drawer	Align	**Average**
`Natural-Lanaguage`	**0.55**	**0.44**	0.63	0.63	0.88	1.00	0.79
`Structured`	0.27	0.35	1.00	0.63	0.88	1.00	**0.88**
`Symbolic`	0.30	0.27	–	–	–	–	–

Optimal Visual State Representation is Task-Dependent.

https://github.com/RAGEN-AI/VAGEN

Reinforcing Explicit Visual State Reasoning

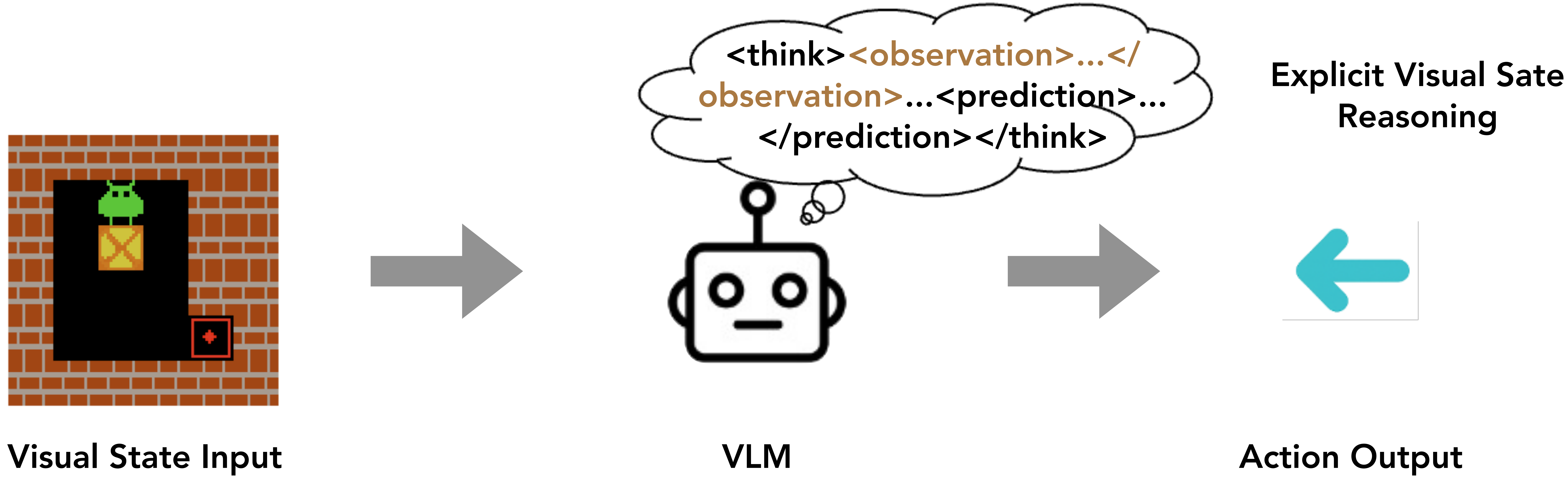
VAGEN



https://github.com/RAGEN-AI/VAGEN

Reinforcing Explicit Visual State Reasoning

VAGEN

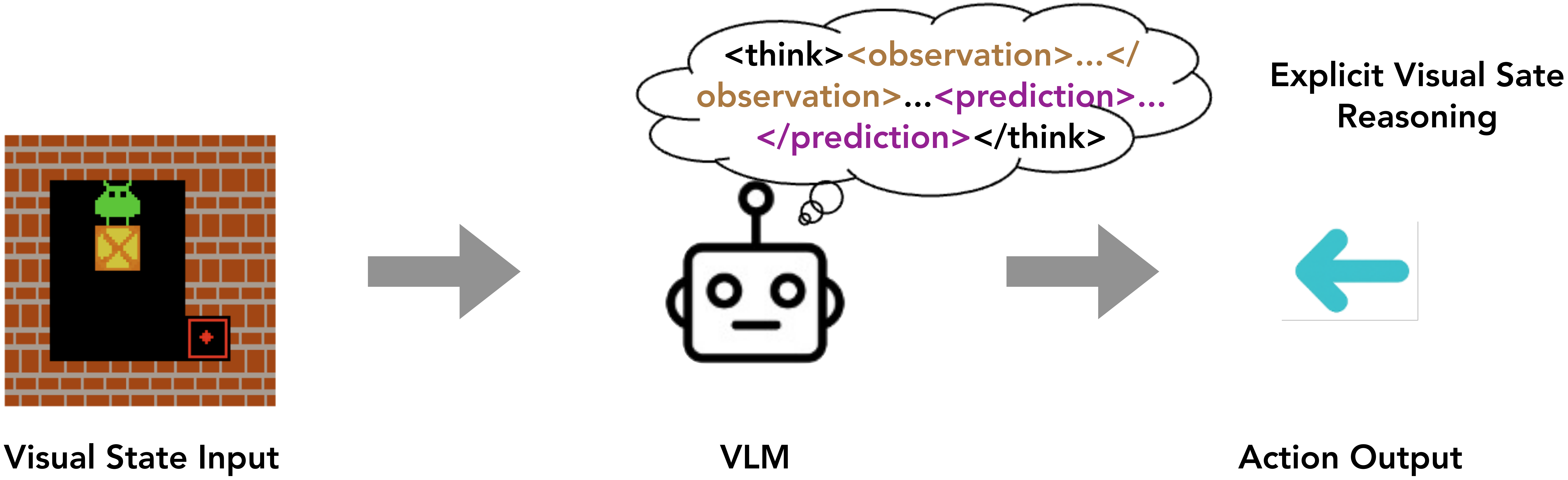


How to verify the correctness of current state verbalization?

https://github.com/RAGEN-AI/VAGEN

Reinforcing Explicit Visual State Reasoning

VAGEN



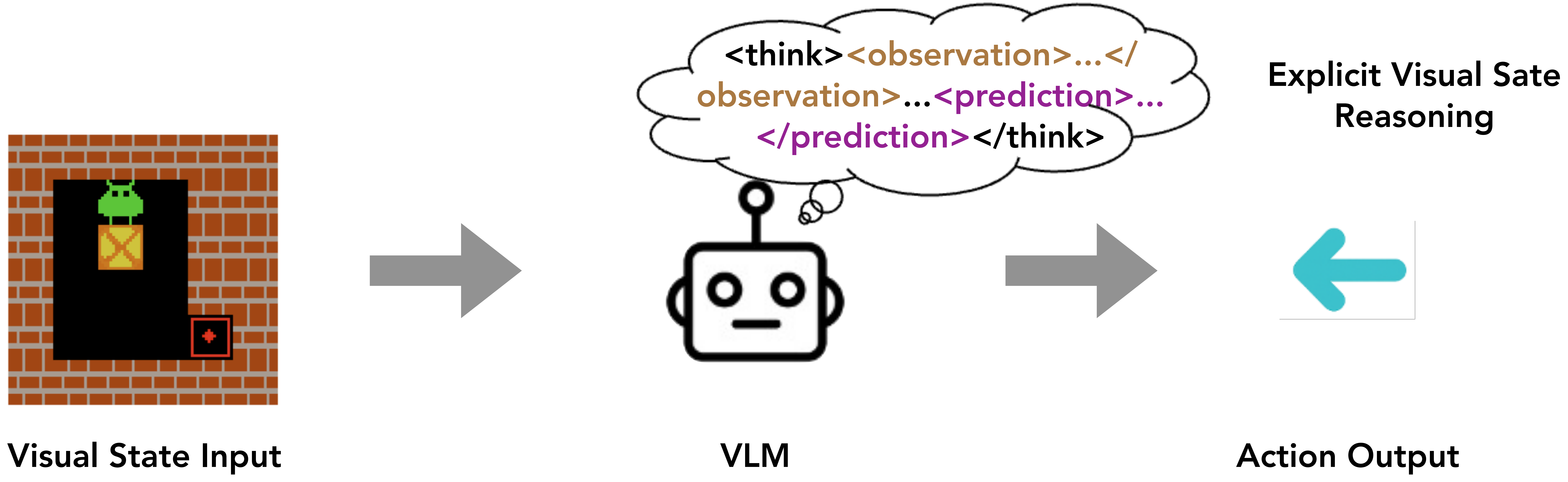
How to verify the correctness of current state verbalization?

How to check the plausibility of next state prediction?

https://github.com/RAGEN-AI/VAGEN

Reinforcing Explicit Visual State Reasoning

VAGEN



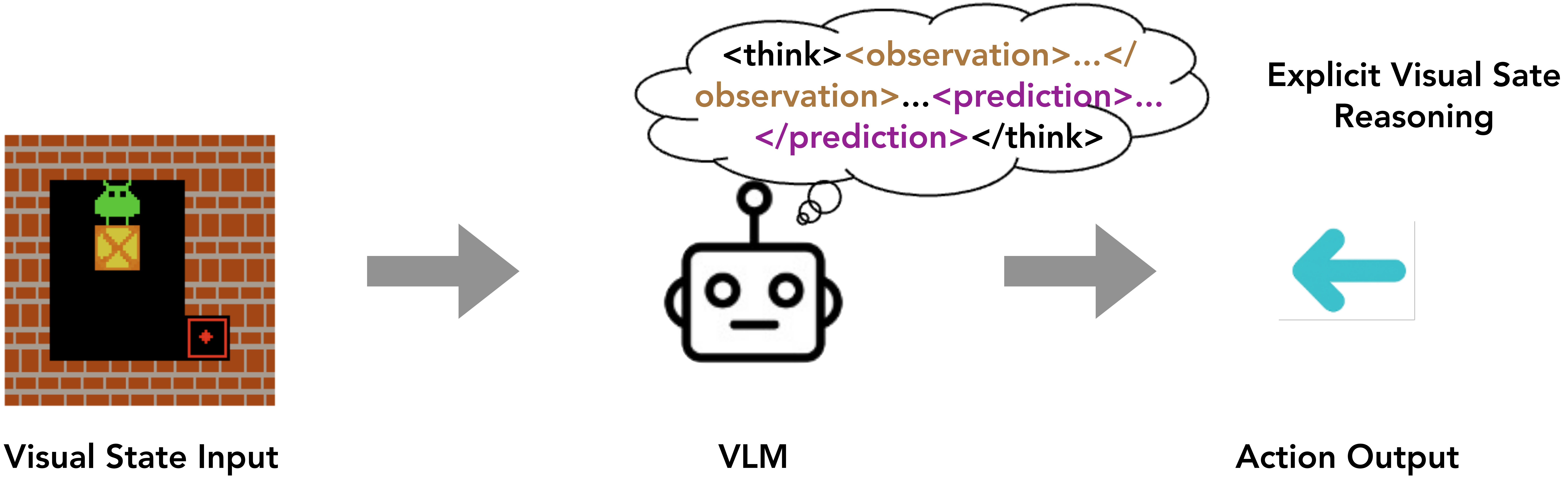
How to verify the correctness of current state verbalization?

How to check the plausibility of next state prediction?

How to assign reward to intermediate reasoning steps (current state+next state) vs. final action?

Reinforcing Explicit Visual State Reasoning

VAGEN



How to verify the correctness of current state verbalization?

How to check the plausibility of next state prediction?

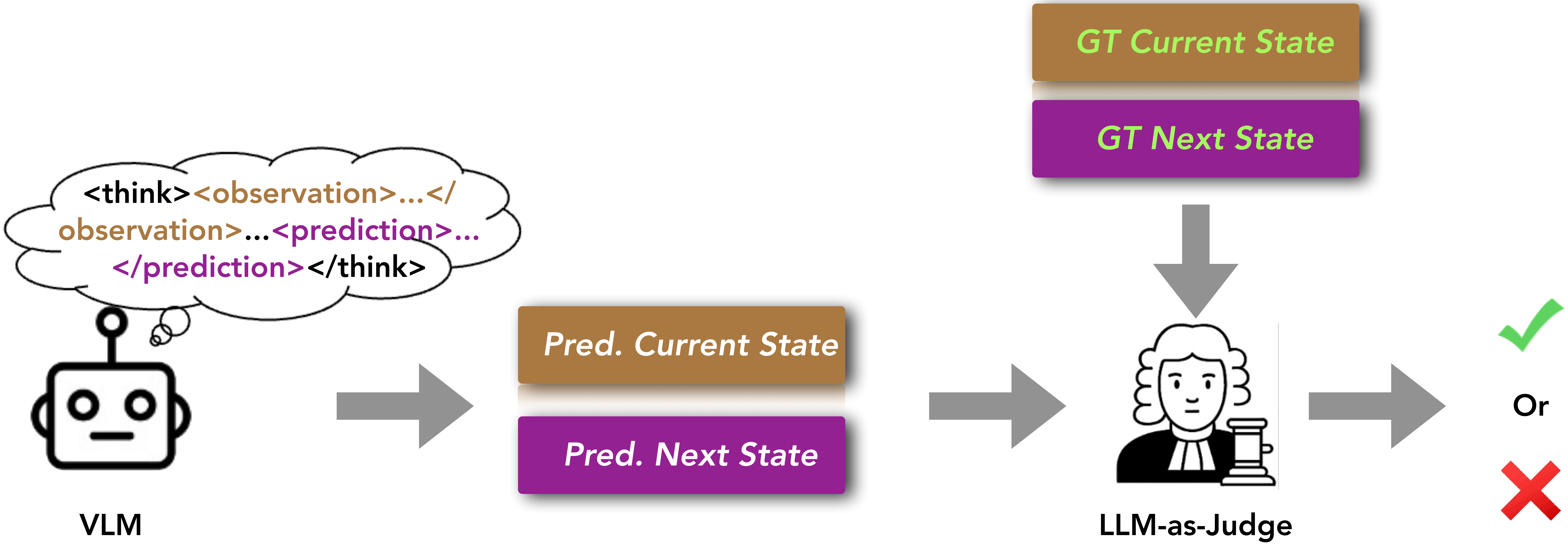
How to assign reward to intermediate reasoning steps (current state+next state) vs. final action?

Visual Reasoning Reward w/ LLM-as-judge

Hierarchical credit assignment w/ Bi-level GAE

Visual Reasoning Reward

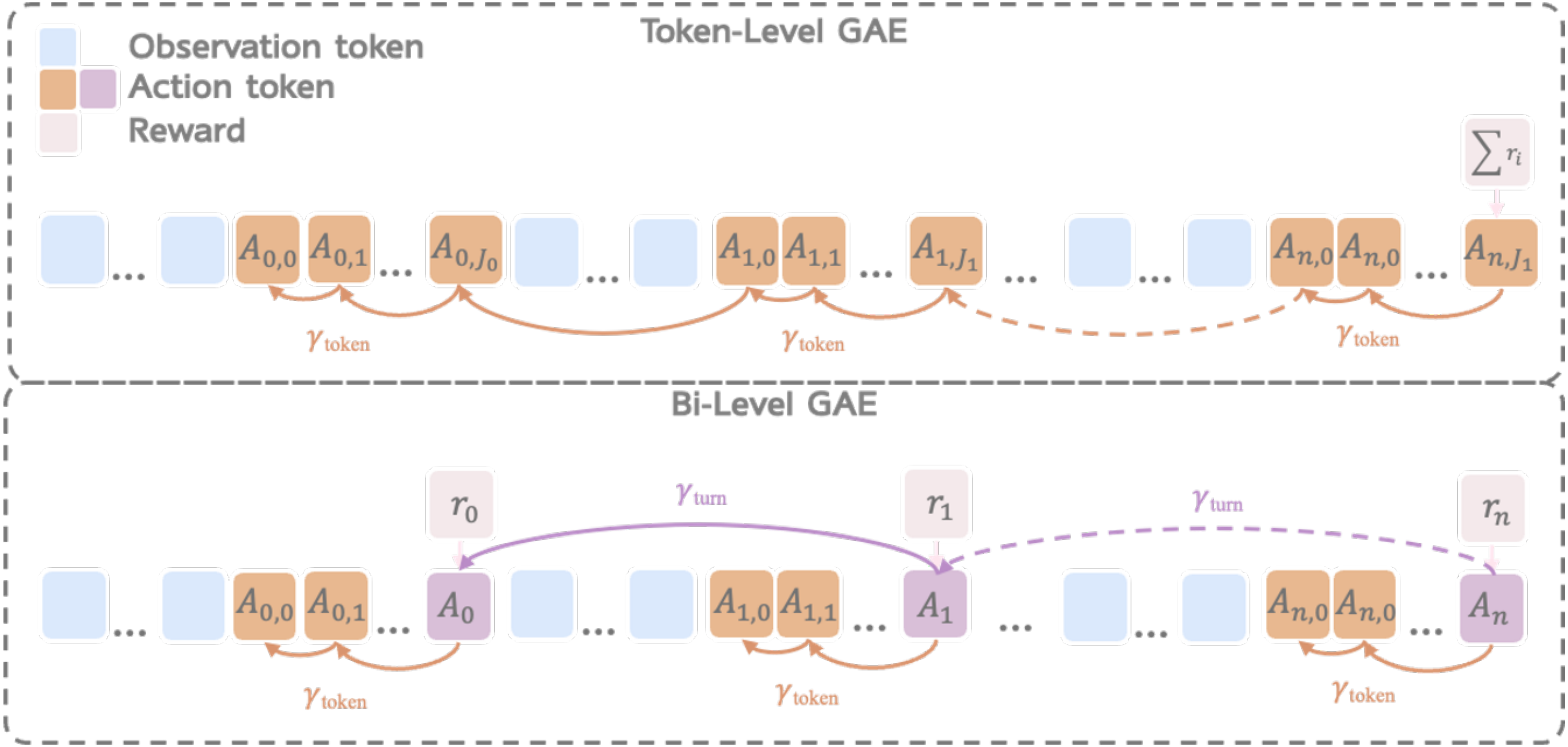
VAGEN



https://github.com/RAGEN-AI/VAGEN

Bi-level General Advantage Estimation

VAGEN



https://github.com/RAGEN-AI/VAGEN

Case Study

VAGEN

	Step 1	VLM Output	Step 2
VR-RL		<think> <observation>There is a plate on the dining table to the right.</observation> <reasoning> First, I should move forward to get closer to the table. Then, I can move to the right to be in front of the table.</reasoning> <prediction>I will to the forward-left of plate and dining table.</prediction> </think> <answer>moveahead, moveright</answer>	
Base RL		<think> <observation>I can see there is a pot on the stove.</observation> <reasoning>To move closer to the pot, I should move forward and then to the right.</reasoning> <prediction>I will closer to the pot.</prediction> </think> <answer>moveahead, moveahead, moveright</answer>	

https://github.com/RAGEN-AI/VAGEN

STARE

EMMA

See. Visual Think. Act.

OpenThinkIMG

"See". Think. Act.

Training Language Agents with Reinforcement Learning

RAGEN

See. Think. Act.

Training Multimodal Agents with Reinforcement Learning

RAGEN VAGEN