

From Large Language Models to Large Agent Models: Agent Training Under A MDP Formulation

Manling Li
Northwestern University
manling.li@northwestern.edu

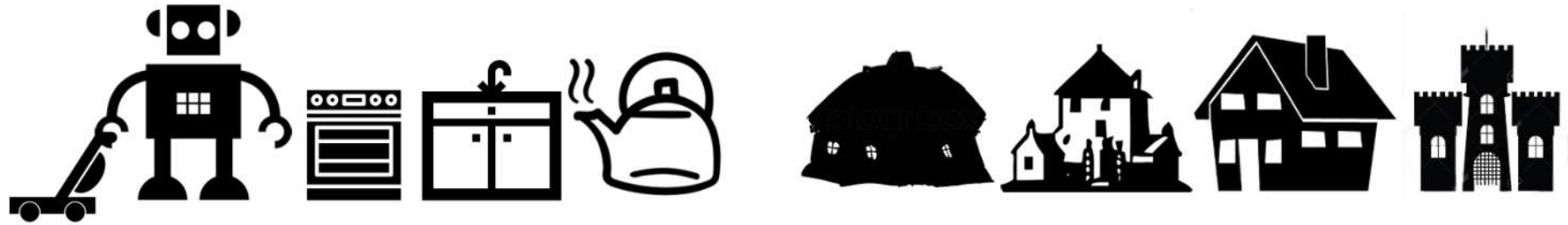
KnowledgeNLP @ NAACL 2025
based on AAAI 2025 New Faculty Highlight



Northwestern
University

What is a generalist agent?

What is a generalist agent

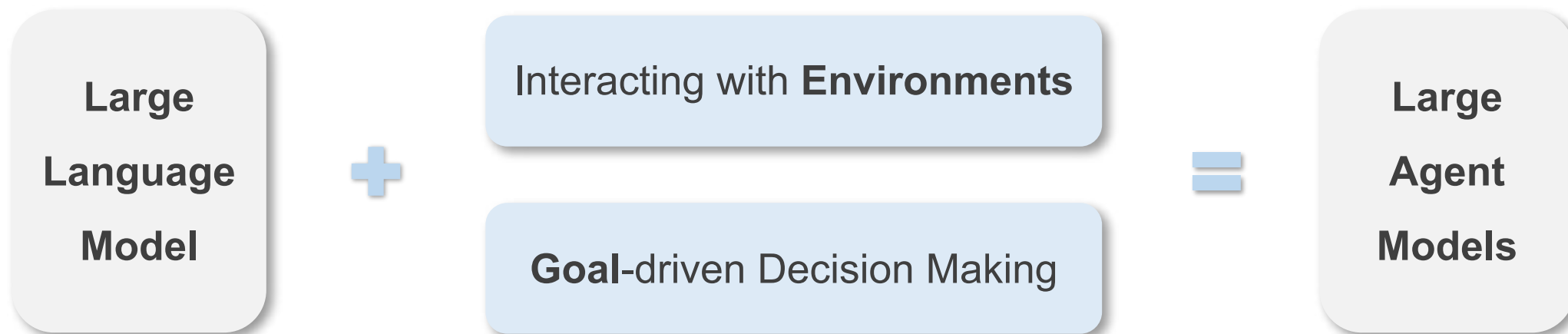


Having an agent that can do many tasks, across many environments.

Goal-driven Decision Making

Interacting with **Environments**

Building Agents





Set up the table.

...

on_top_of (turkey, table)

...





Set up the table.

...

on_top_of (turkey, table)

...



Use the plates.

...

on_top_of (turkey, plate)

on_top_of (plate, table)

...



Trajectory

...

on_top_of (turkey, plate)

on_top_of (plate, table)

...



What is “embodied decision making”?

Can we leverage MDP as a guiding principle to categorize “foundation models”?

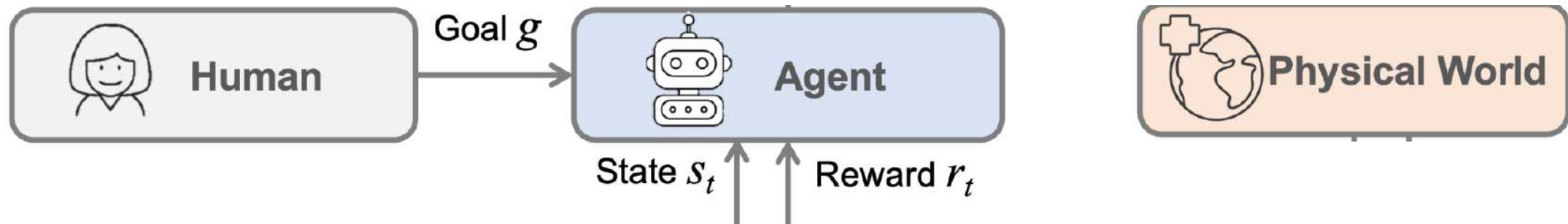
Let us go back to MDPs (Markov Decision Processes)



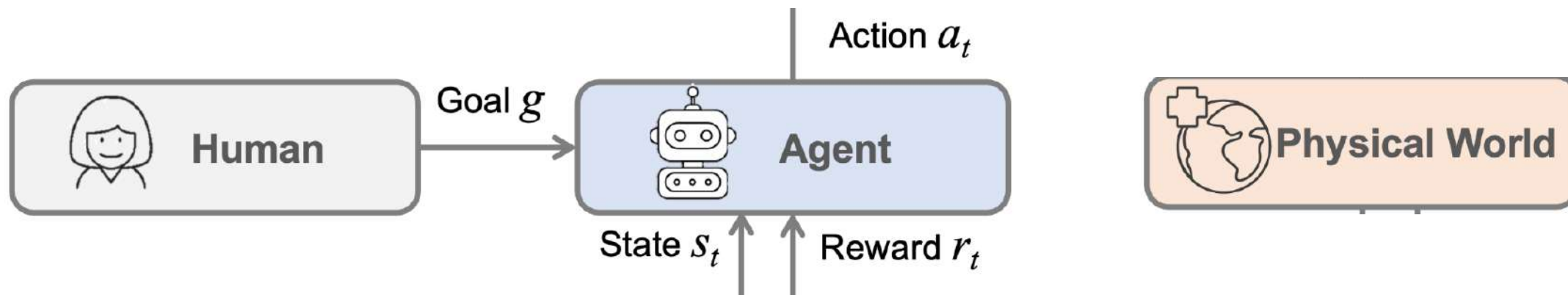
Let us go back to MDPs (Markov Decision Processes)



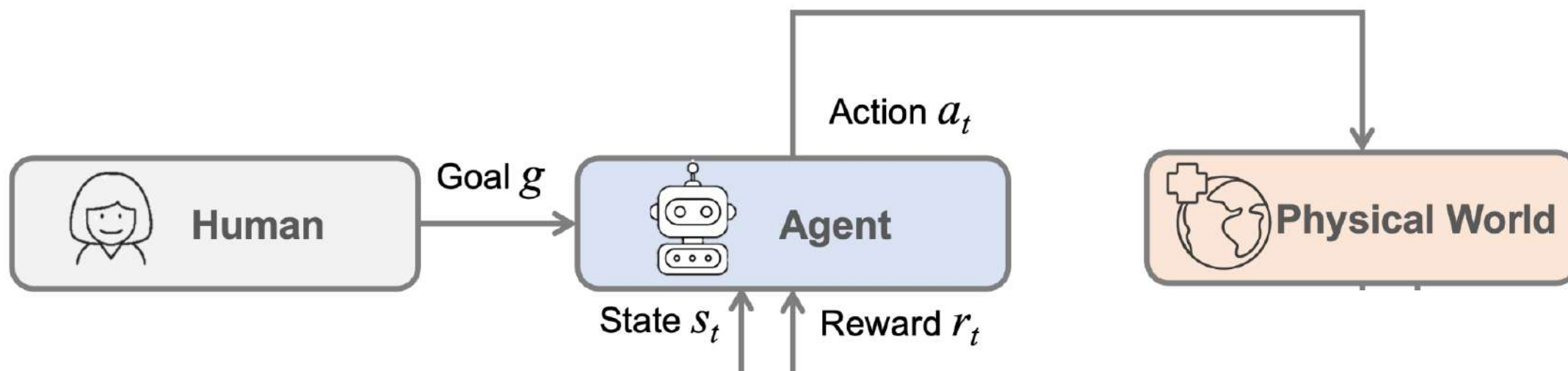
Let us go back to MDPs (Markov Decision Processes)



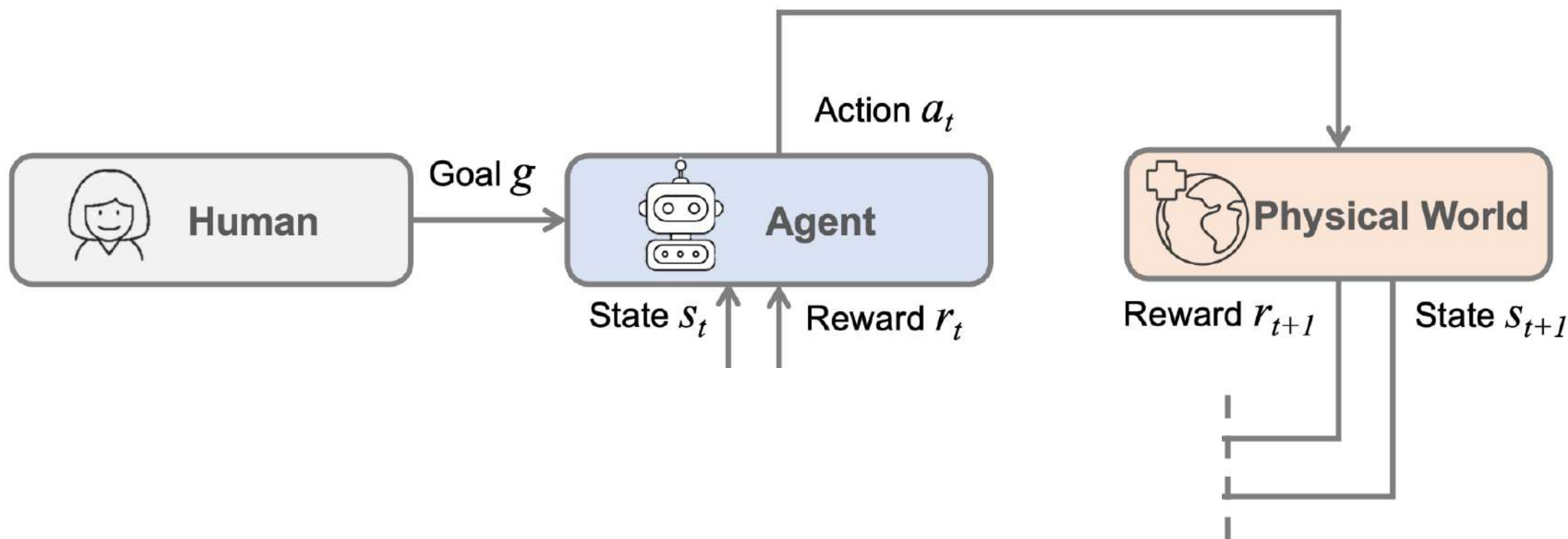
Let us go back to MDPs (Markov Decision Processes)



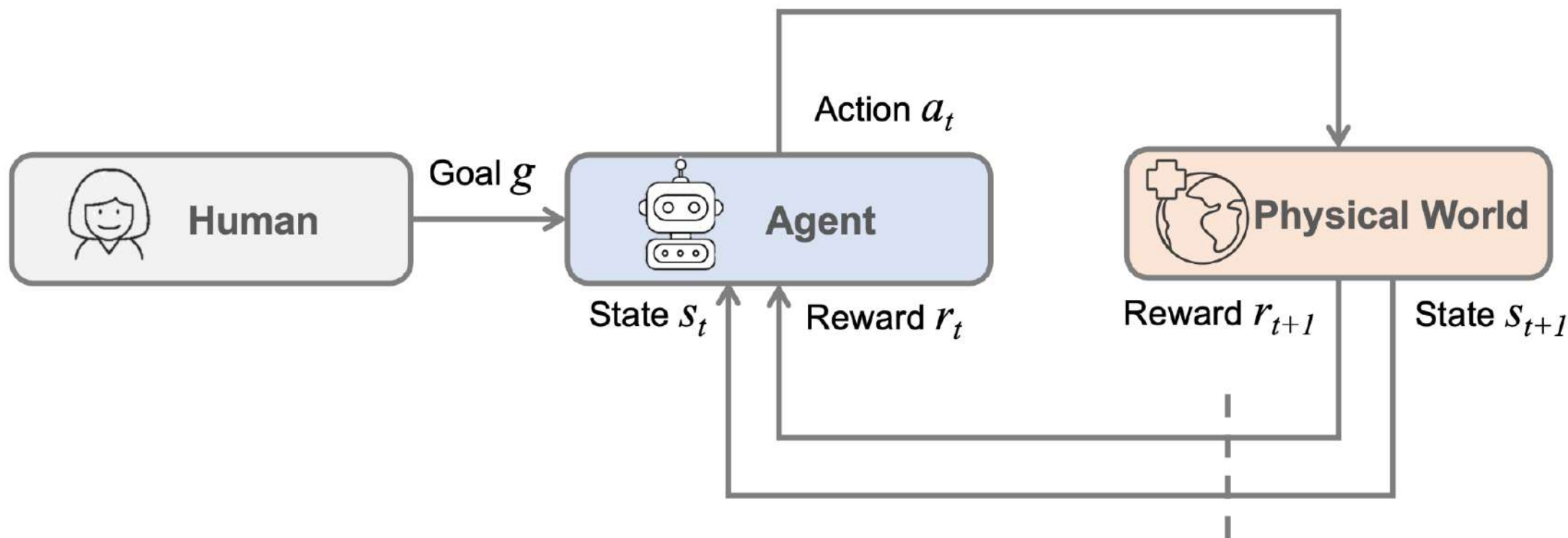
Let us go back to MDPs (Markov Decision Processes)



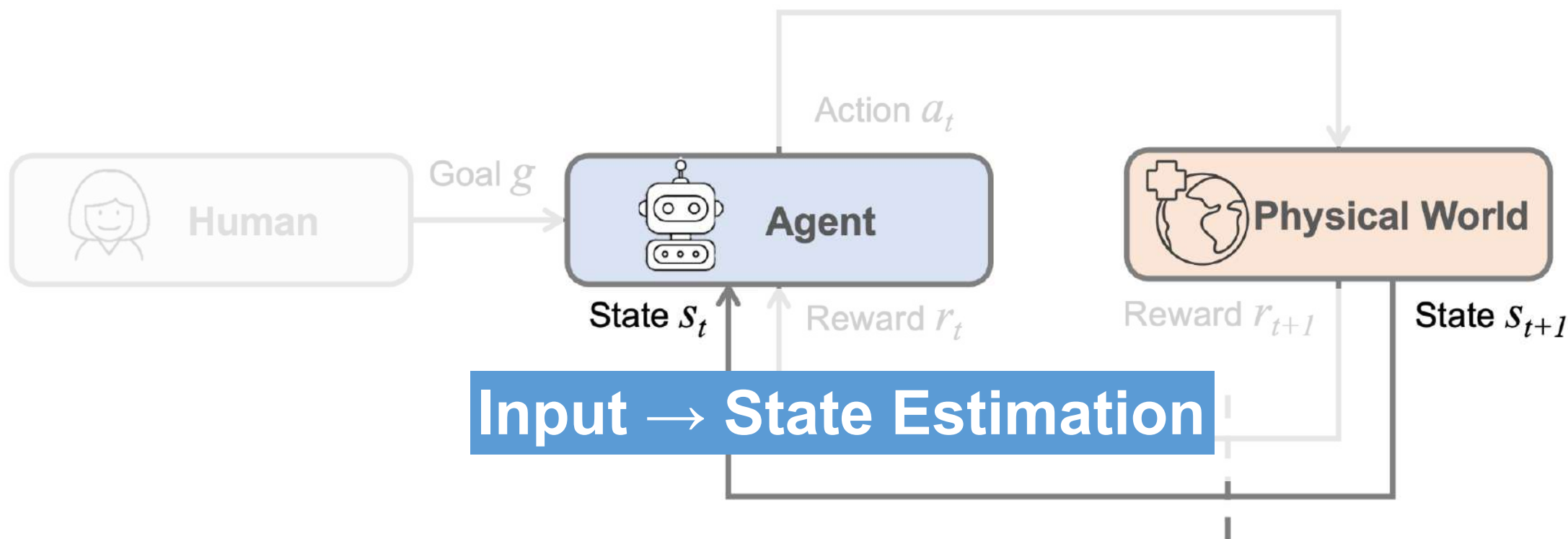
Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Perception / State Estimation

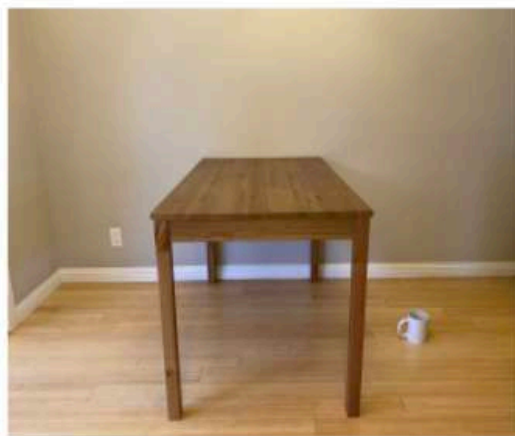
$$o \rightarrow s$$

Environment Interaction



Cup is on top of the table.

Cup is on top of the table.



Cup is on top of the table.

Cup is on top of the table.



Model	Whats- Up	COCO- spatial	GQA- spatial	Avg
CLIP ViT-B/32	31.0	47.4	46.9	41.8
CLIP ViT-L/14	26.1	49.5	47.3	41.0
NegCLIP	34.4	46.9	46.0	42.4
RoBERTaCLIP	25.1	50.0	49.8	41.6
CoCa	29.4	46.7	47.1	41.0
XVLM 4M	31.5	61.7	58.7	50.6
XVLM 16M	41.9	65.0	58.2	55.0
BLIP 14M	38.5	54.0	49.8	47.5
BLIP 129M	30.4	49.3	49.0	42.9
BLIP2-ITM	37.6	53.0	49.8	46.8
BLIP2-ITC	29.0	53.7	51.0	44.6
FLAVA	30.5	52.6	51.7	44.9
CoCa-Caption	24.1	48.6	49.5	40.8
XVLM-Flickr30K	44.3	65.2	61.4	56.9
XVLM-COCO	42.1	71.0	68.1	60.4
BLIP-Flickr30K	33.8	54.2	48.9	45.6
BLIP-COCO	32.8	51.4	51.4	45.2
BLIP-VQA	47.8	62.0	58.4	56.0
Random / Text-only	25.0	50.0	50.0	41.7
Human Estimate	100.0	97.3	99.0	98.8

What is missing: Physical World State Changes

Current **LLMs** still **fall short** on understanding concepts involving complex **physical interactions**.

Place a **blue chair** behind a **red chair**. Then, put a **yellow chair** behind the **blue chair**. Then, put a **book** on top of the chair that is in front of the **blue chair**.

Question: What chair is the book on?

Answer: The book is on the yellow chair.

GPT4 failed
example

Mental models, which can be viewed as **internal representations of the physical world**, enable **humans** to understand such concepts with ease.

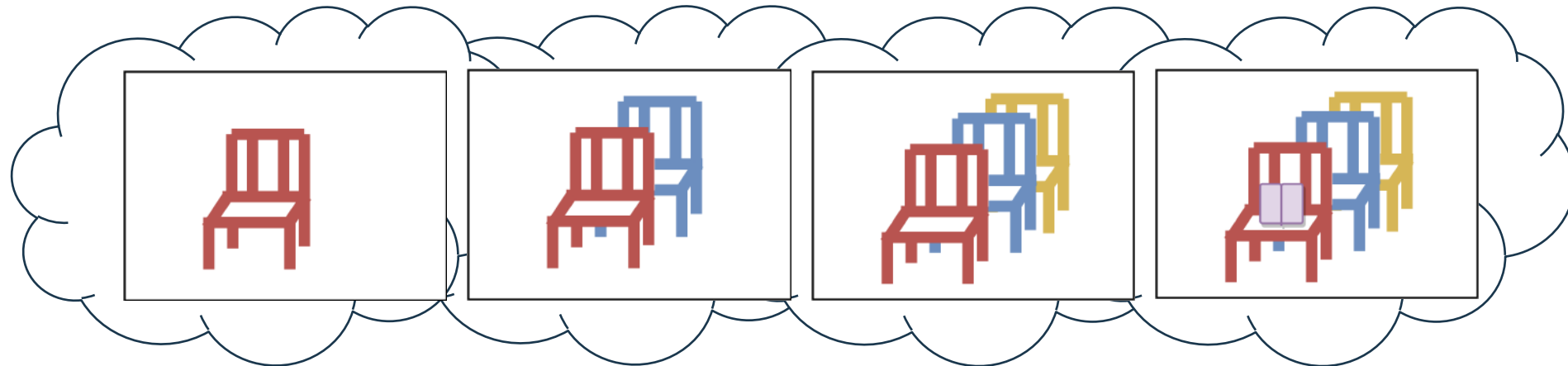
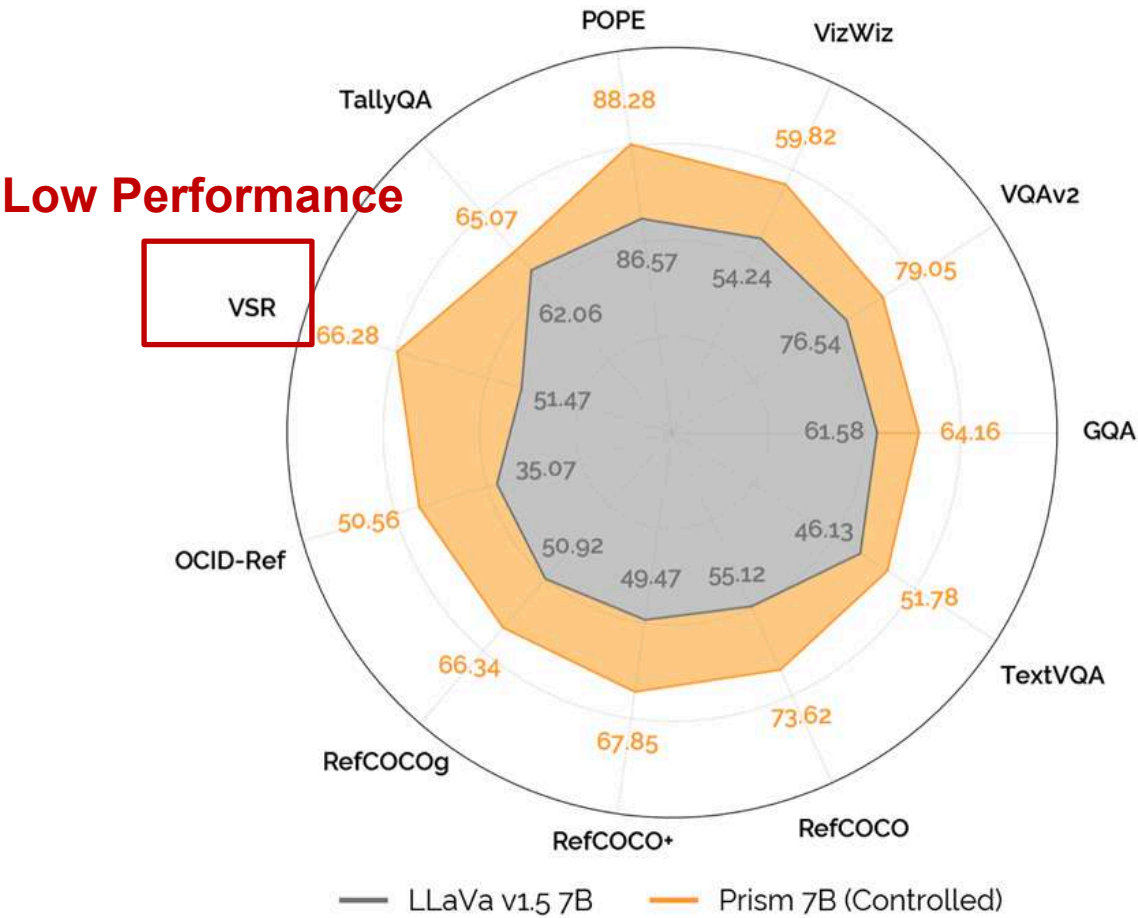


Illustration
Of
Mental
Simulation

Low Performance on Visual Spatial Reasoning



Video-Language Foundation Models
Prismatic VLMs (2024)

What we know about physical world

Semantics



Object Concept

Relationship

Action

Planning

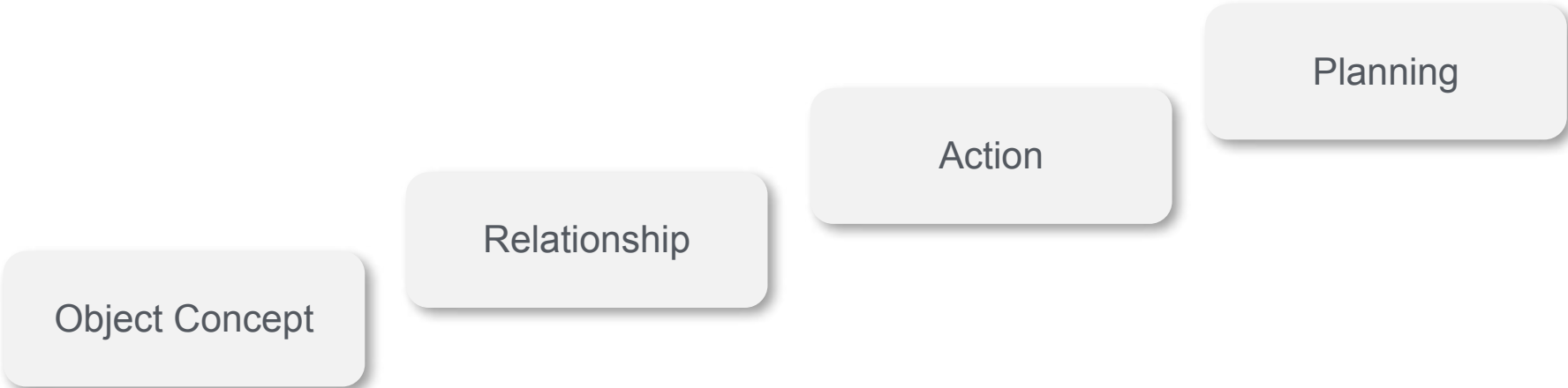
Long-Horizon



Missing knowledge about physical world

Horizon

Semantic



Long-Horizon

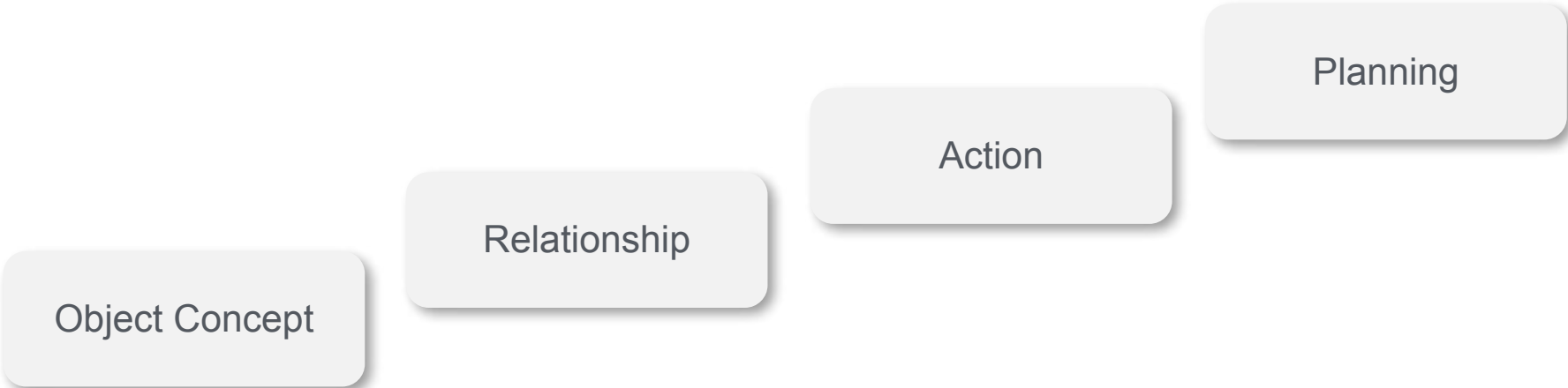
?

Geometric

Missing knowledge about physical world

Horizon

Semantic



Long-Horizon

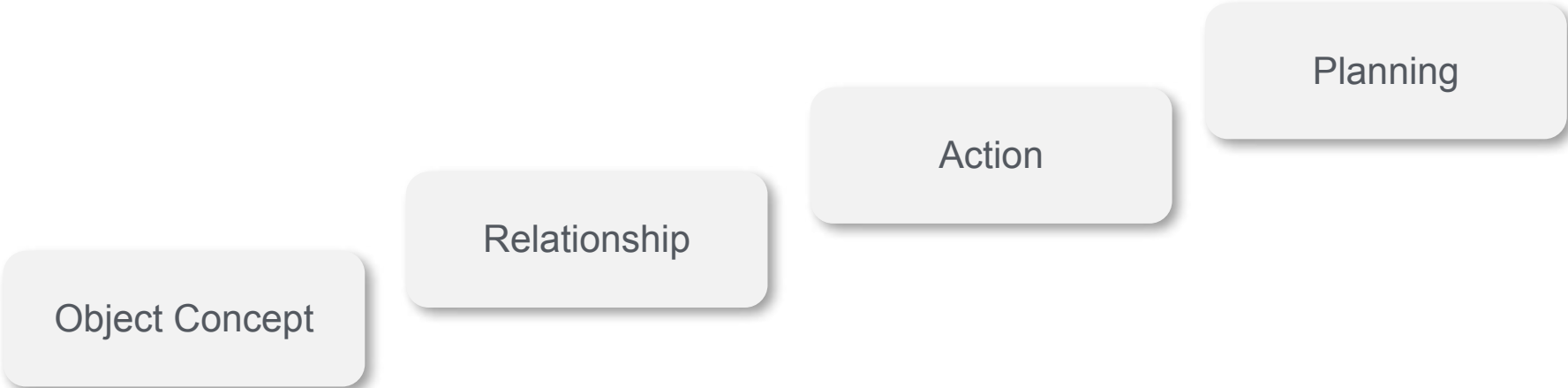
?

Geometric

Missing knowledge about physical world

Horizon

Semantic



Long-Horizon

Shape / Color ...



Pose / Orientation...

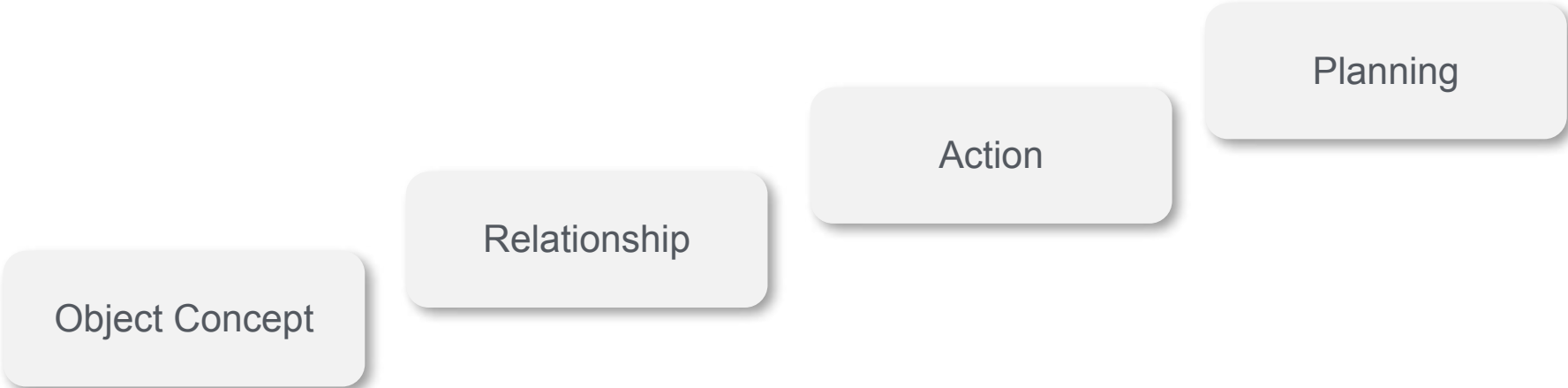
Geometric Features

Geometric

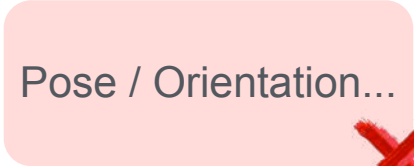
Missing knowledge about physical world

Horizon

Semantic



Long-Horizon



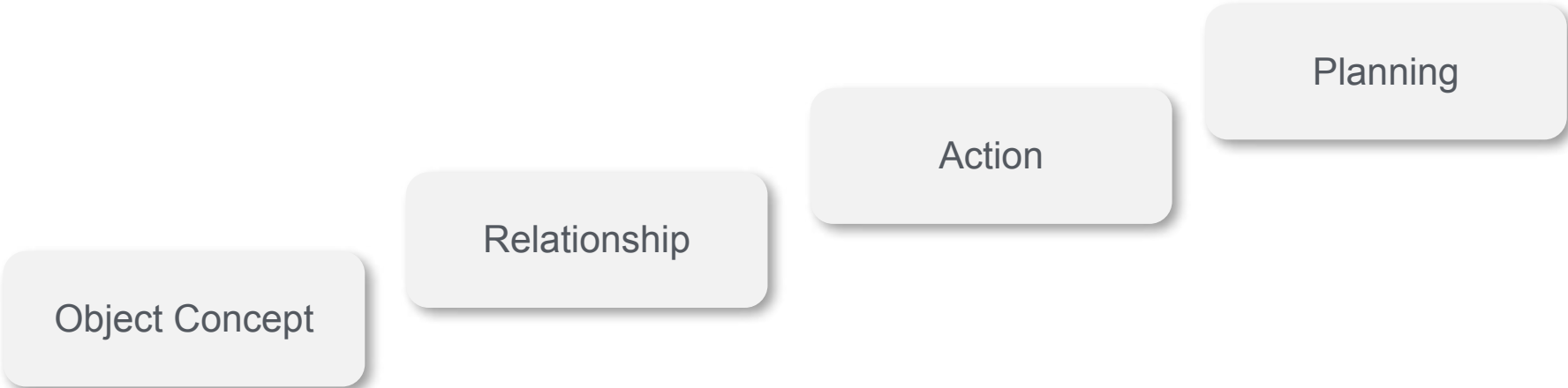
Geometric Features

Geometric

Missing knowledge about physical world

Horizon

Semantic



Long-Horizon

Shape / Color ...

Spatial Relation

Pose / Orientation...

Egocentrism

Geometric Features

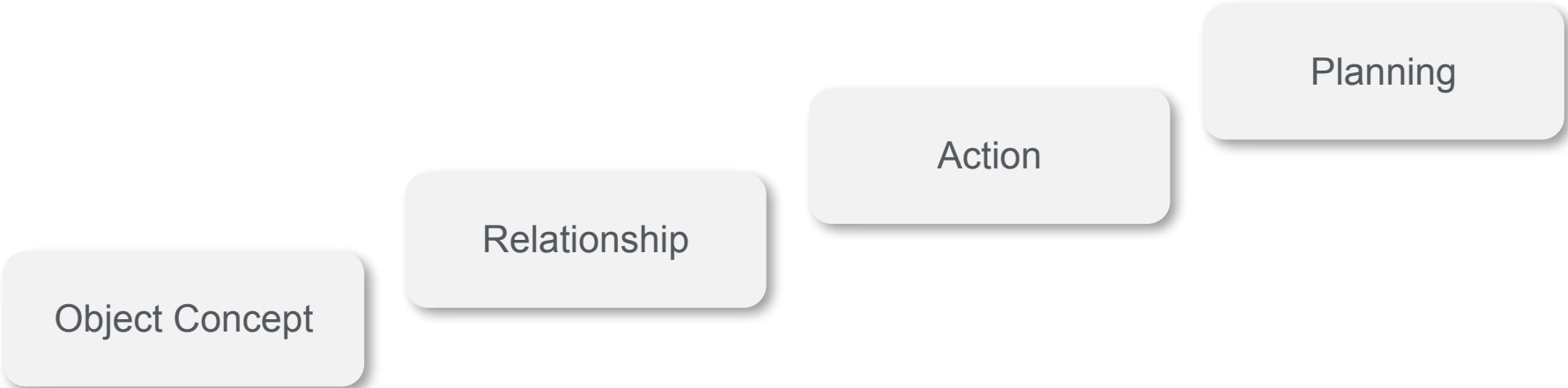
Geometric Structures

Geometric

Missing knowledge about physical world

Horizon

Semantic



Long-Horizon

Shape / Color ...

Spatial Relation

Pose / Orientation...

Egocentrism

Geometric Features

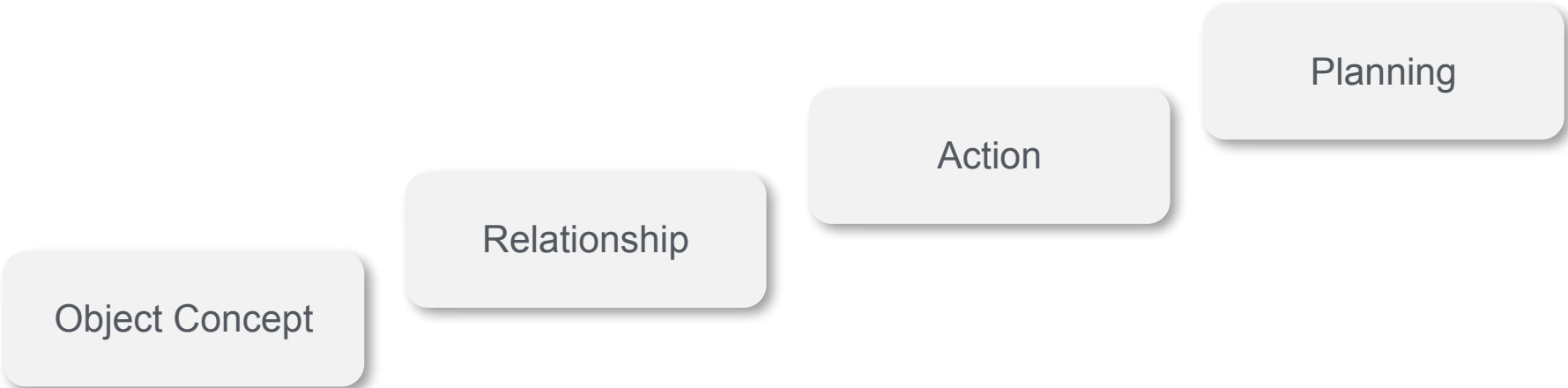
Geometric Structures

Geometric

Missing knowledge about physical world

Horizon

Semantic



Object Concept

Relationship

Action

Planning

Shape / Color ...

Spatial Relation

State Changes

Transition Model

Spatial Reasoning

Pose / Orientation...

Egocentrism

Mental Simulation

Navigation

Episodic Memory

Geometric Features

Geometric Structures

Laws of Physics

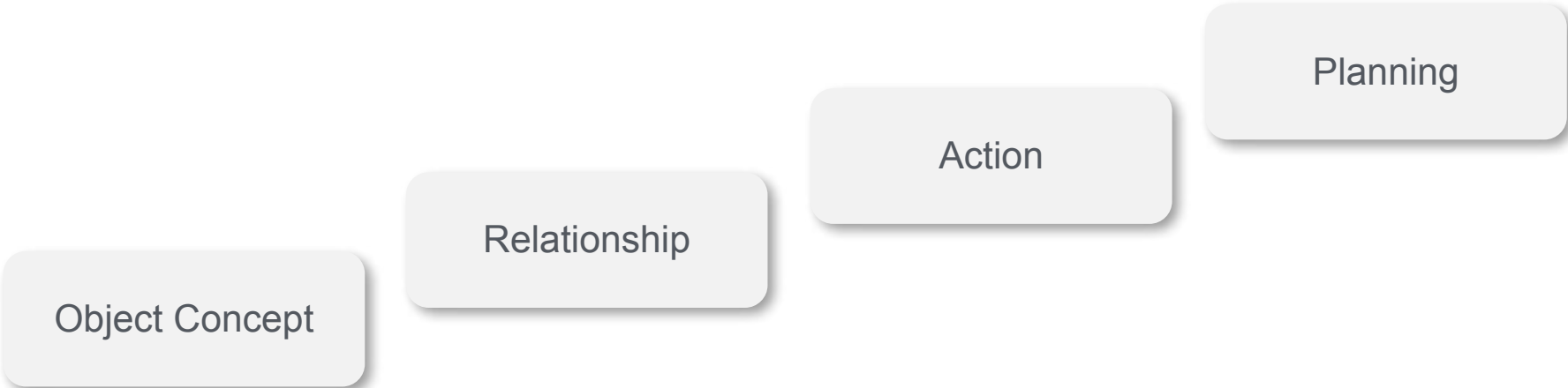
Geometric

Long-Horizon

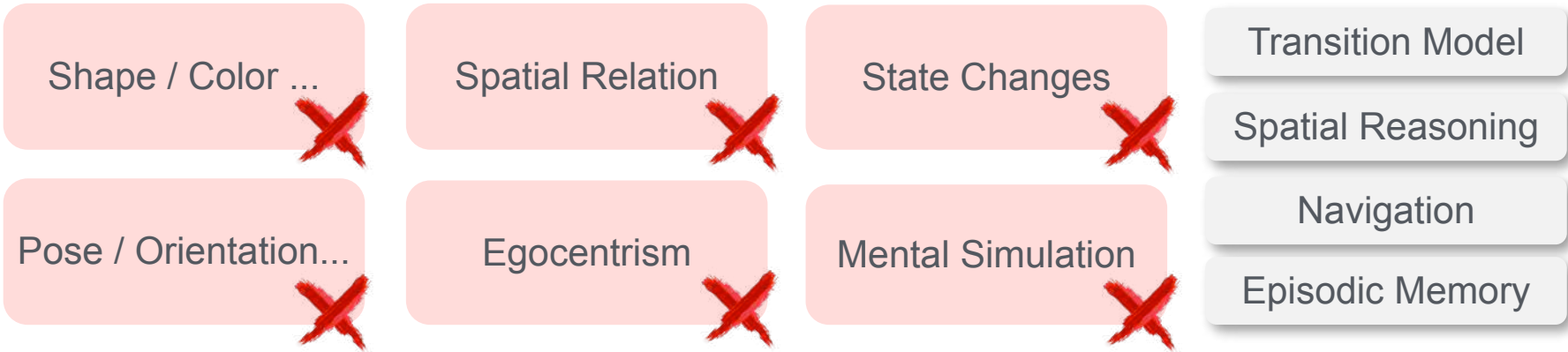
Missing knowledge about physical world

Horizon

Semantic



Long-Horizon



Geometric

Geometric Features

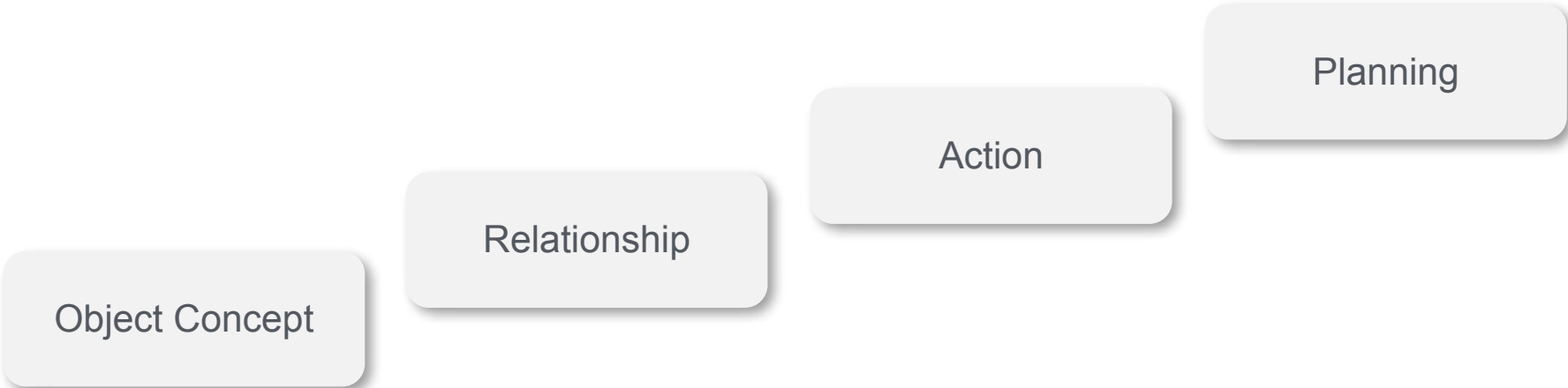
Geometric Structures

Laws of Physics

Missing knowledge about physical world

Horizon

Semantic



Object Concept

Relationship

Action

Planning

Shape / Color ...

Spatial Relation

State Changes

Transition Model

Spatial Reasoning

Pose / Orientation...

Egocentrism

Mental Simulation

Navigation

Episodic Memory

Geometric Features

Geometric Structures

Laws of Physics

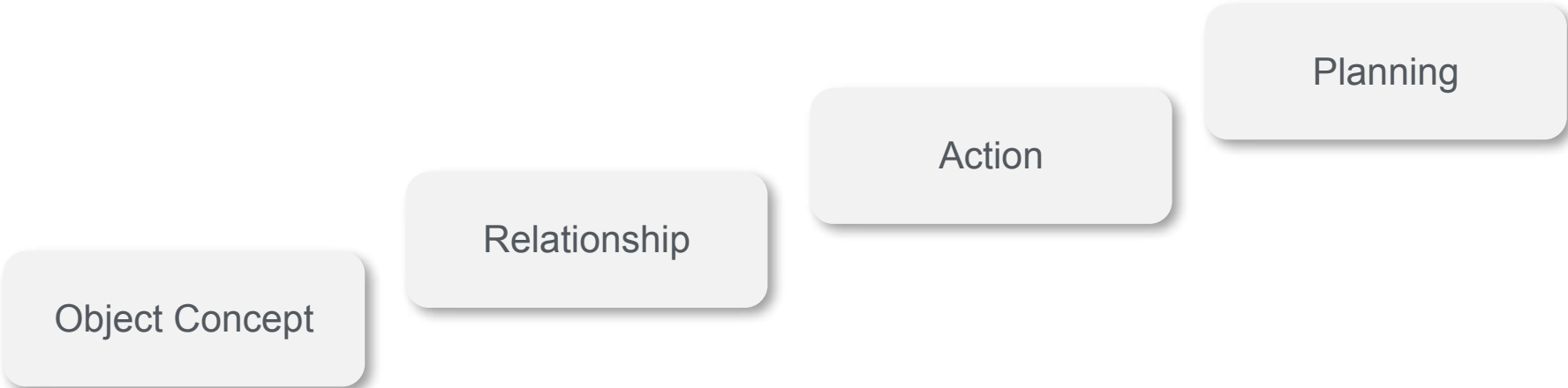
Geometric

Long-Horizon

Missing knowledge about physical world

Horizon

Semantic



Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ...

Spatial Relation

State Changes

Transition Model

Spatial Reasoning

Pose / Orientation...

Egocentrism

Mental Simulation

Navigation

Episodic Memory

Geometric Features

Geometric Structures

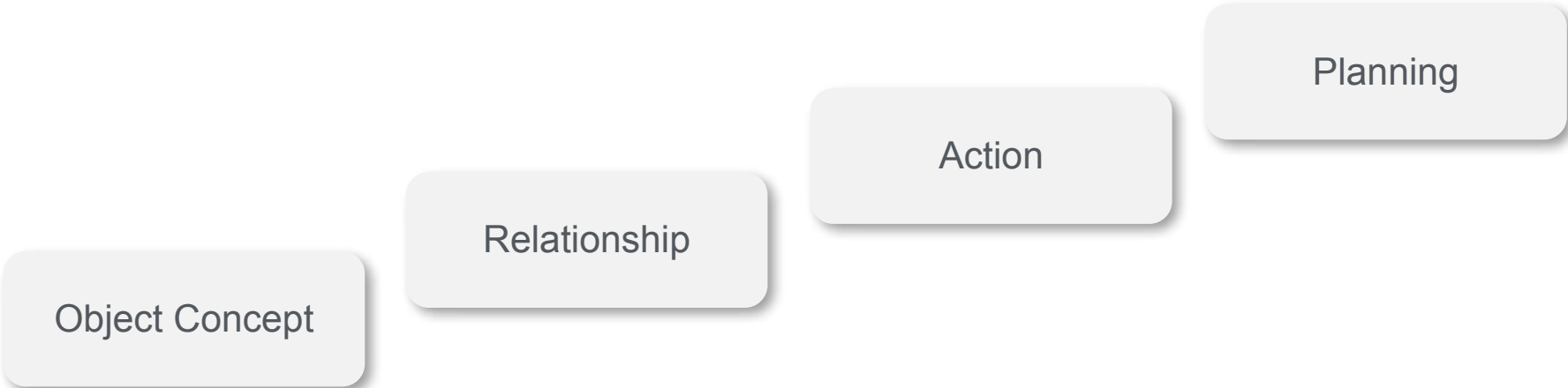
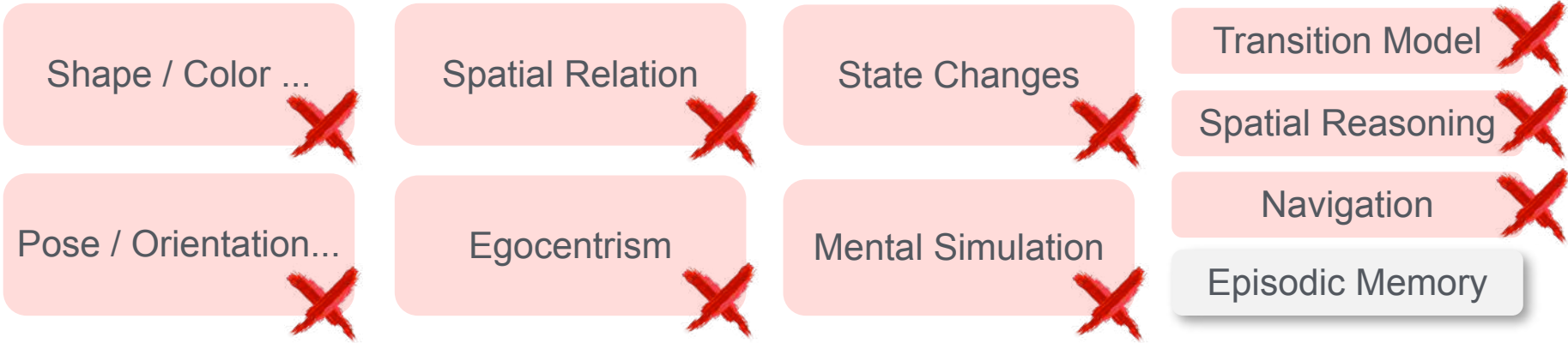
Laws of Physics

Geometric

Missing knowledge about physical world

Horizon

Semantic



Geometric

Long-Horizon

Geometric Features

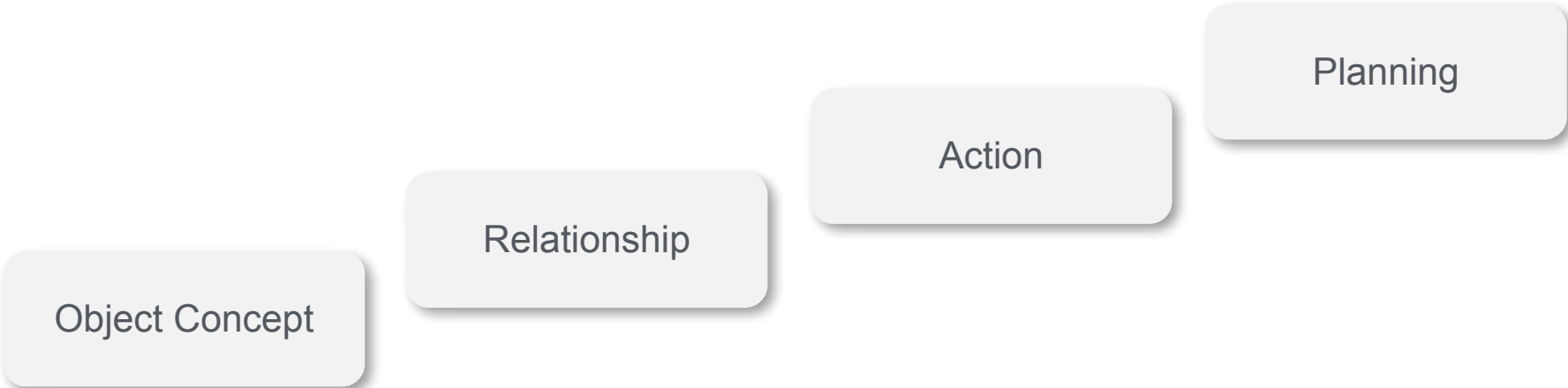
Geometric Structures

Laws of Physics

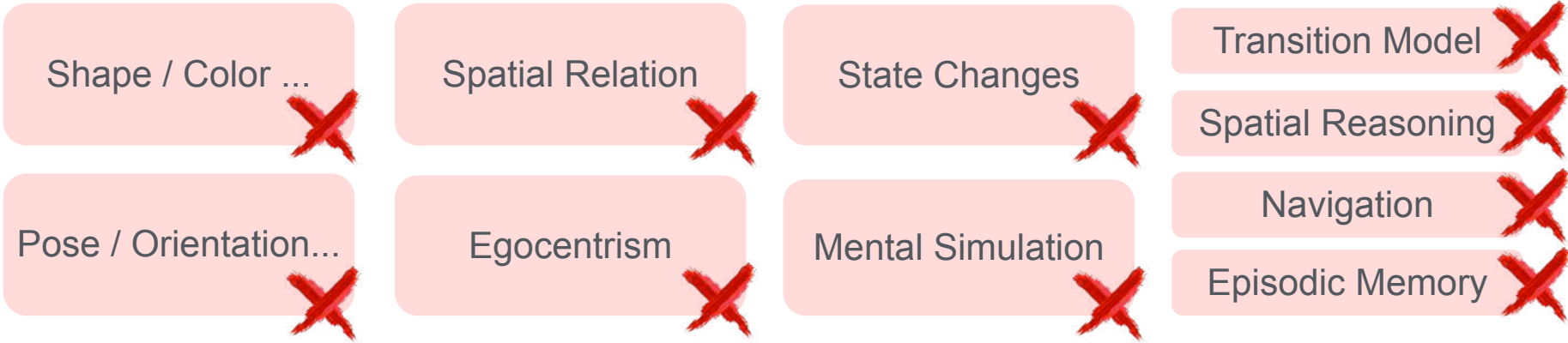
Missing knowledge about physical world

Horizon

Semantic



Long-Horizon

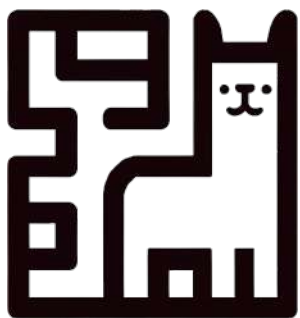


Geometric Features

Geometric Structures

Laws of Physics

Geometric



Why Is Spatial Reasoning **Hard** in VLMs



Shiqi Chen



Tongyao Zhu



Ruochen Zhou



Jinghan Zhang



Siyang Gao



Juan Carlos Niebles



Mor Geva



Junxian He



Jiajun Wu



Manling Li

Spatial Reasoning is a **bottleneck** for current VLMs

Where is the book in relation to the candle?



VLM
(e.g.LLaVA)

Left

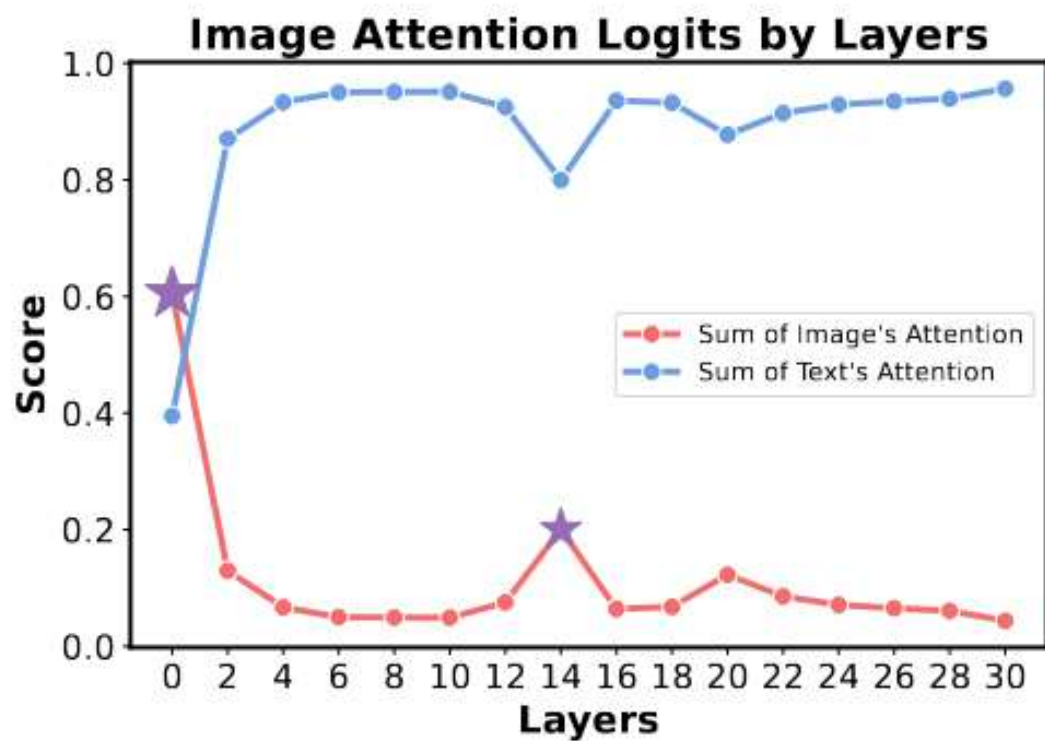
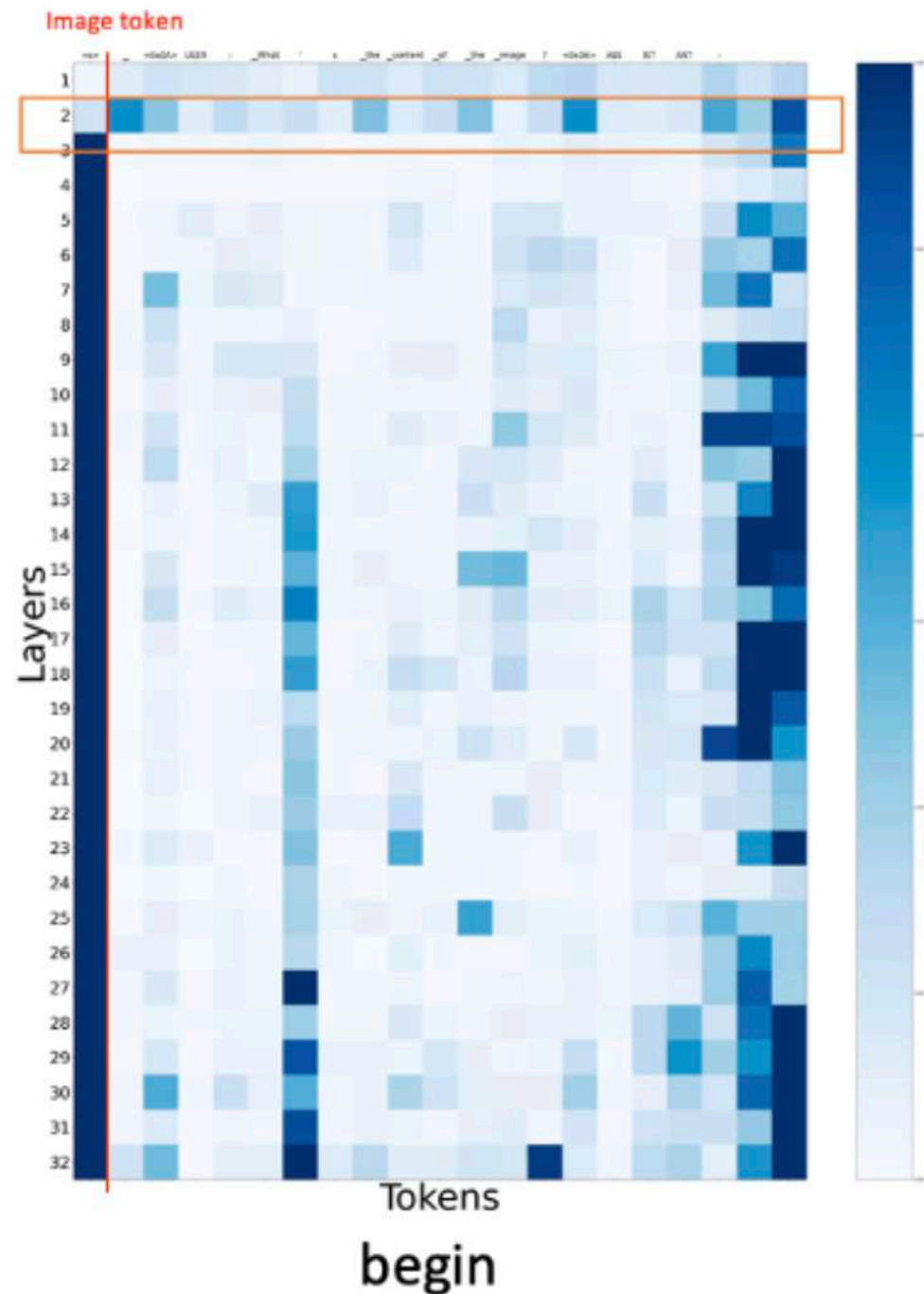


Where is the cat in relation to the person?

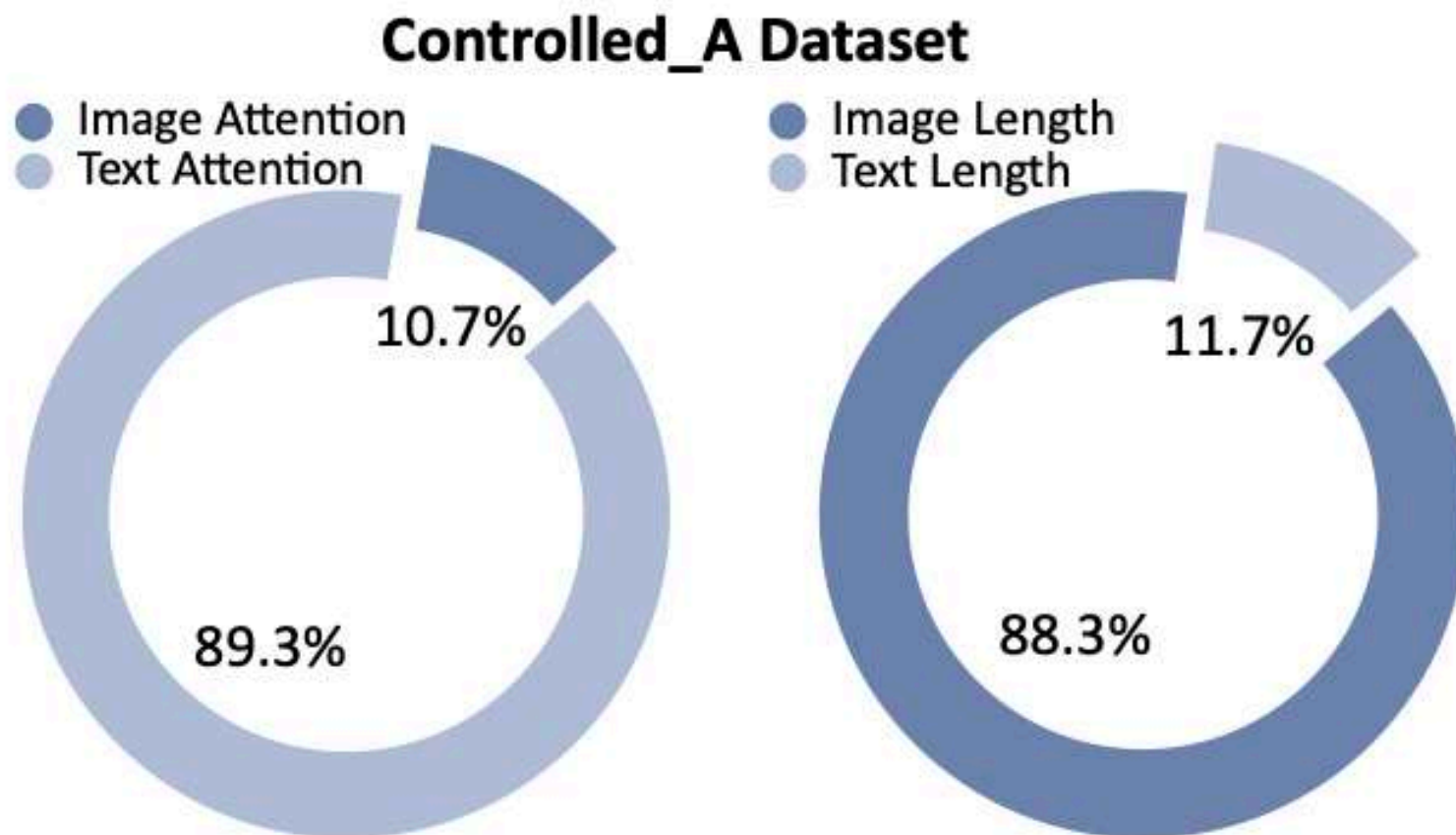
Above



Open Up the Model

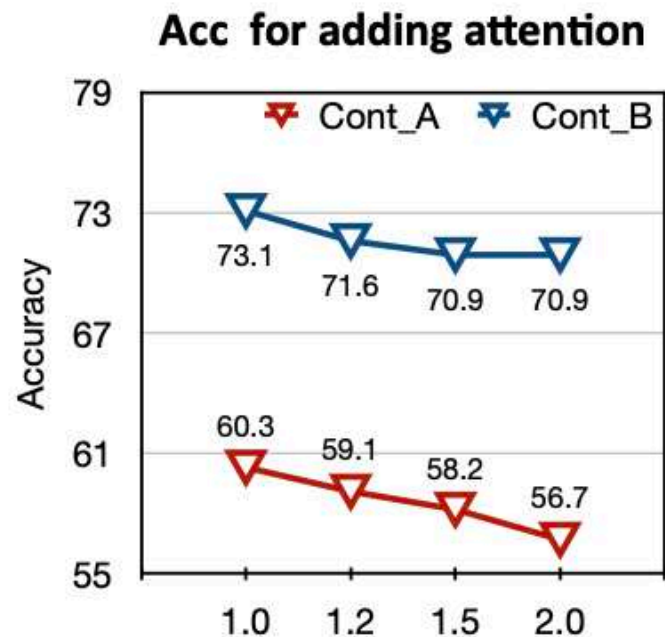
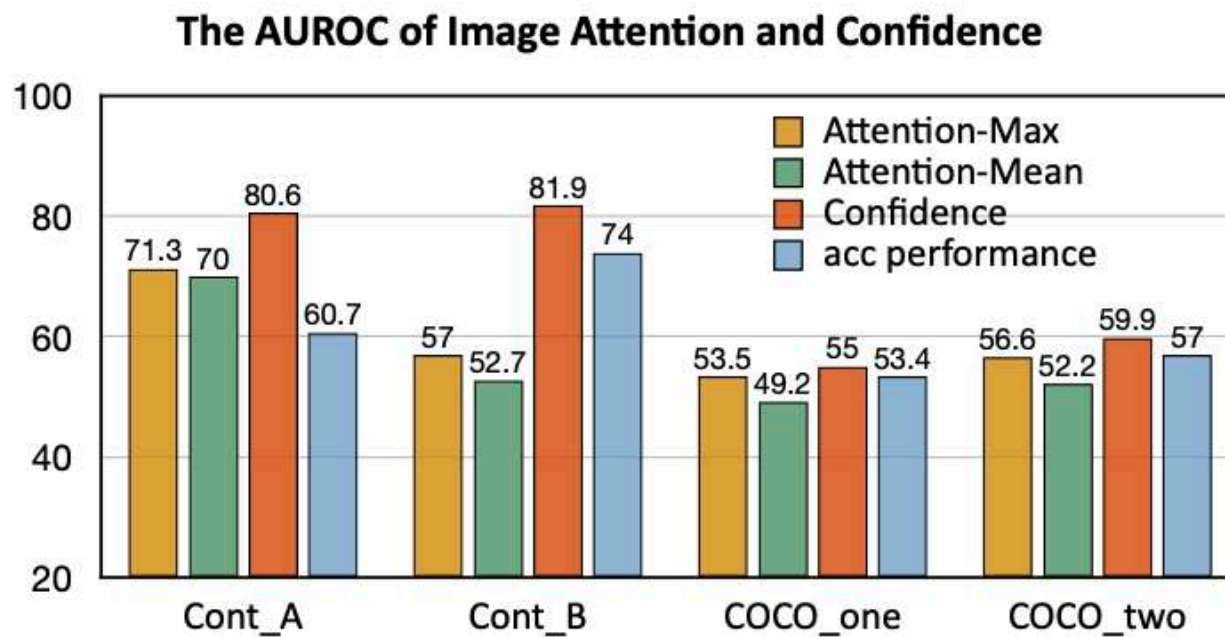


The sparsity of Image's attention scores



Let us simply add more attention

Accuracy Not Improved if the Model Sees the Image More



We open up the model and find the **Distribution** is the key.
VLM would **SEE** 🔍 the **wrong place** or **missing something!**

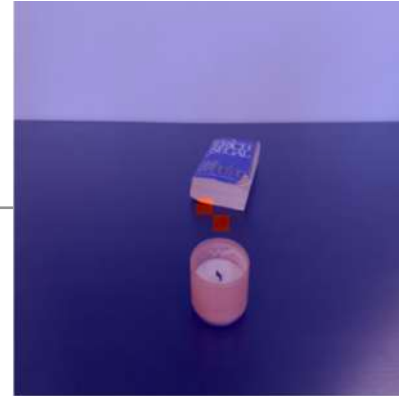
Where is the book in
relation to the candle?



VLM
(e.g.LLaVA)



Attention on image



✗ Incorrect focus!



Where is the cat in relation
to the person?



Attention on image



✗ Insufficient focus!

We intervene attention's distribution **adaptively** given **model's self-confidence**!

Where is the book in relation to the candle?



Unfamiliar!

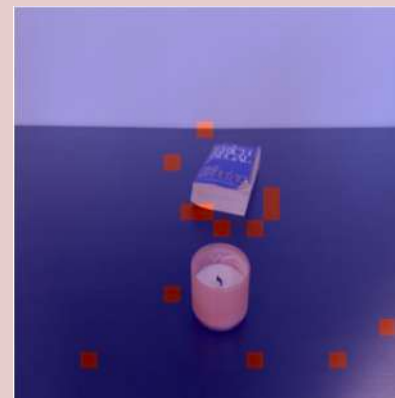
Left

Confidence: **0.27**

low confidence

Smoother!

Look at other patches!



Behind



high confidence

Above

Confidence: **0.38**

Familiar!

Sharper!

Look more closely!



Right



Where is the cat in relation to the person?

Results

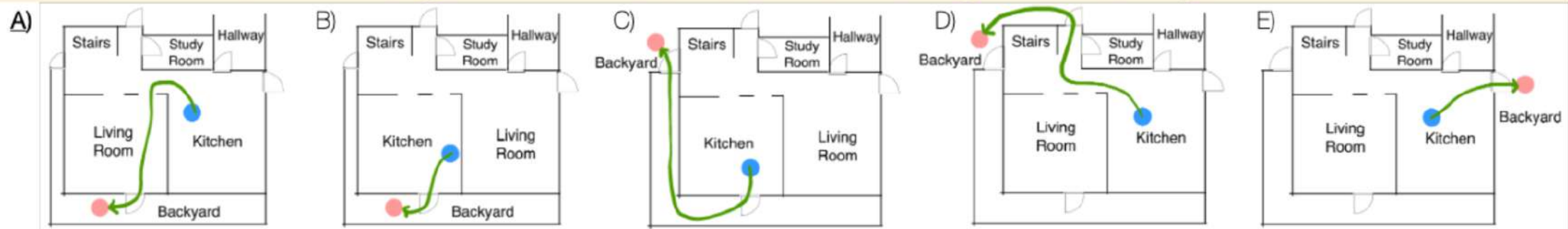
Model	Controlled_A			Controlled_B		
	Acc	Pair Acc	Set Acc	Acc	Pair Acc	Set Acc
LLaVA-1.5	60.3	40.6	0.0	73.1	41.6	3.7
+VCD	61.5 $\uparrow 1.2$	39.4 $\downarrow 1.2$	0.0	73.4 $\uparrow 0.3$	42.2 $\uparrow 0.6$	3.7
+Dola	61.2 $\uparrow 0.9$	41.6 $\uparrow 1.0$	0.0	73.4 $\uparrow 0.3$	42.2 $\uparrow 0.6$	3.7
+SCALINGVIS	64.5 $\uparrow 4.2$	40.6	0.0	75.2 $\uparrow 2.1$	44.6 $\uparrow 3.0$	9.8 $\uparrow 6.1$
+ADAPTVIS	84.9 $\uparrow 24.6$	61.2 $\uparrow 20.6$	30.3 $\uparrow 30.3$	83.8 $\uparrow 10.7$	55.7 $\uparrow 14.1$	18.3 $\uparrow 14.6$
LLaVA-1.6	48.2	37.6	0.0	63.0	39.1	3.7
+VCD	61.8 $\uparrow 13.6$	41.8 $\uparrow 4.2$	10.9 $\uparrow 10.9$	65.4 $\uparrow 2.4$	41.6 $\uparrow 2.5$	7.3 $\uparrow 3.6$
+Dola	48.2	37.6	0.0	62.7 $\downarrow 0.3$	39.1	3.7
+SCALINGVIS	97.0 $\uparrow 48.8$	76.4 $\uparrow 38.8$	54.5 $\uparrow 54.5$	73.4 $\uparrow 10.4$	48.9 $\uparrow 9.8$	15.9 $\uparrow 12.2$
+ADAPTVIS	98.2 $\uparrow 50.0$	78.8 $\uparrow 41.2$	57.0 $\uparrow 57.0$	73.4 $\uparrow 10.4$	48.9 $\uparrow 9.8$	15.9 $\uparrow 12.2$

HourVideo

1-Hour Video-Language Understanding



Room-to-Room Navigation: How can the camera wearer get to the backyard from the kitchen?



Object Retrieval: How can the camera wearer retrieve the motorcycle from the kitchen?

- A) Exit the kitchen towards the stairs and exit through the door. The motorbike is outside.
- B) Exit the kitchen towards the stairs and exit through the door. The motorbike is outside.
- C) Exit the kitchen and turn left. Walk through the living room and go through the door into the backyard. The motorbike is on the right.
- D) Exit the kitchen to the living room and turn left. Go through the door to the backyard; the motorbike is on the right.
- E) Exit the kitchen and turn left. Walk down the hallway and turn right before the stairs. Exit the door and the motorbike is outside.

LongVideo-Haystack

<https://huggingface.co/datasets/LVHaystack/LongVideoHaystack>

Re-thinking Temporal Search for Long-Form Video Understanding

Jinhui Ye^{1*†}, Zihan Wang^{2*}, Haosen Sun², Keshigeyan Chandrasegaran¹,
Zane Durante¹, Cristobal Eyzaguirre¹, Yonatan Bisk³, Juan Carlos Niebles¹, Ehsan Adeli¹,
Li Fei-Fei¹, Jiajun Wu¹, Manling Li^{1,2}

¹Stanford University ²Northwestern University ³Carnegie Mellon University

<https://longvideohaystack.github.io>

CVPR 25: LV-Haystack, T* for Temporal Search

4 4 3 (work w/ Keshik)

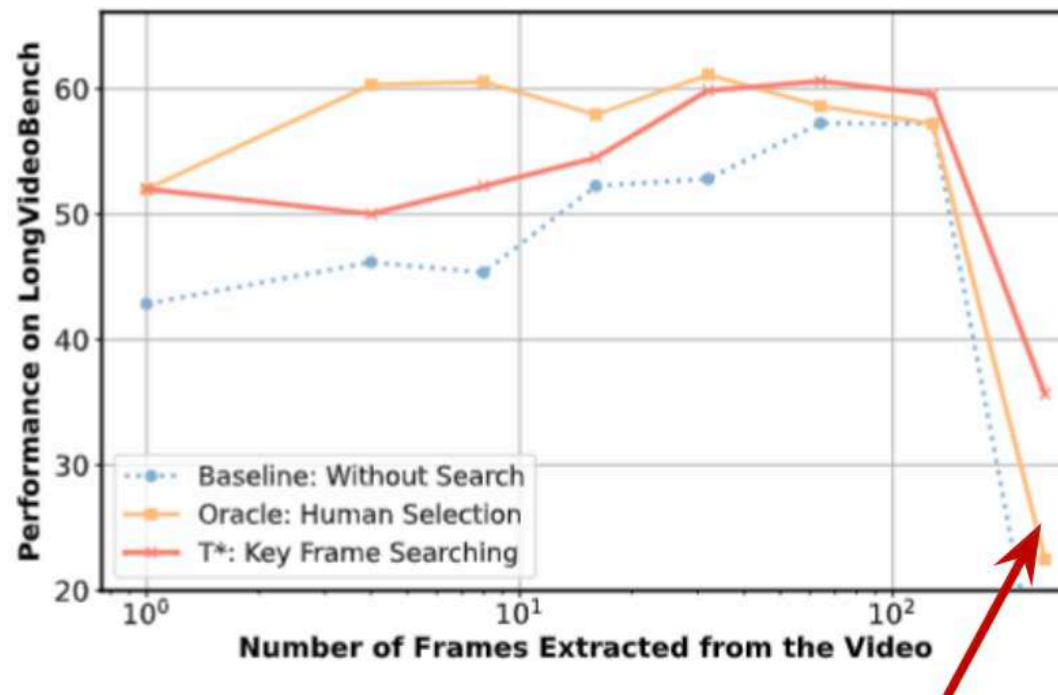


Figure 5. Performance improvement with increasing search frames.

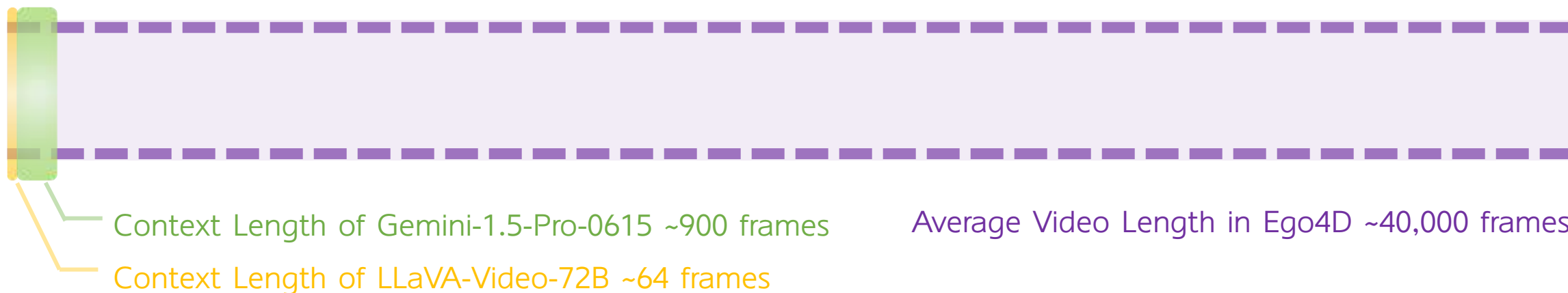
Medium-sized models (7b) **struggle**
to handle more than 100 frames



Video

Question: Where was the white trash can before I raised it?

Needle in the Long Video Haystack



Needle in the Long Video Haystack



 **Datasets:**  LVHaystack / **LongVideoHaystack**

Haystack-Ego4D

988 videos

432 hours

15,092 QA pairs

23,800 frames

Haystack-LVBench

246 videos

57.7 hours

602 QA pairs

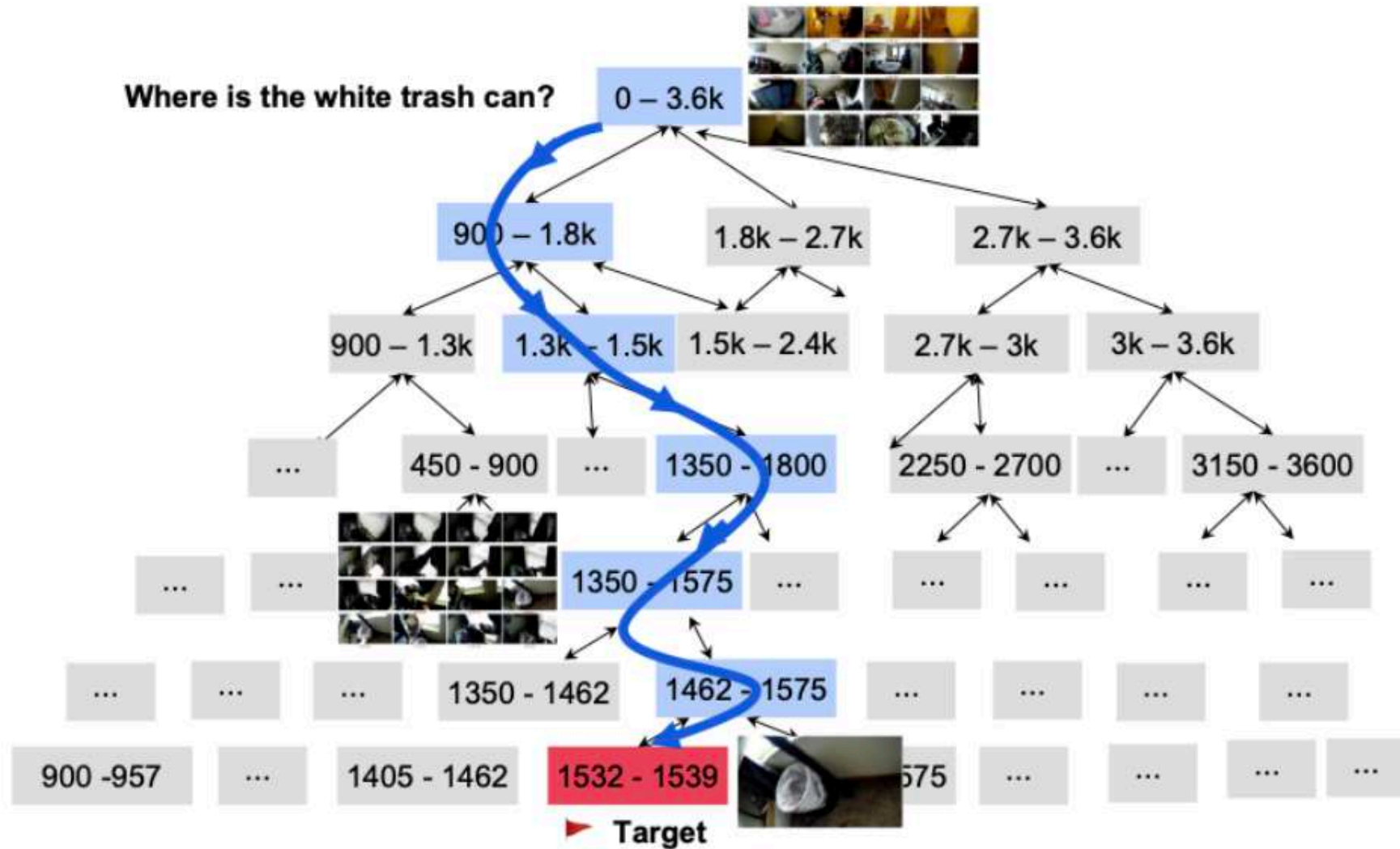
1,070 frames

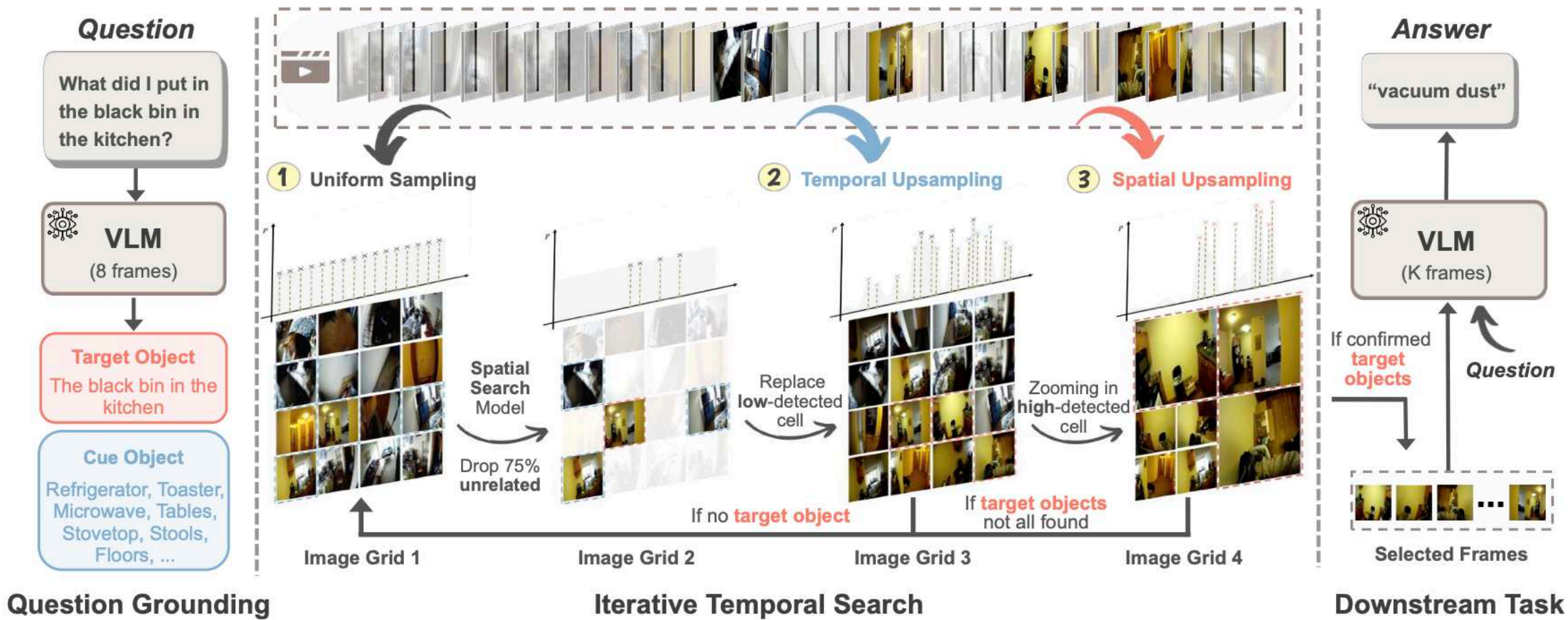
Needle in the Long Video Haystack

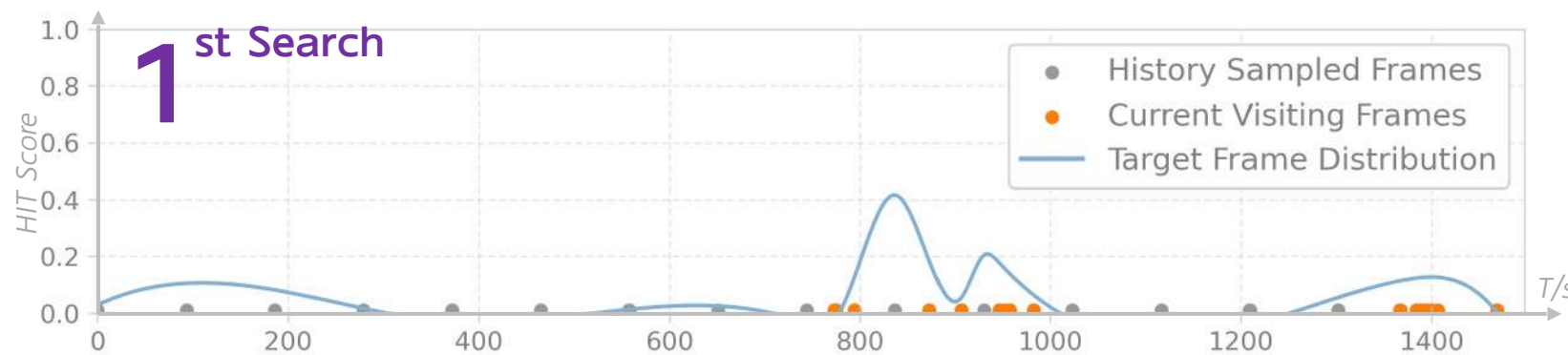
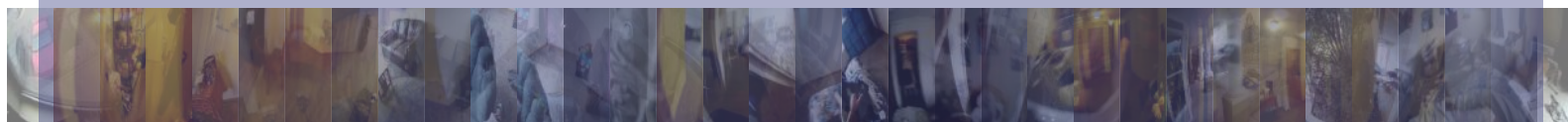
T^*

a light-weighted plug-in for **temporal searching**

T*: Temporal Searching







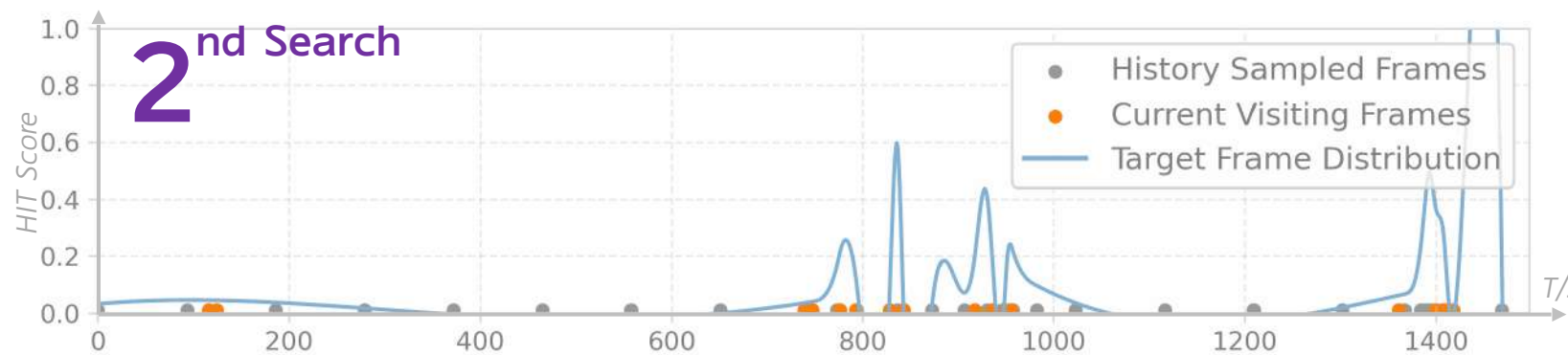
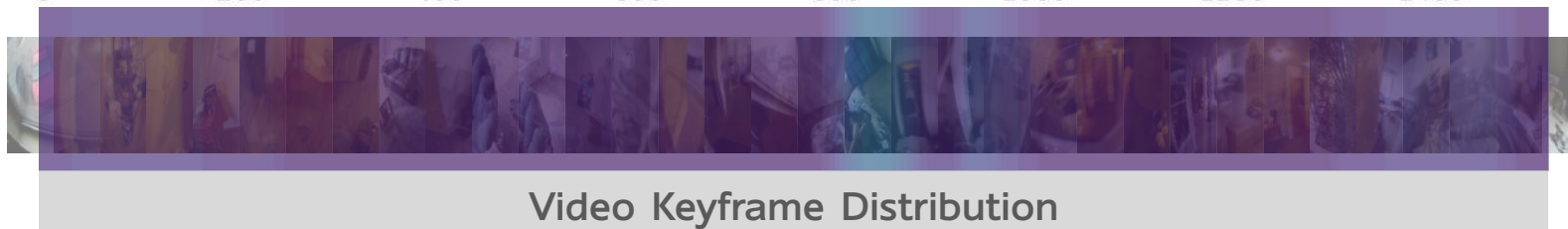
Video Keyframe Distribution

Frame
Sampling

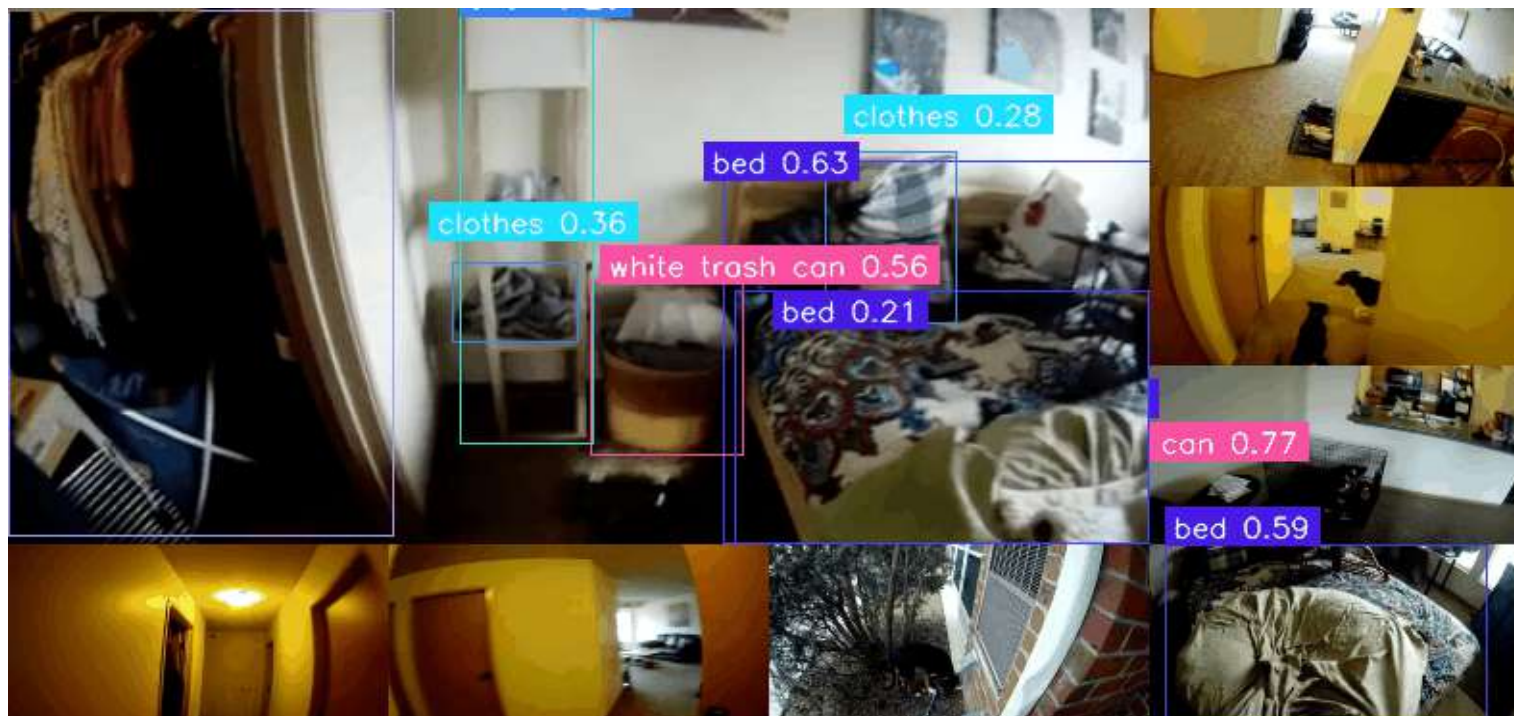


Sampled Frames

Distribution
Updating

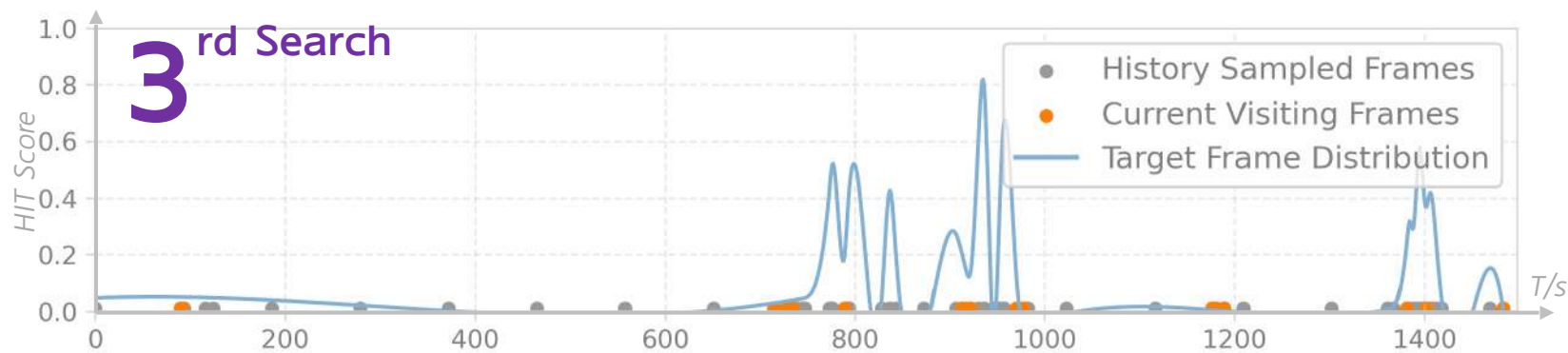
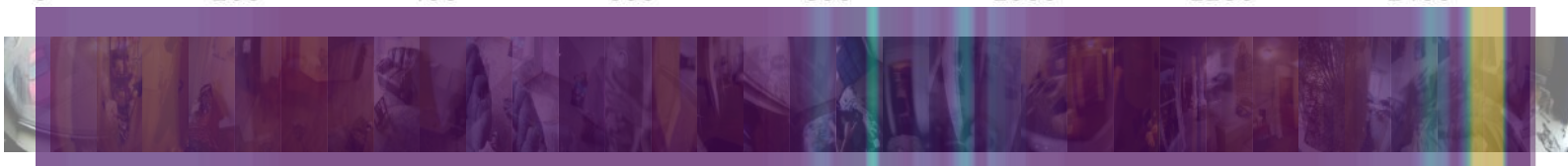


Frame
Sampling



Sampled Frames

Distribution
Updating



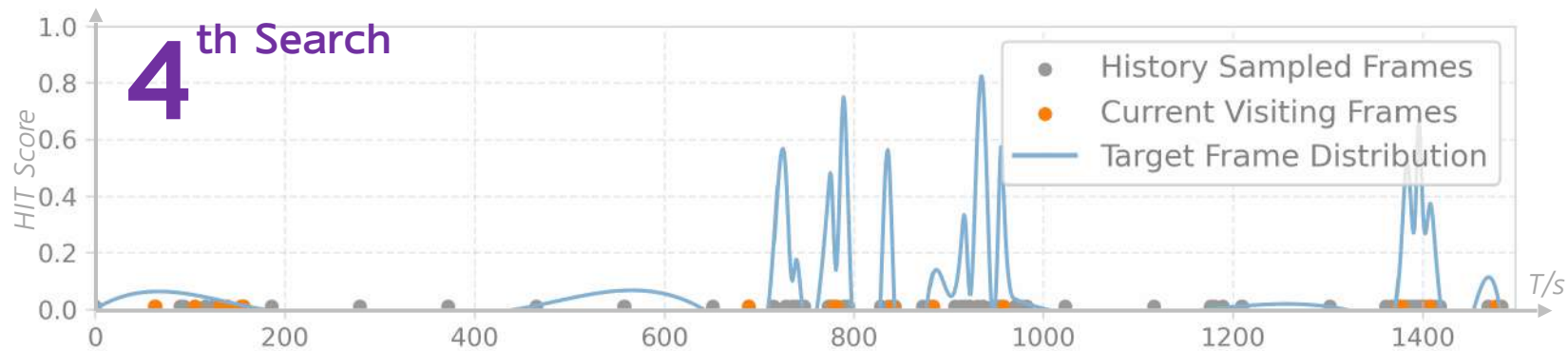
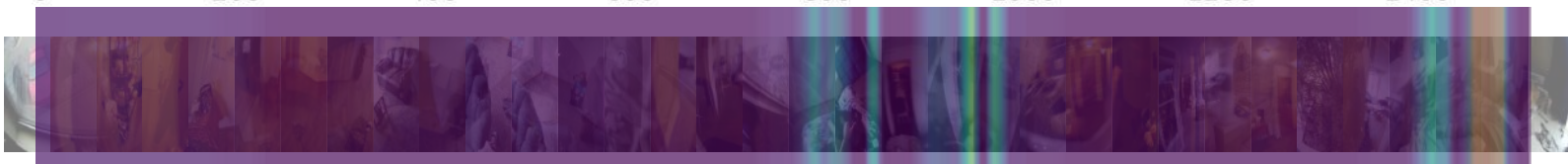
Video Keyframe Distribution

Frame
Sampling



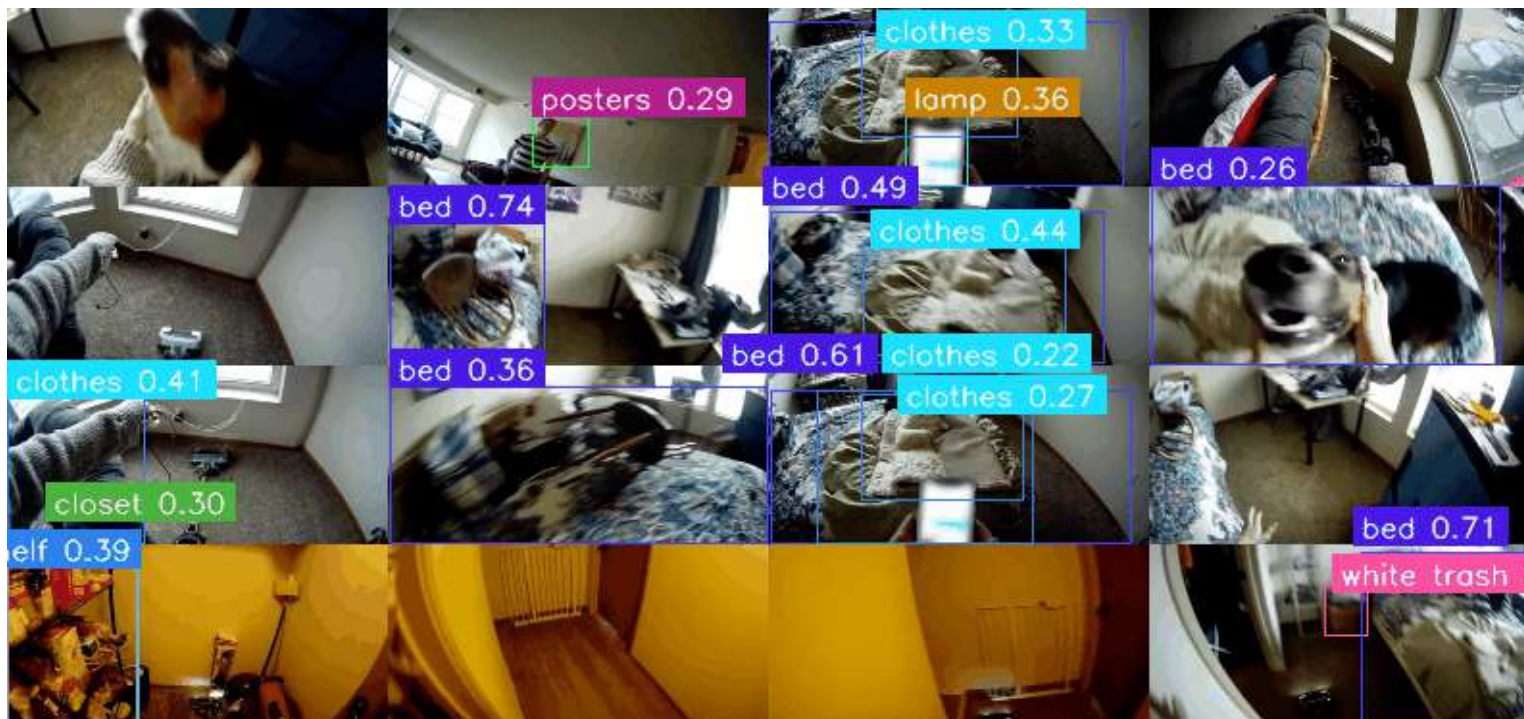
Sampled Frames

Distribution
Updating



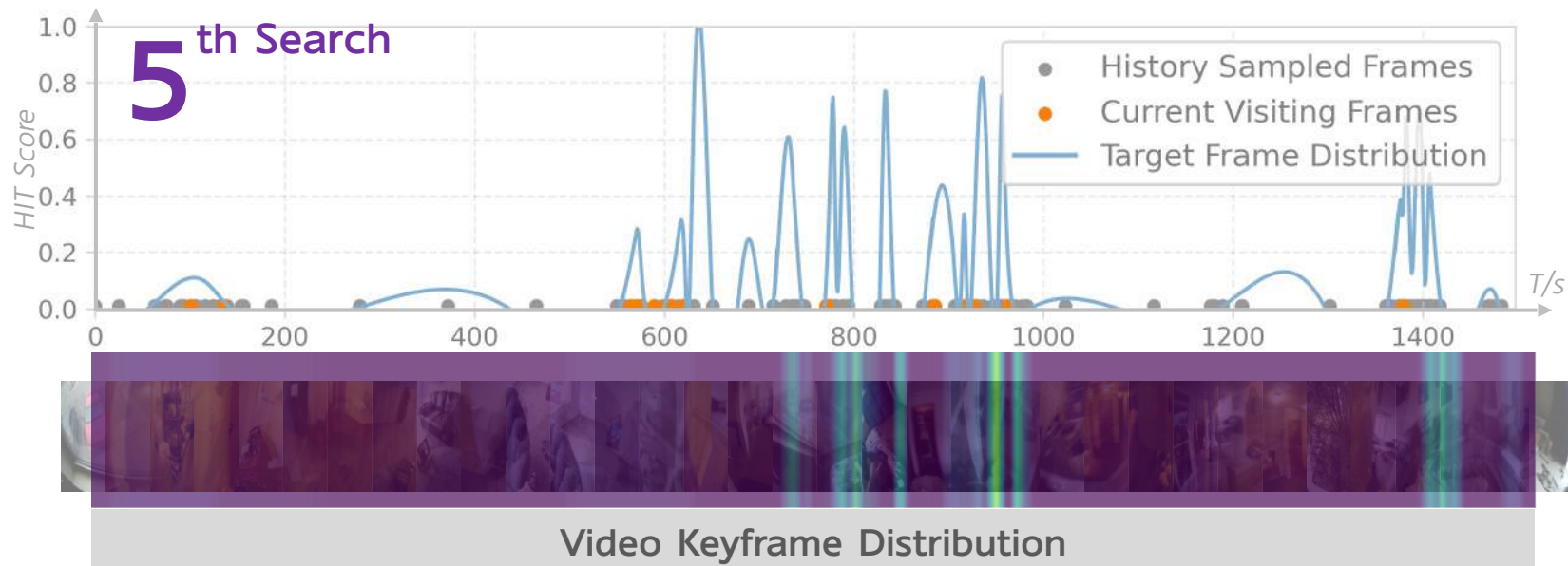
Video Keyframe Distribution

Frame
Sampling



Sampled Frames

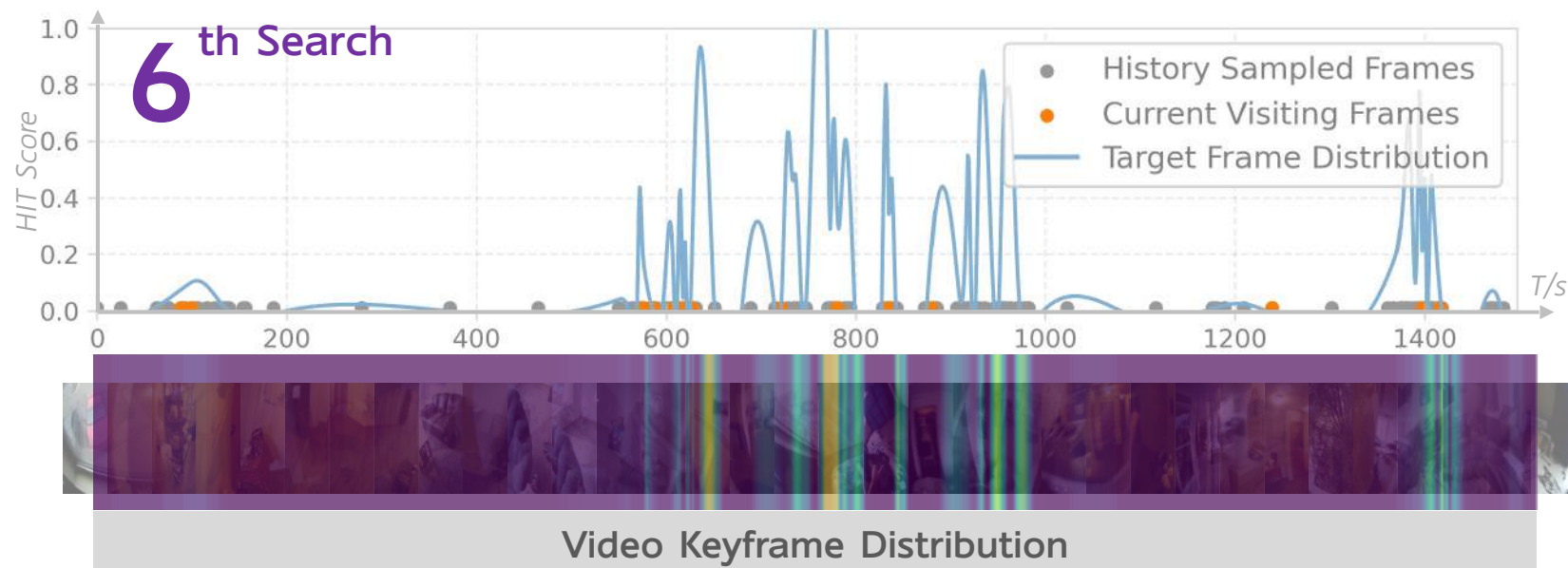
Distribution
Updating



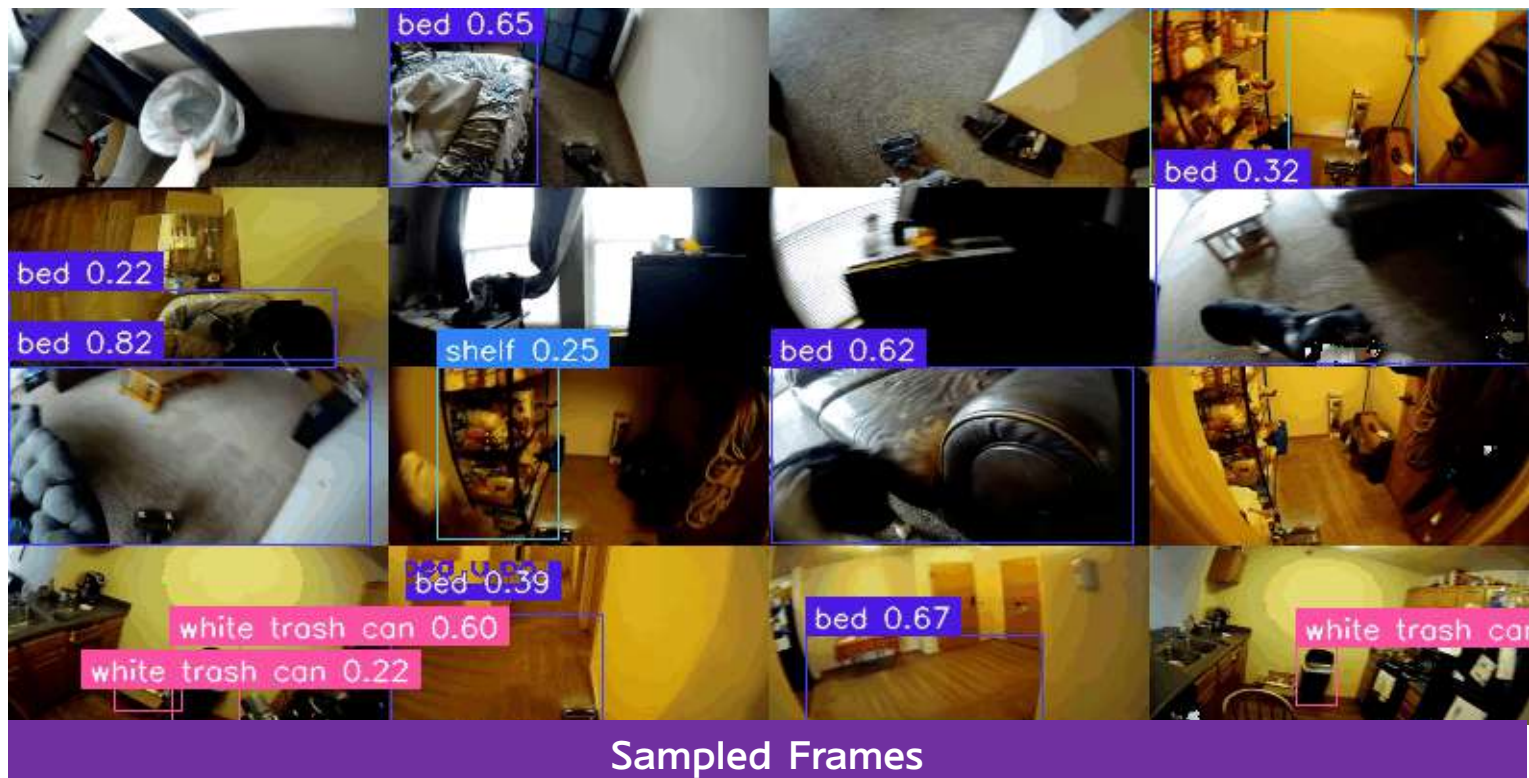
Frame
Sampling



Distribution
Updating

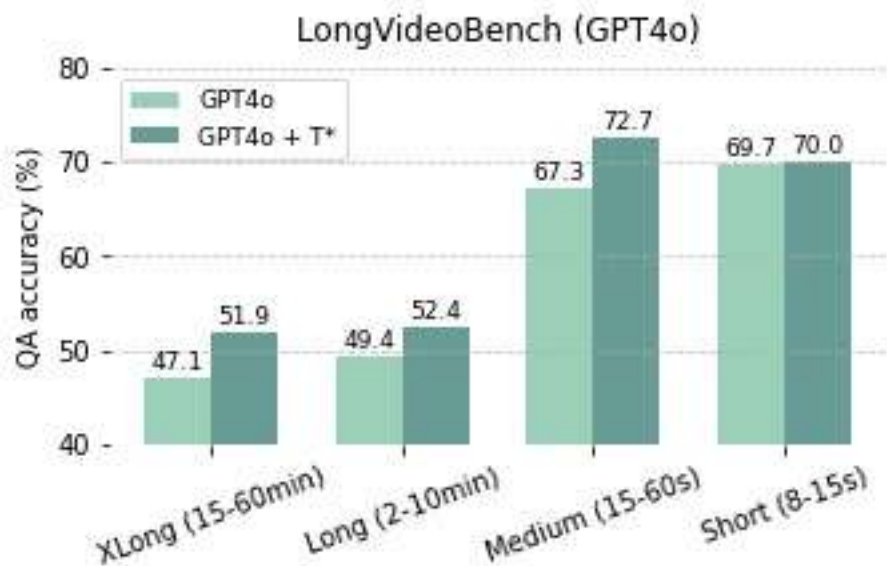
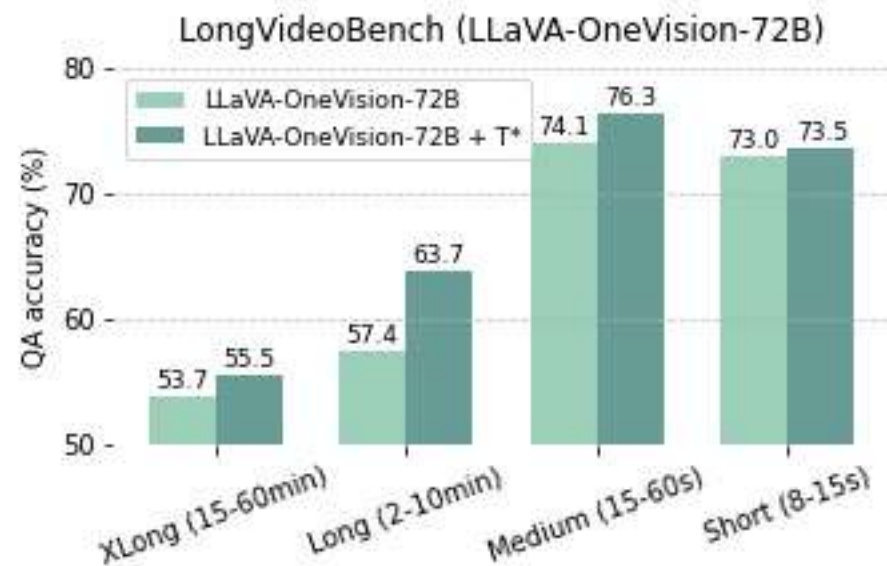
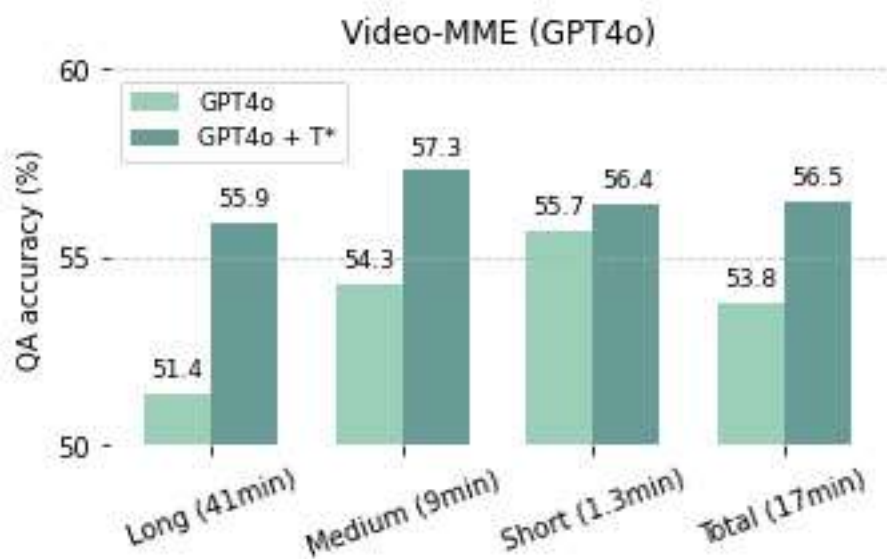
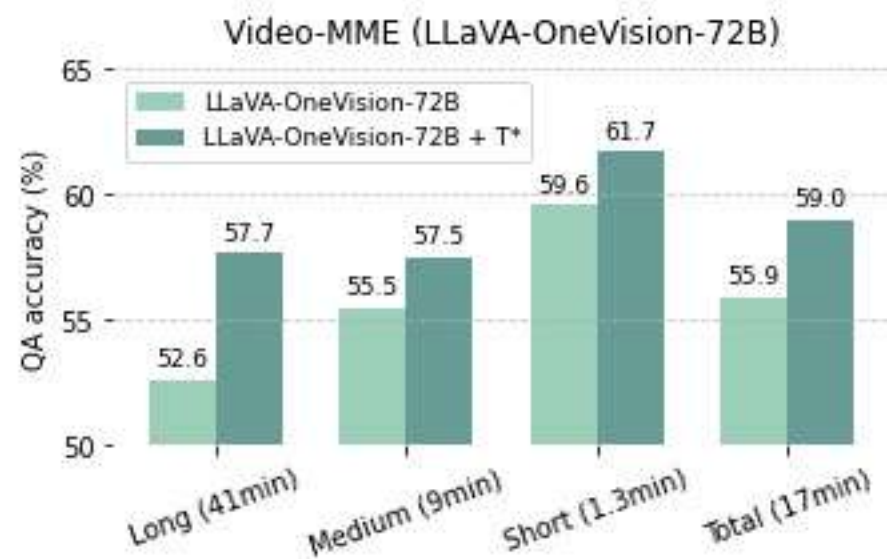


Frame
Sampling

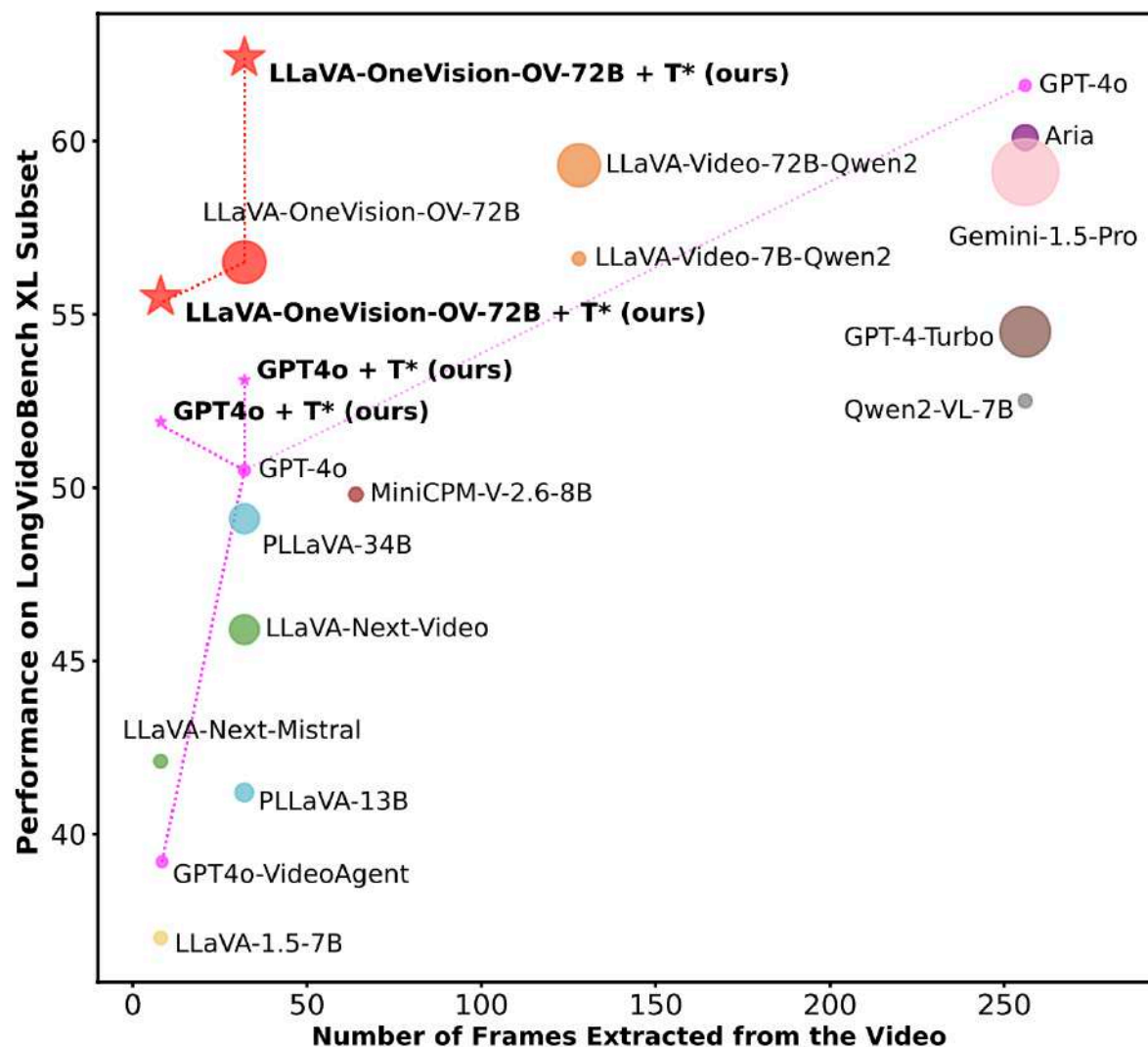


Distribution
Updating

T*: Temporal Searching



T* for Temporal Search



Re-thinking Temporal Search for Long-Form Video Understanding

Jinhui Ye^{1*}

Zihan Wang^{2*}

Haosen Sun²

Keshigeyan Chandrasegaran¹

Zane Durante¹

Cristobal Eyzaguirre¹

Yonatan Bisk³

Juan Carlos Niebles¹

Ehsan Adeli¹

Li Fei-Fei¹

Jiajun Wu¹

Manling Li²

¹Stanford University ²Northwestern University ³Carnegie Mellon University

<https://longvideohaystack.github.io>



LVHaystack

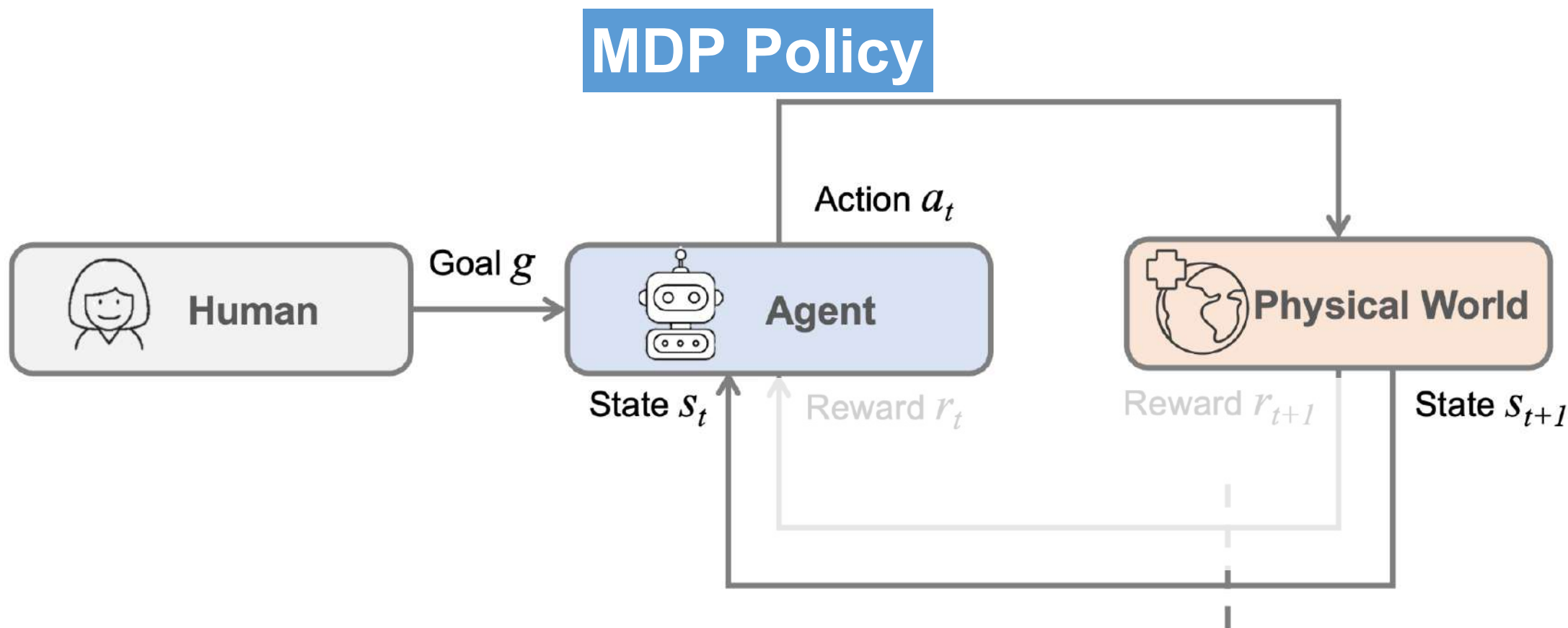


T*



Live Demo

Let us go back to MDPs (Markov Decision Processes)



Policy

$$\pi(o, g) \rightarrow a$$



RAGEN:

Training Agents by Reinforcing Reasoning

Zihan Wang*, Kangrui Wang*, Qineng Wang*, Pingyue Zhang*, Linjie Li*,

Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Monica Lam, Yiping Lu,

Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, Manling Li

Key idea: MDP as sequence prediction



Goal

State

Action

Reward

Why multi-turn RL + LLM Agents?



Memory Across Turns



Stochastic Environments

Why multi-turn RL + LLM Agents?

Other Methods

Static Tasks



Math

Single-turn

Non-Stochastic



Code

Single-turn

Non-Stochastic

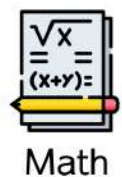
Why multi-turn RL + LLM Agents?

Other Methods

SFT



Static Tasks

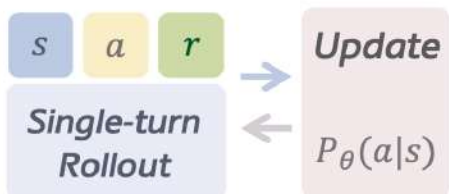


Single-turn
Non-Stochastic



Single-turn
Non-Stochastic

Single-turn RL



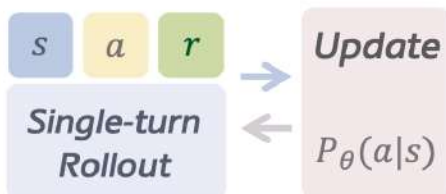
Why multi-turn RL + LLM Agents?

Other Methods

SFT



Single-turn RL



Static Tasks



RAGEN (Ours)

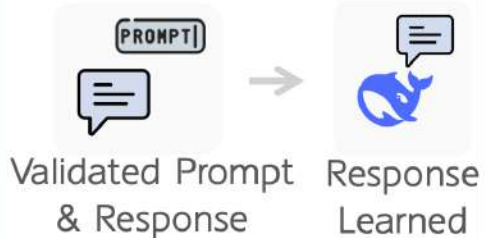
Dynamic Tasks



Why multi-turn RL + LLM Agents?

Other Methods

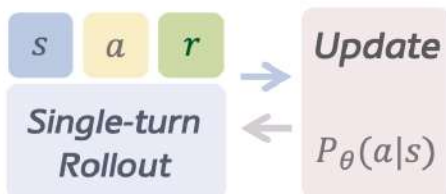
SFT



Static Tasks

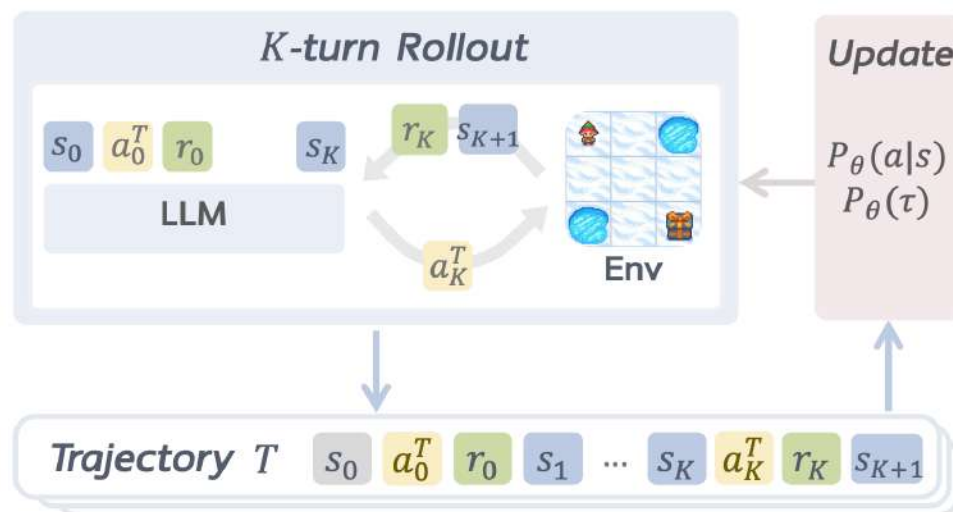


Single-turn RL



RAGEN (Ours)

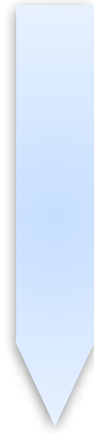
StarPO (State-Thinking-Actions-Reward Policy Optimization)



Dynamic Tasks



Optimizing each single action



Optimizing entire multi-turn trajectory

Optimizing entire multi-turn trajectory

Objective Function:

$$J_{\text{StarPO}}(\theta) = E_{M, \tau \sim \pi_{\theta}}[R(\tau)]$$

where τ is a full trajectory and $R(\tau)$ is the cumulative reward over the trajectory.

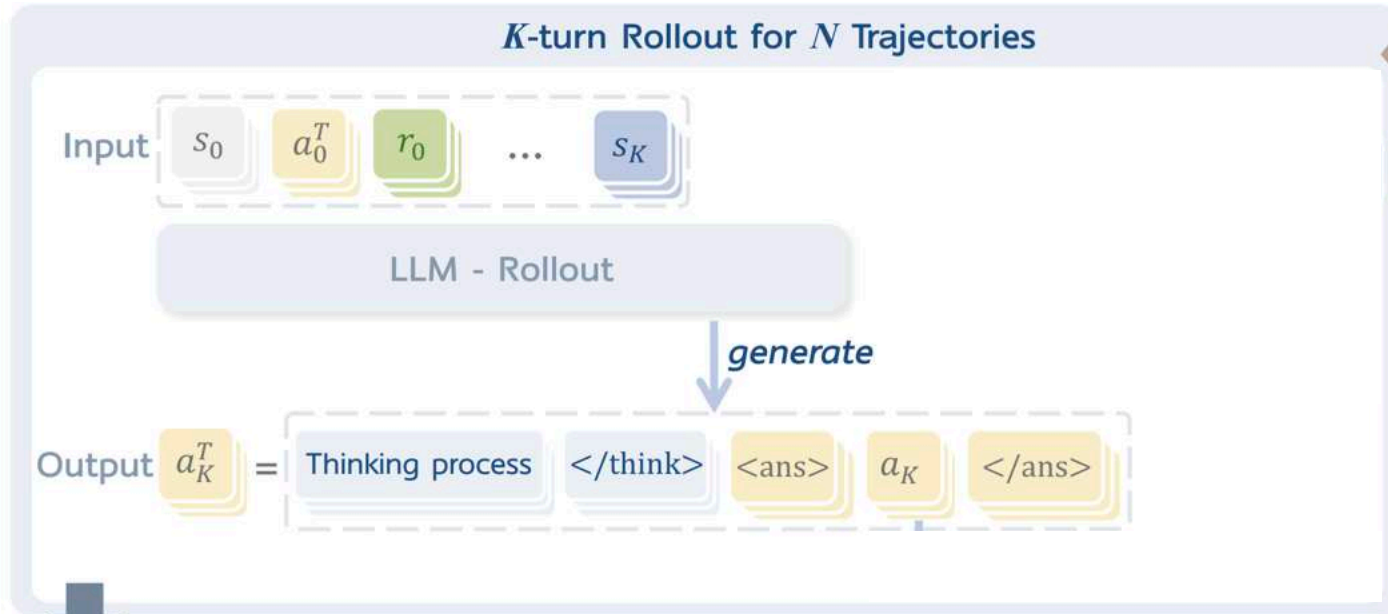
StarPO: State-Thinking-Action-Reward Policy Optimization



1 Rollout Generation

Agent generates multiple trajectories (N) starting from initial states (P).

StarPO: State-Thinking-Action-Reward Policy Optimization



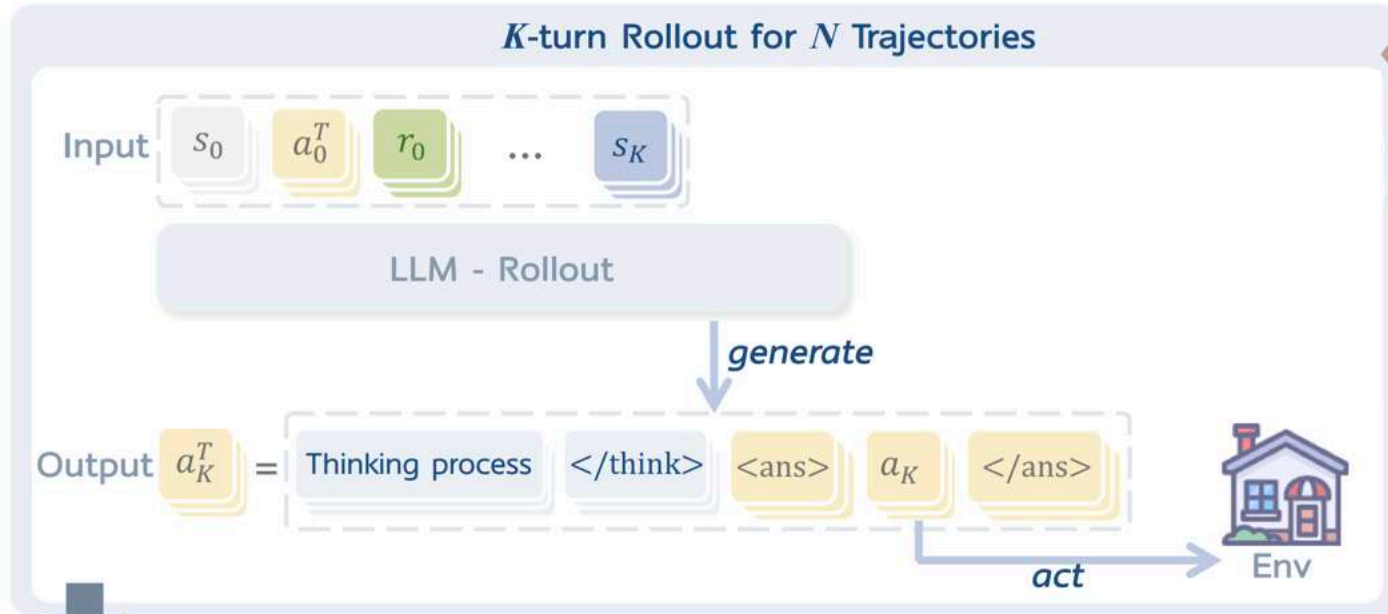
2

Structured Output

Each action a_t^T includes reasoning and executable action a_t :

$$a_t^T = \langle \text{think} \rangle \dots \langle \text{answer} \rangle a_t \langle \text{answer} \rangle$$

StarPO: State-Thinking-Action-Reward Policy Optimization



1 Rollout Generation

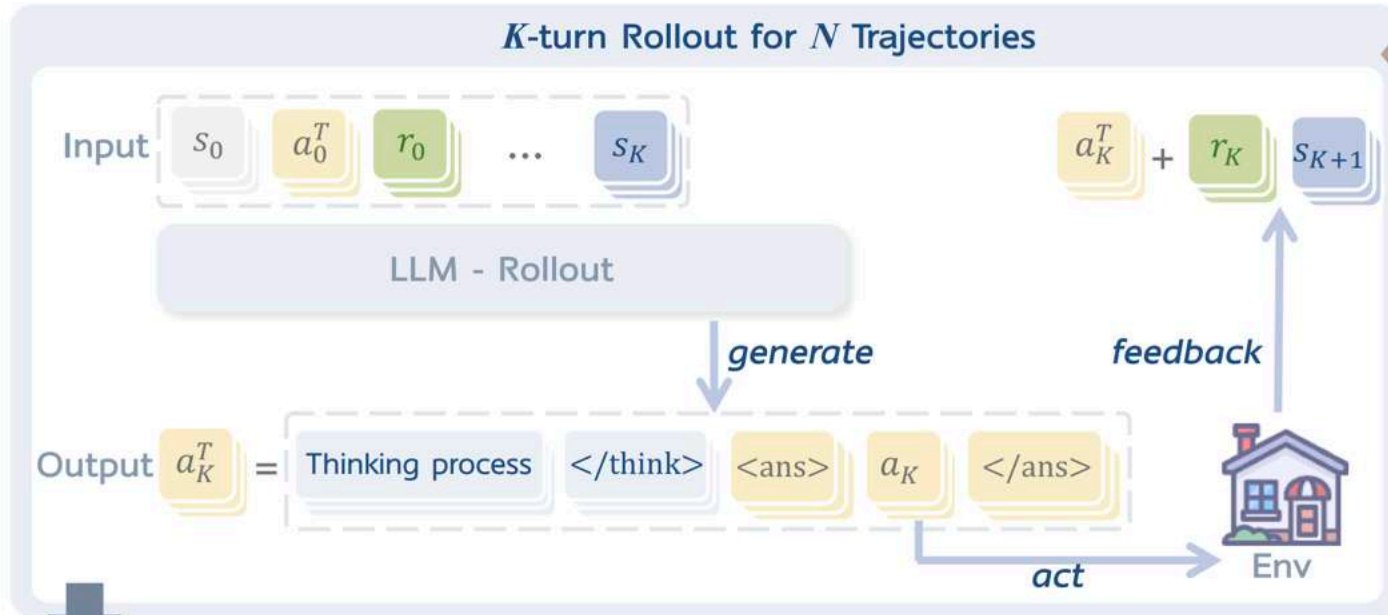
Agent generates multiple trajectories (N) starting from initial states (P).

2 Structured Output

Each action a_t^T includes reasoning and executable action a_t :

$$a_t^T = \langle \text{think} \rangle \dots \langle \text{answer} \rangle a_t \langle \text{answer} \rangle$$

StarPO: State-Thinking-Action-Reward Policy Optimization



1 Rollout Generation

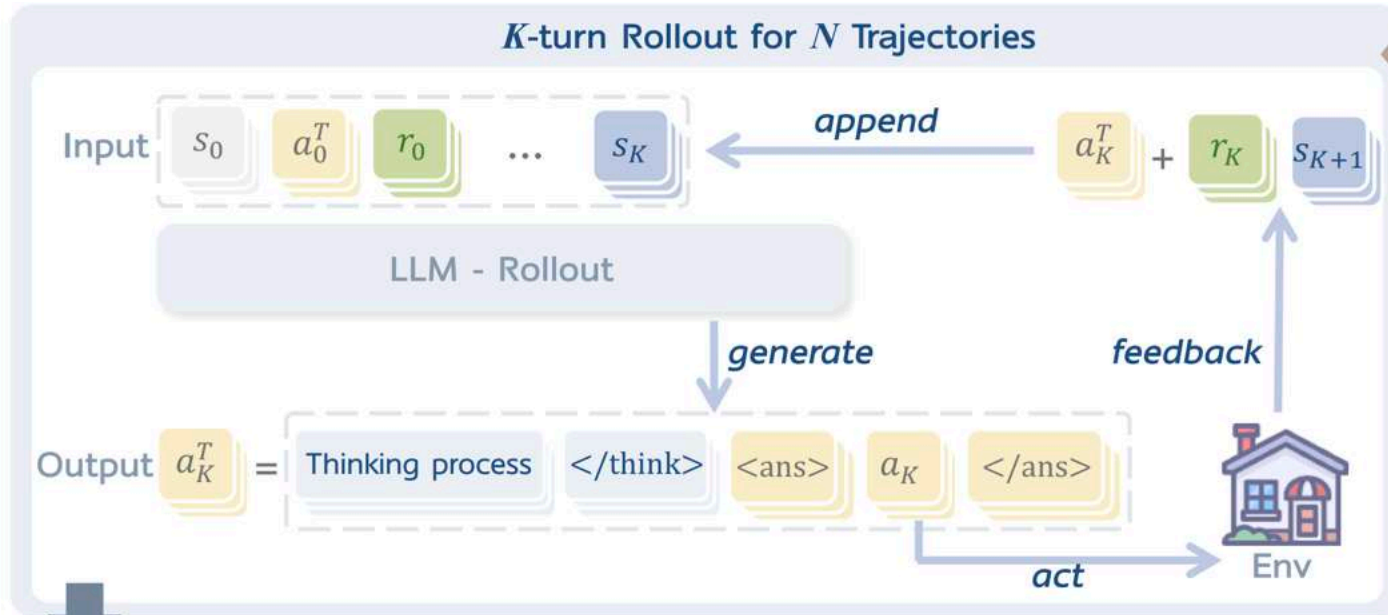
Agent generates multiple trajectories (N) starting from initial states (P).

2 Structured Output

Each action a_t^T includes reasoning and executable action a_t :

$$a_t^T = \langle \text{think} \rangle \dots \langle \text{answer} \rangle a_t \langle \text{answer} \rangle$$

StarPO: State-Thinking-Action-Reward Policy Optimization



1 Rollout Generation

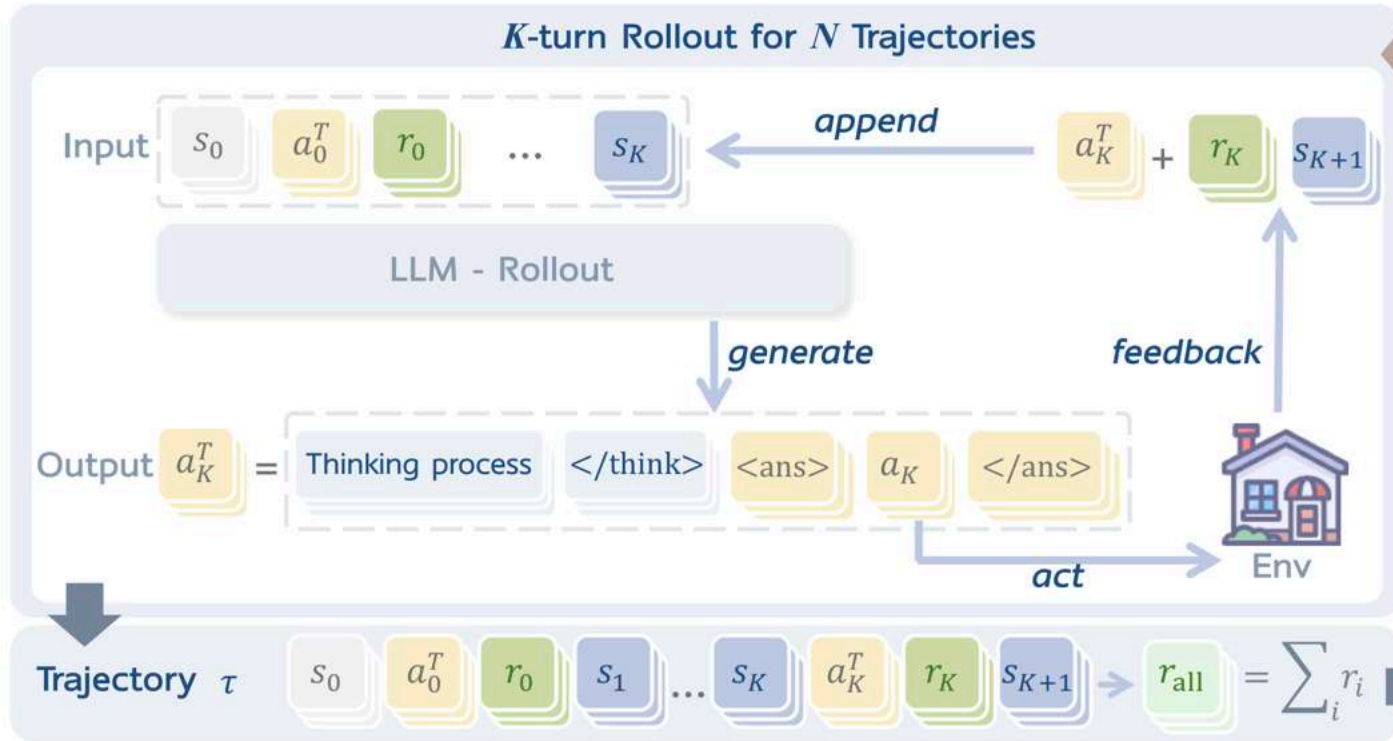
Agent generates multiple trajectories (N) starting from initial states (P).

2 Structured Output

Each action a_t^T includes reasoning and executable action a_t :

$$a_t^T = \langle \text{think} \rangle \dots \langle \text{answer} \rangle a_t \langle \text{answer} \rangle$$

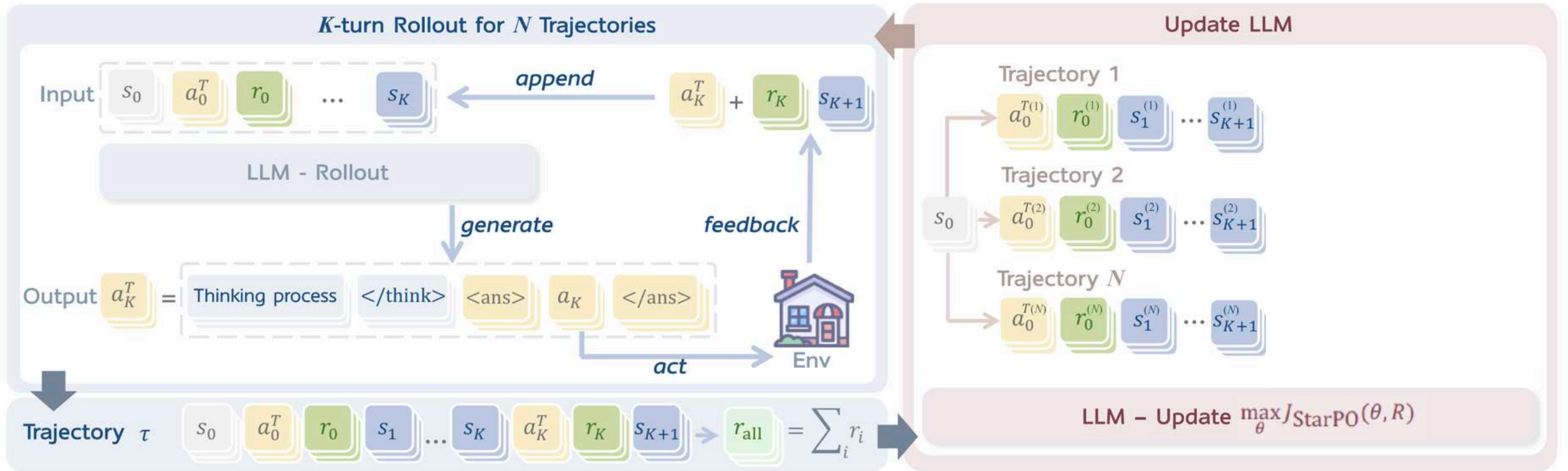
StarPO: State-Thinking-Action-Reward Policy Optimization



3 Trajectory

Full sequence of interactions $\tau = \{s_0, a_0^T, r_0, s_1, \dots, s_K\}$.

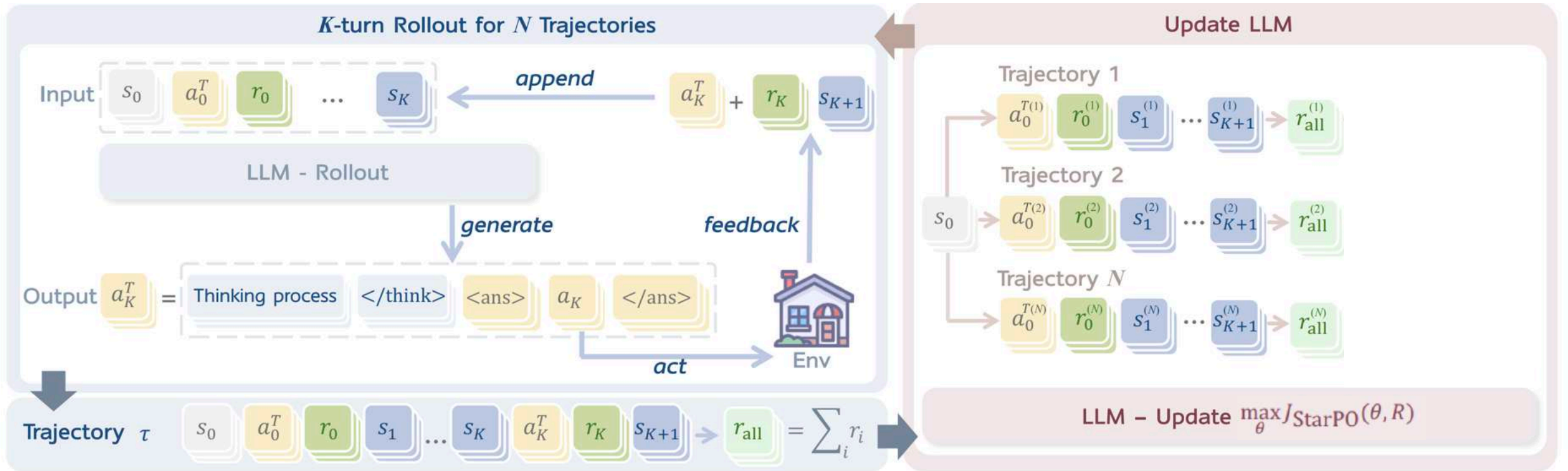
StarPO: State-Thinking-Action-Reward Policy Optimization



4 Interleaving

Rollouts and updates are performed iteratively.

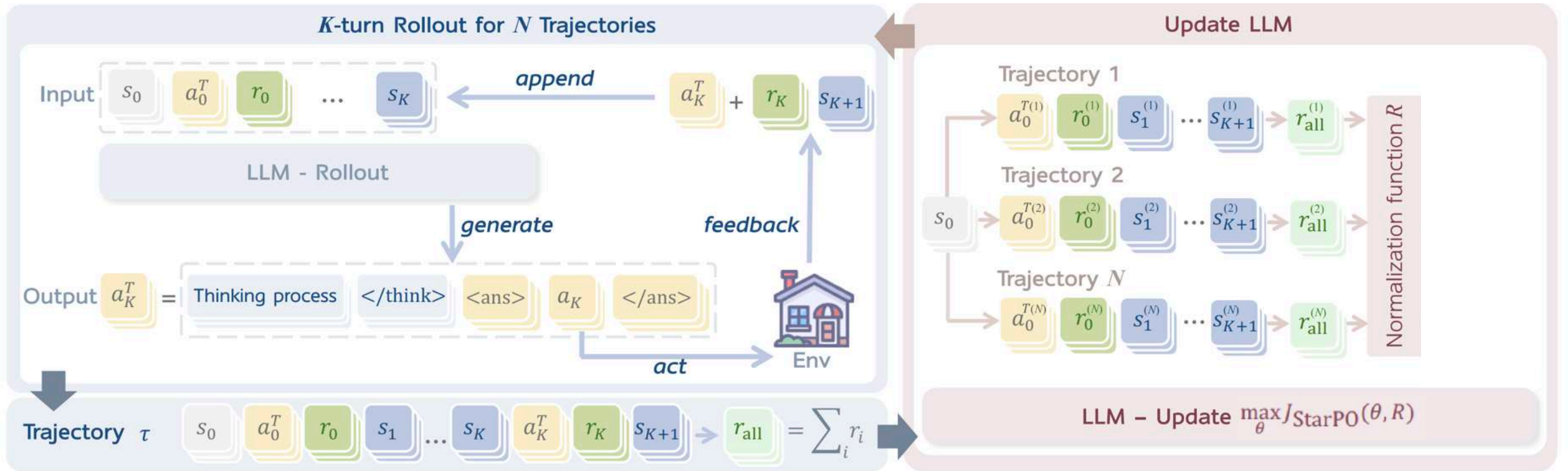
StarPO: State-Thinking-Action-Reward Policy Optimization



4 Interleaving

Rollouts and updates are performed iteratively.

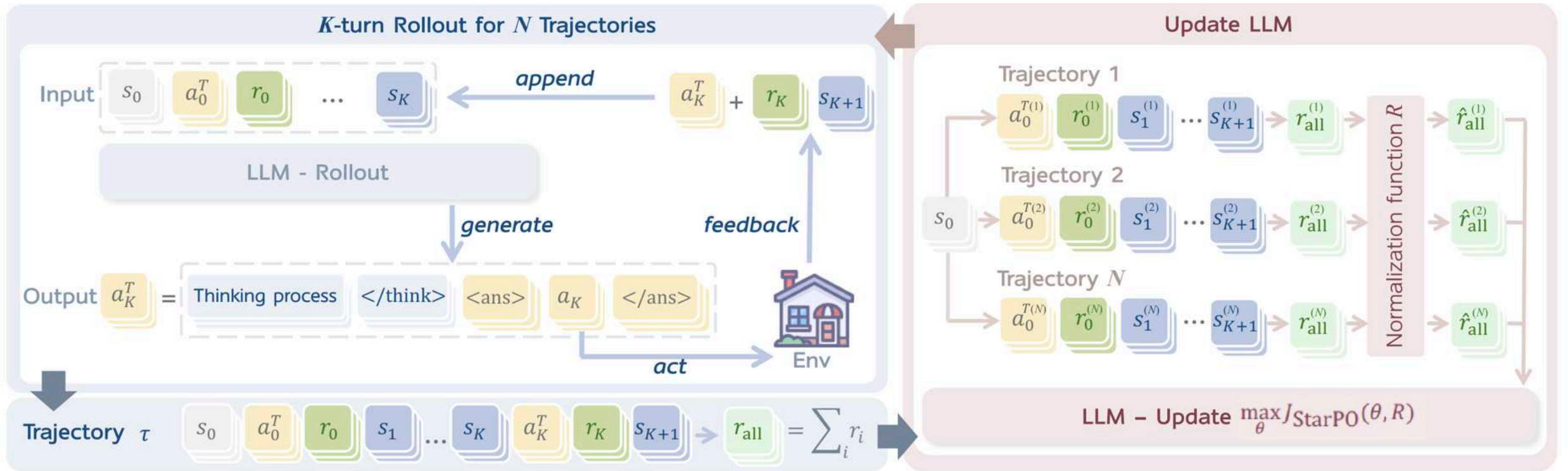
StarPO: State-Thinking-Action-Reward Policy Optimization



4 Interleaving

Rollouts and updates are performed iteratively.

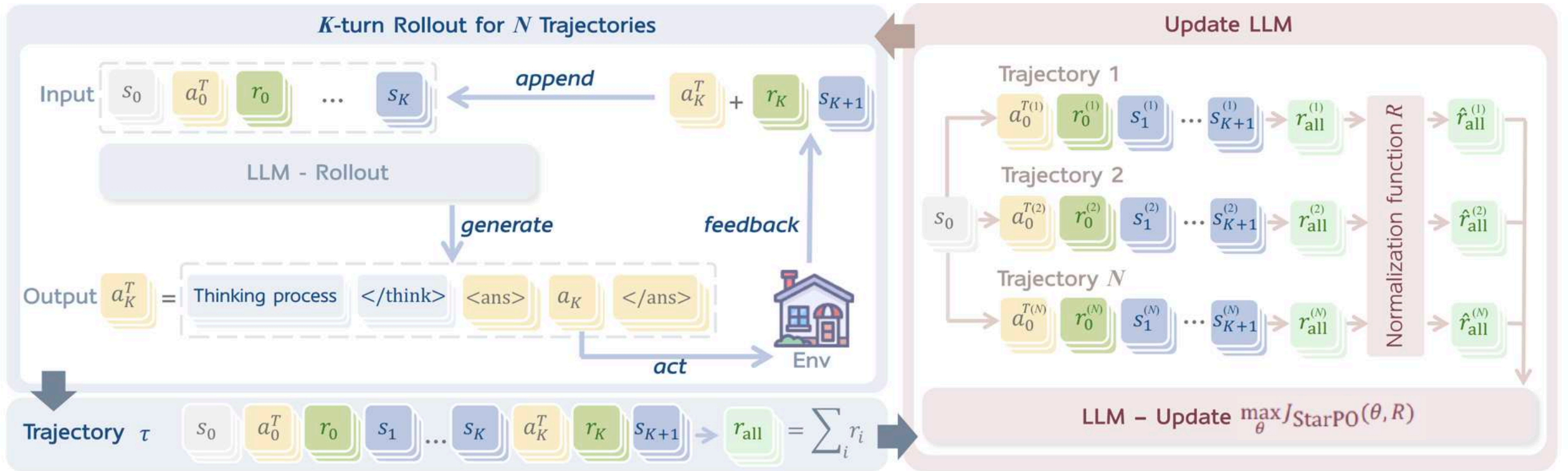
StarPO: State-Thinking-Action-Reward Policy Optimization



4 Interleaving

Rollouts and updates are performed iteratively.

StarPO: State-Thinking-Action-Reward Policy Optimization



4 Interleaving

Rollouts and updates are performed iteratively.

Researching why multi-turn RL collapse?

We use a **Controlled Environment**



Single-turn
Stochastic

Bandit



Sokoban

Multi-turn
Non-Stochastic



Frozen Lake

Multi-turn
Stochastic

Researching why multi-turn RL collapse?

We use a **Controlled Environment**



Single-turn
Stochastic



Sokoban

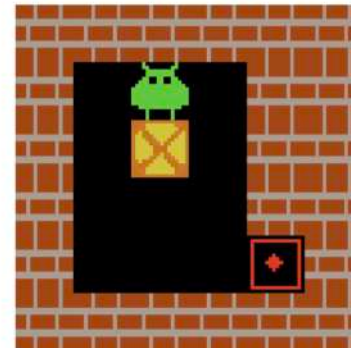
Multi-turn
Non-Stochastic



Frozen Lake

Multi-turn
Stochastic

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



Experiment Settings



Model & Framework

- Model: Qwen-2.5 (0.5B) LLM
- Framework: StarPO (PPO & GRPO variants)



Hardware & Iterations

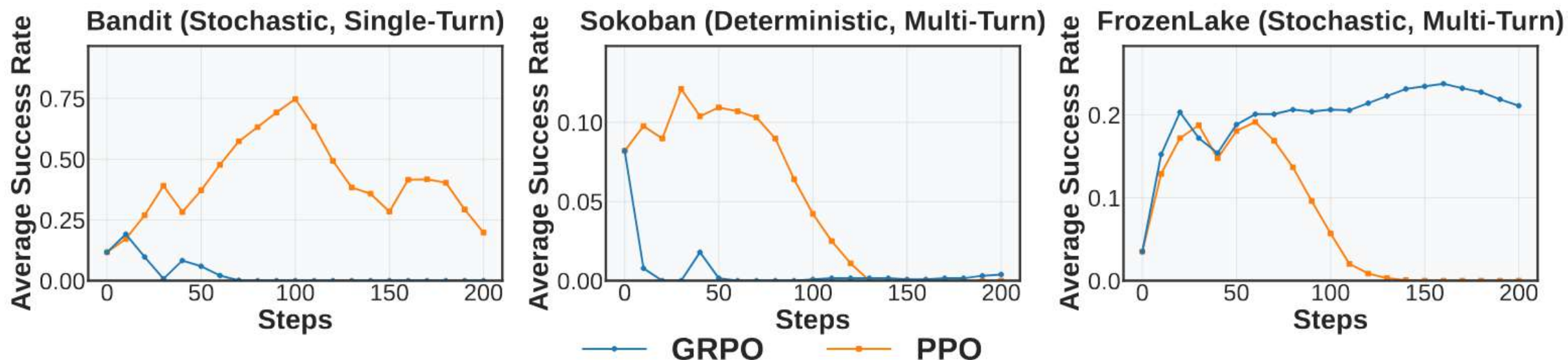
- Hardware: H100 GPUs
- Iterations: 200 rollout-update cycles



Batch & Trajectory

- Batch: $P = 8$ prompts, $N = 16$ rollouts/prompt
- Trajectory: Max 5 turns, 10 actions/turn

Finding 1: Instability in Multi-Turn RL



Different from Single-Turn

PPO > GRPO: critic help resist early-stage degradation

Finding 2: The "Echo Trap" Failure Mode

Step	Model Output
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.

What happens when collapse? **Repeating!**

Finding 2: The "Echo Trap" Failure Mode



Model Behavior Evolution

Early-stage:

Diverse reasoning (symbolic, exploratory).

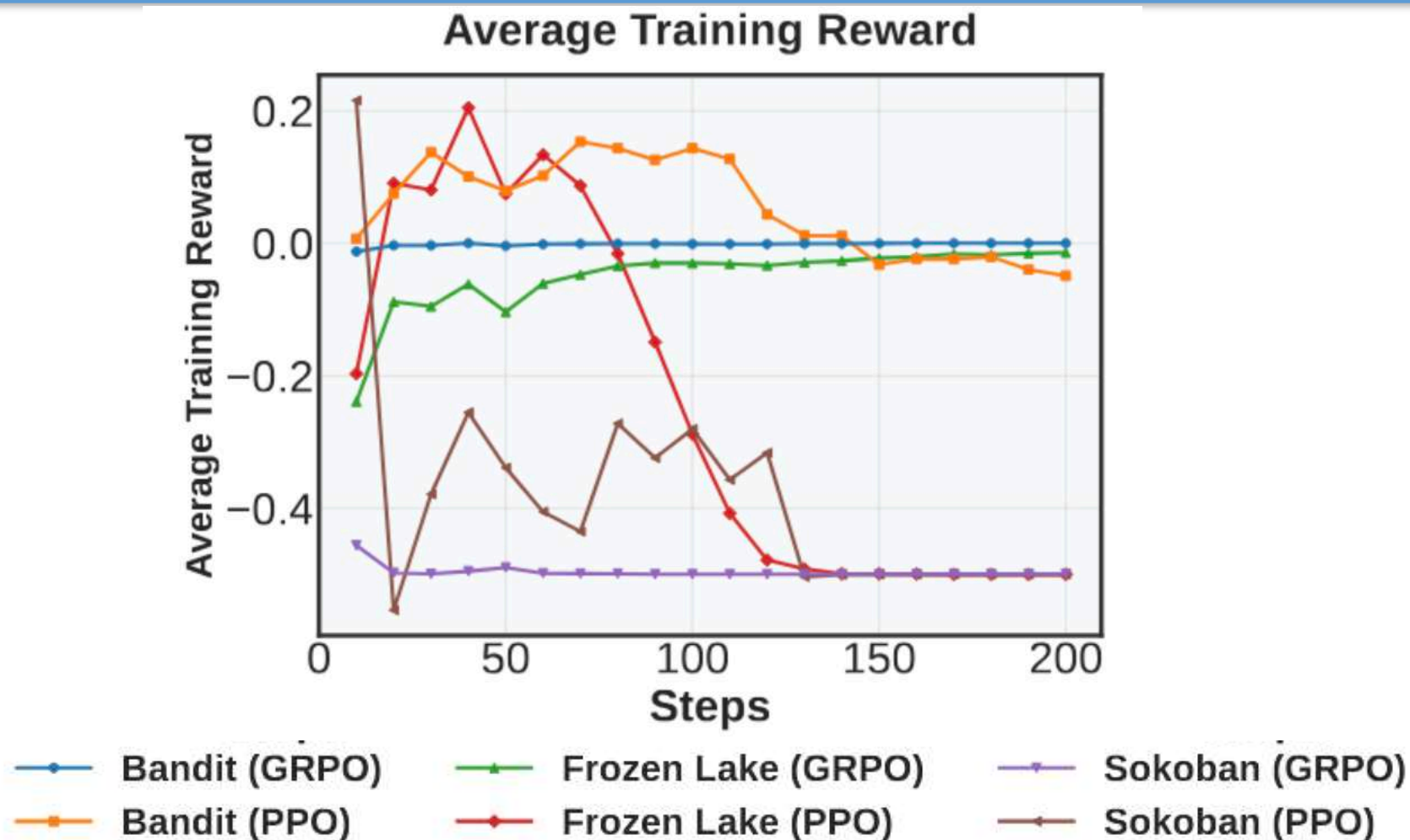
Late-stage:

Repetitive, deterministic templates (e.g., "choosing Dragon" without justification).

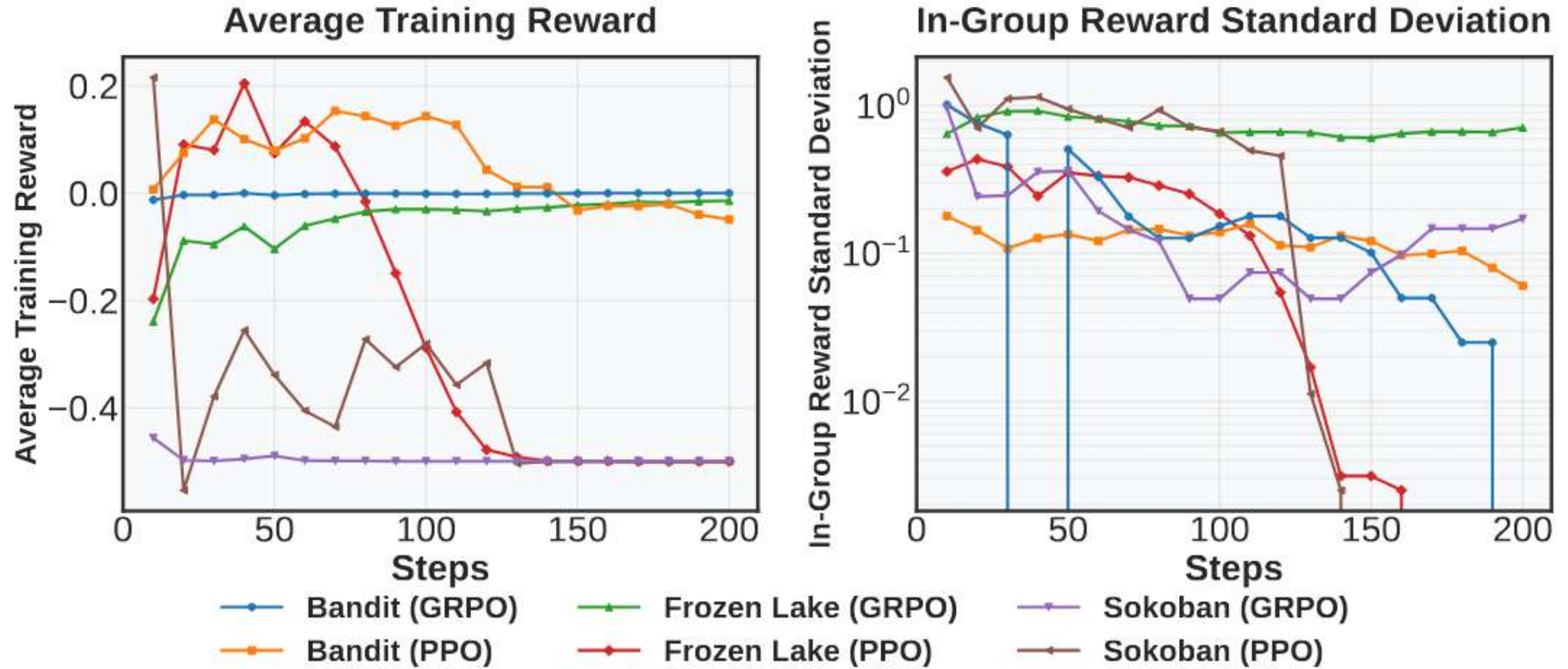
RL reinforces superficial patterns, suppressing exploration

Finding 3: Collapse Indicators

Ideal Reward: increase steadily

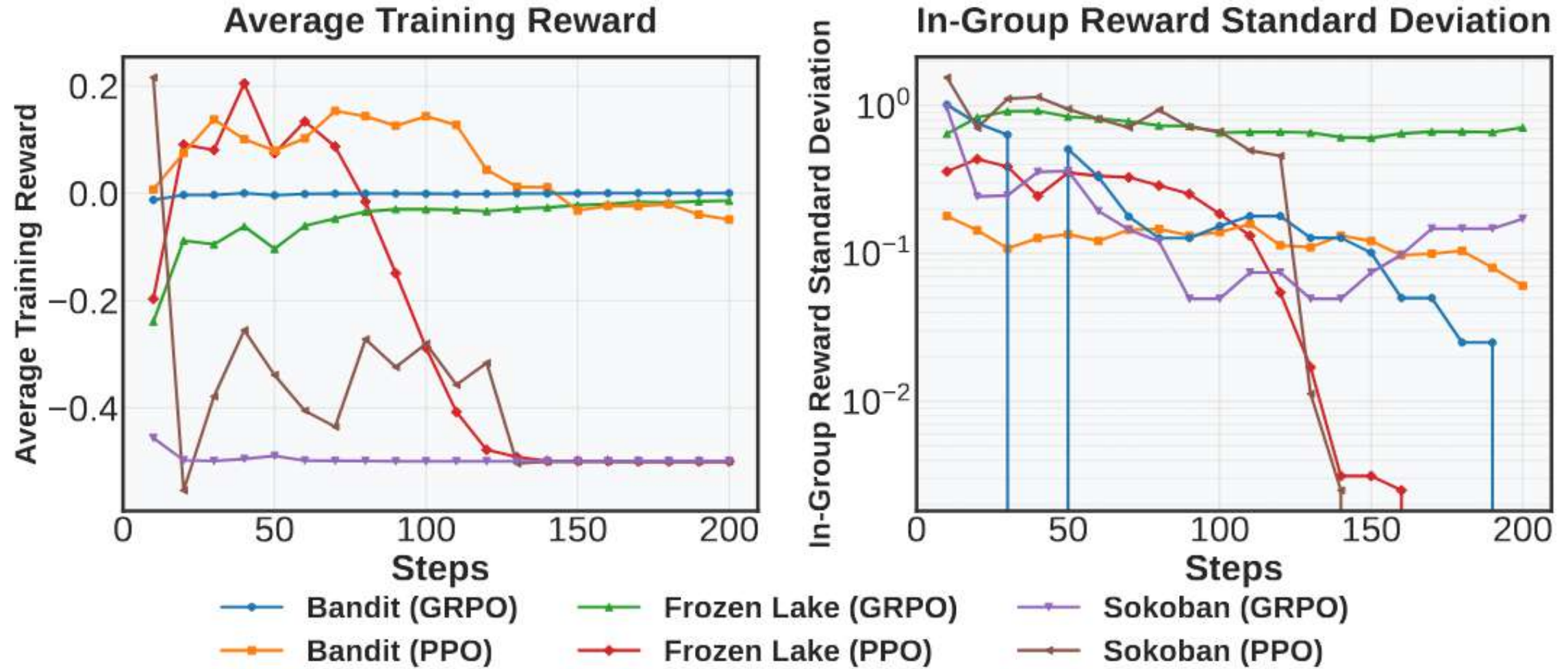


Finding 3: Collapse Indicators



Reward Standard Deviation: Drops sharply before performance degrade.

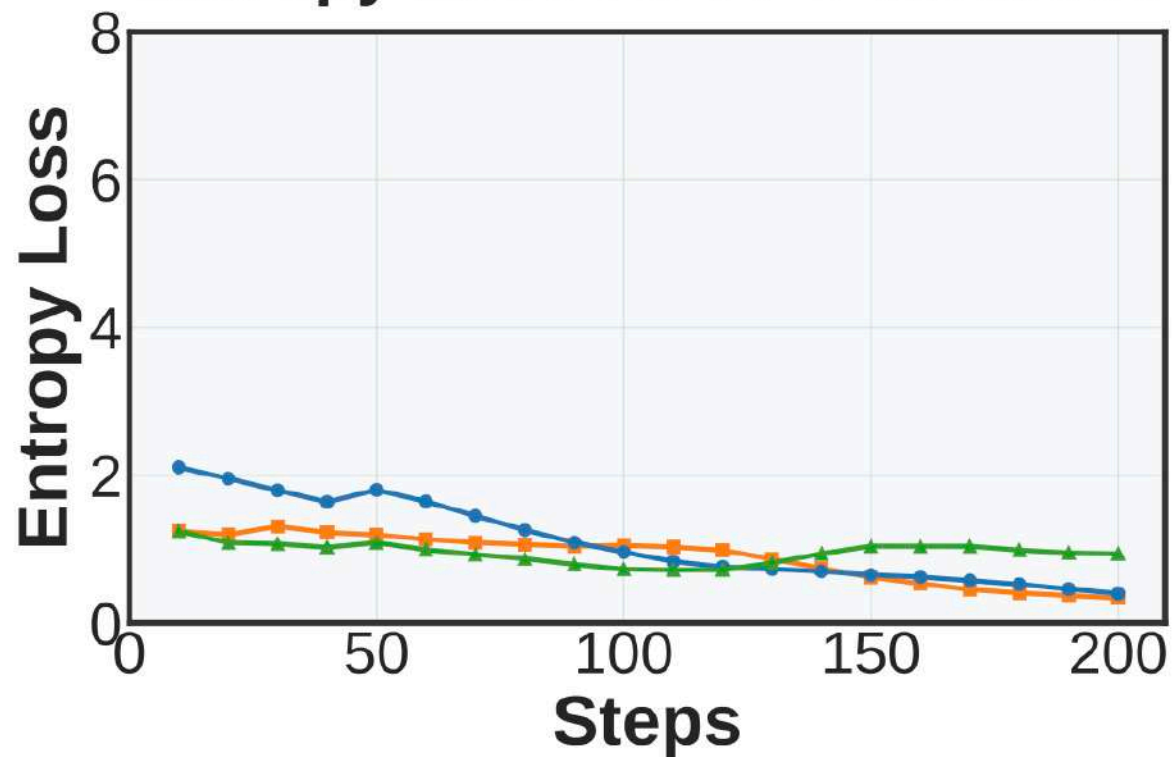
Finding 3: Collapse Indicators



Reward Standard Deviation: Drops sharply before performance degrade.

Finding 3: Collapse Indicators

Entropy Loss for PPO Methods

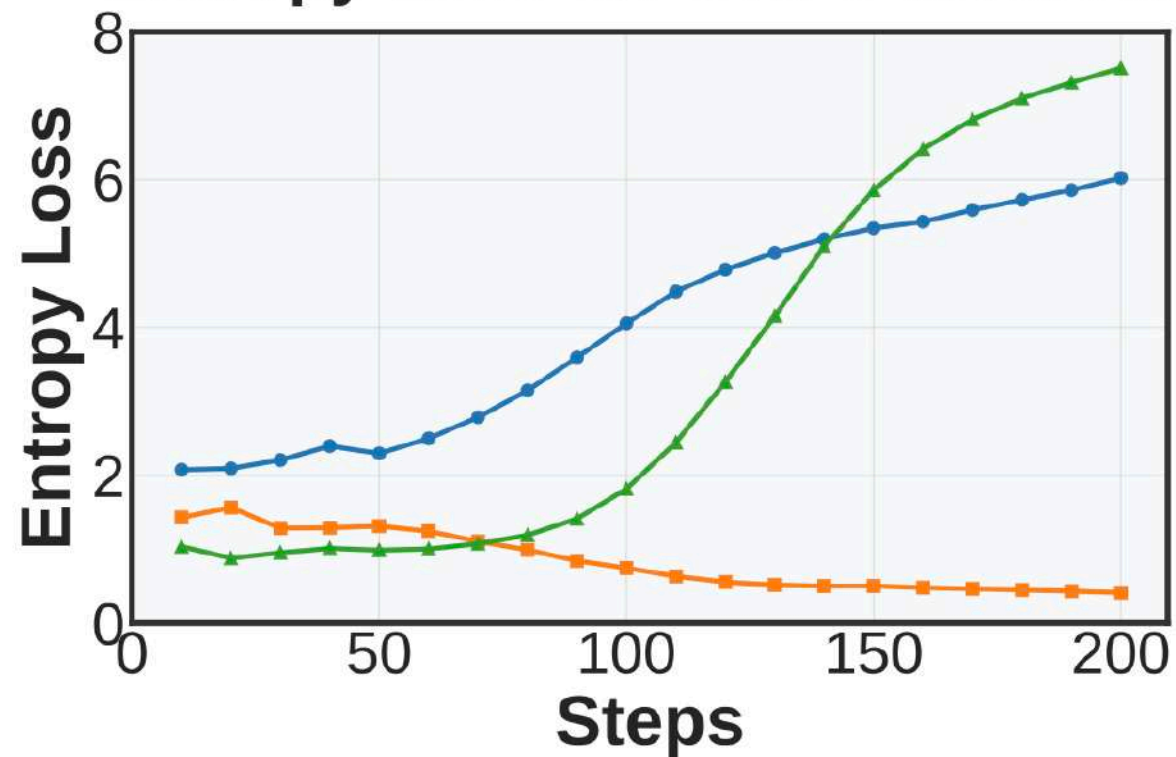


—●— Bandit

—■— Frozen Lake

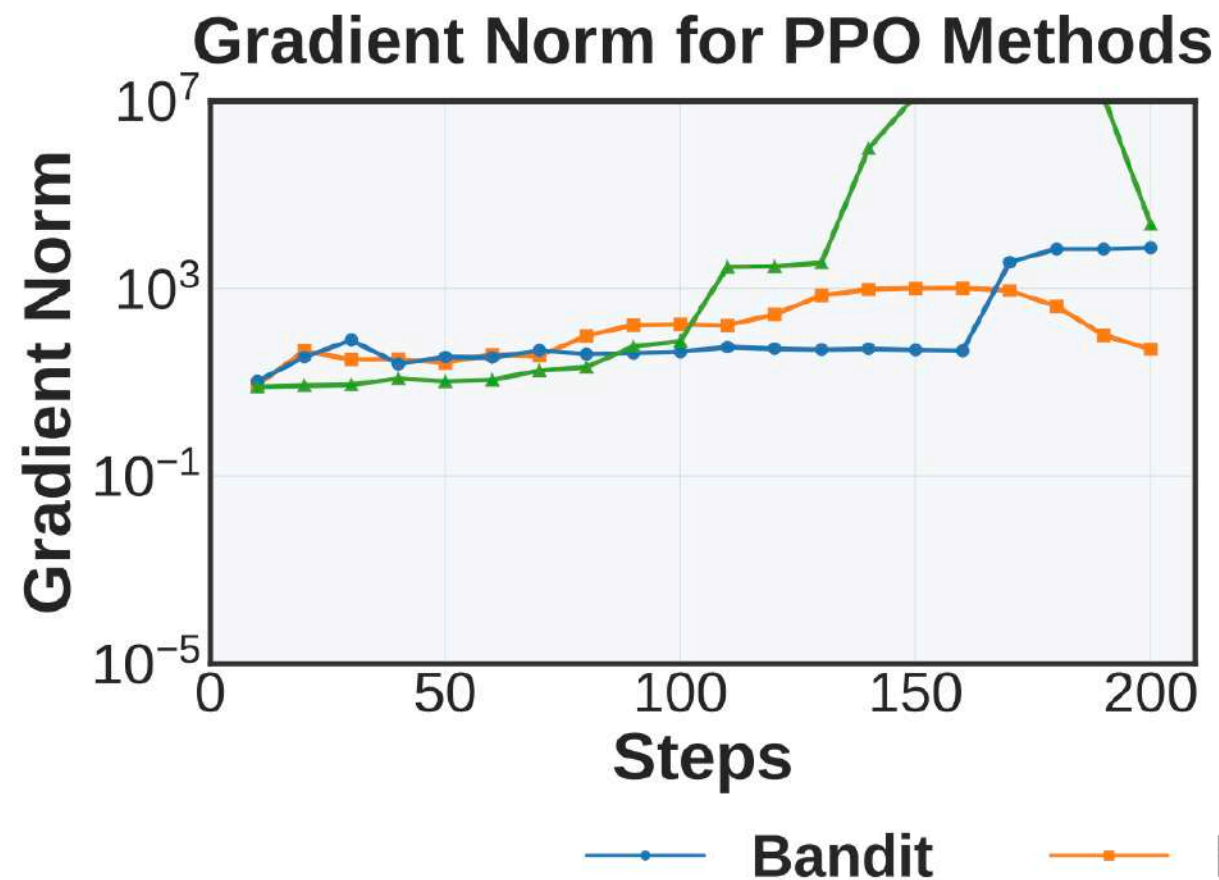
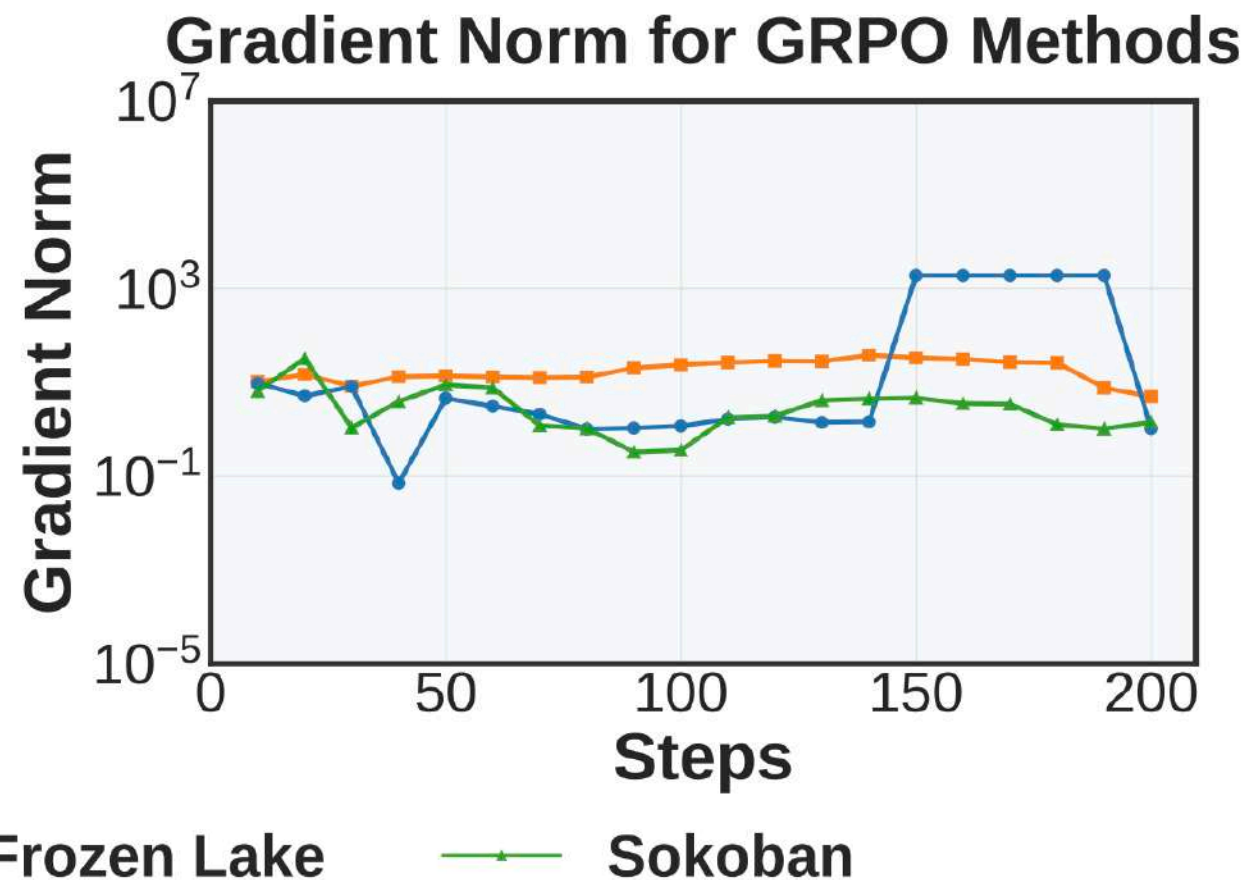
—▲— Sokoban

Entropy Loss for GRPO Methods



Output Entropy: Sharp drop indicates overconfidence, narrow reasoning patterns.

Finding 3: Collapse Indicators



Gradient Norm: plateaus and spikes confirm training instability..

Finding 3: Collapse Indicators

⚠ Early Warning Signs

- **Reward Standard Deviation:** Drops sharply before performance degrades (Fig 6).
- **Output Entropy:** Sharp drop indicates overconfidence, narrow reasoning paths.

✖ Confirmation Signs

- **Average Reward:** Plateaus or drops.
- **Gradient Norm:** Spikes indicate instability and irreversible collapse.

StarPO-S: Stabilized StarPO



Core Idea

Focus on training on more informative samples.

1

Trajectory Filtering

Retain only highly uncertain prompts (high reward variance).

2

KL Term Removal

Eliminate KL divergence penalty in PPO.

3

Decoupled Clipping (Asymmetric)

Use higher upper bound for PPO clipping.

Finding 4: Uncertainty Filtering Improves Stability

 Method: Sort prompts by reward variance and retain only the most uncertain ones.

Uncertainty Filtering

Sort prompts by reward variance (Std) from rollouts. Retain top $p\%$ most uncertain prompts.

$$U(\pi_{\theta}, M, s_0) = \text{Std}_{\tau \sim \pi_{\theta}(\cdot | s_0)}[R(\tau)]$$

Effect (Fig 7)

- Filtering low-variance rollouts significantly delays or eliminates collapse (esp. PPO).
- Improves training efficiency (fewer updates needed).
- Aggressive filtering (e.g., 25% retention) is effective.

StarPO-S: Stabilized StarPO

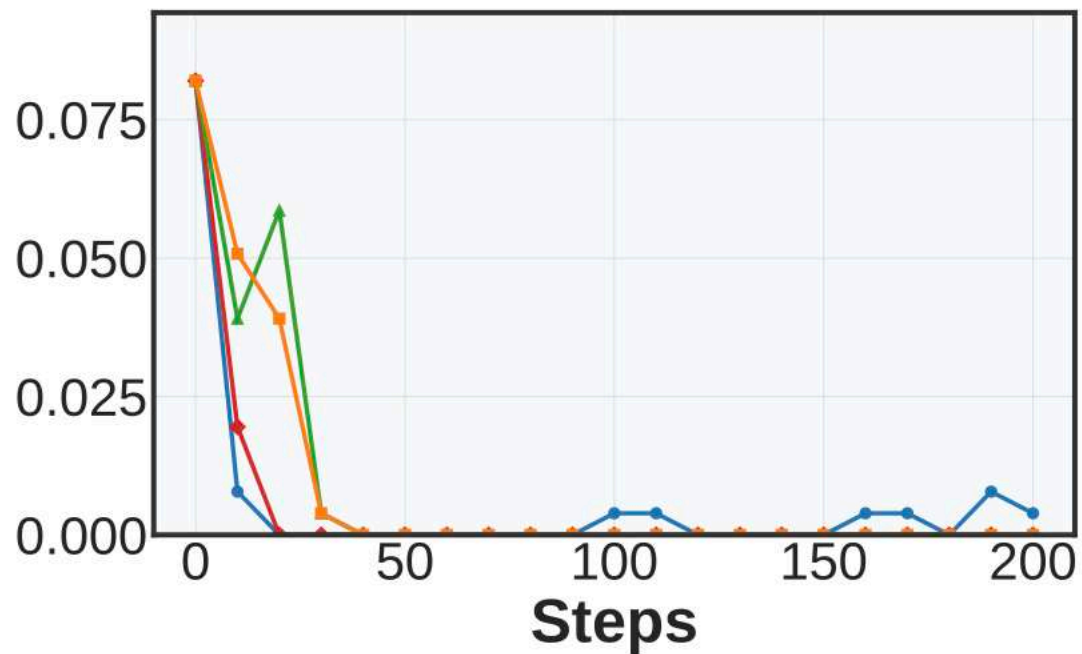


Filtering out low-variance trajectories reduces collapse risk.

Collapse is largely mitigated when more than half of the trajectories are filtered.

StarPO-S: Stabilized StarPO

Success Rate on Sokoban (GRPO)



—●— All Rollouts —●— Keep 25% —●— Keep 50% —●— Keep 75%

Critic is important

Finding 5: KL Removal & Asymmetric Clipping



Methods (from DAPO, adapted for multi-turn):



KL Removal

Eliminate KL divergence penalty in PPO objective.



Clip-Higher (Asymmetric)

Decouple PPO clipping bounds.

$$\epsilon_{high} = 0.28, \epsilon_{low} = 0.2$$

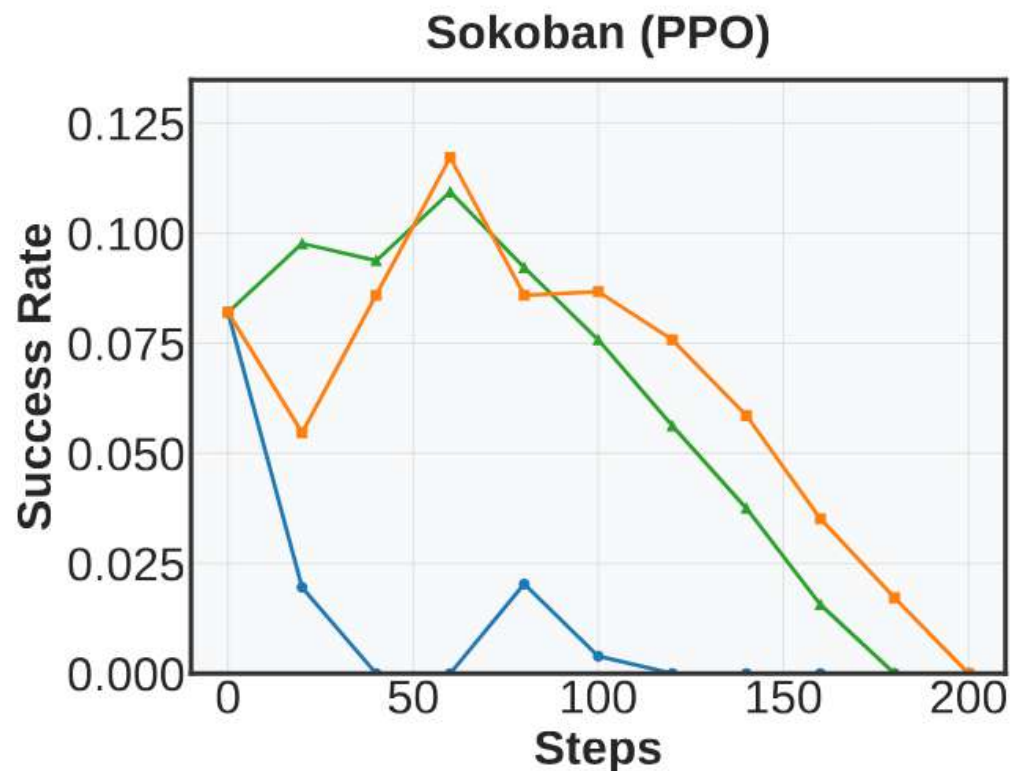
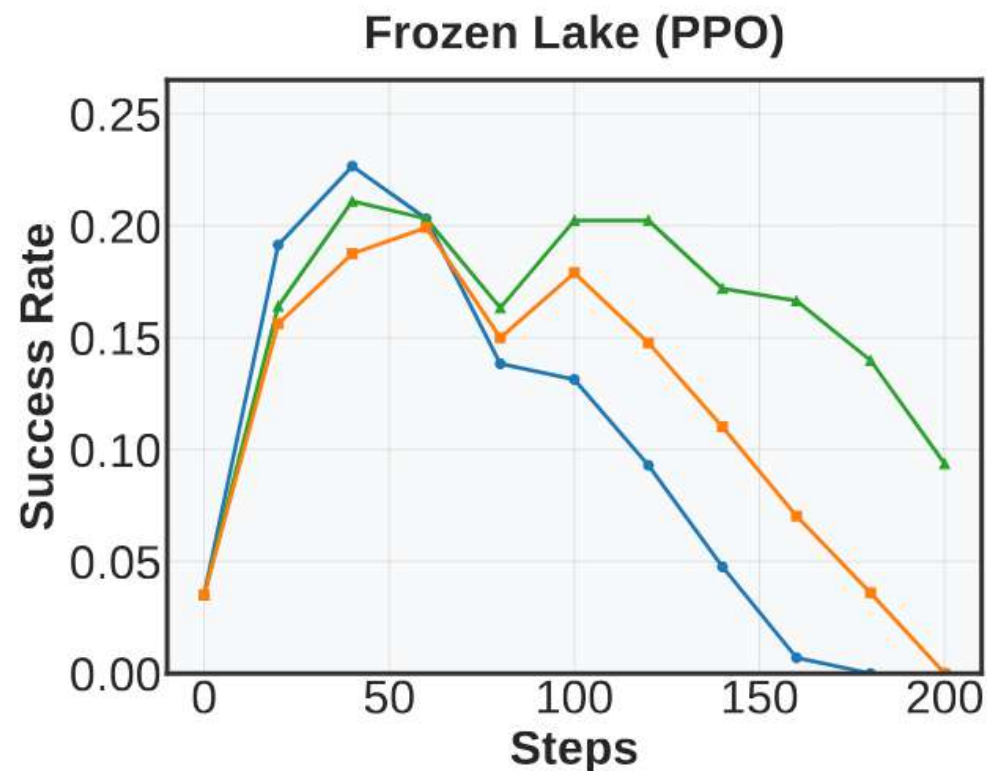
Note: Higher upper bound for better exploration



Effect (Fig 8)

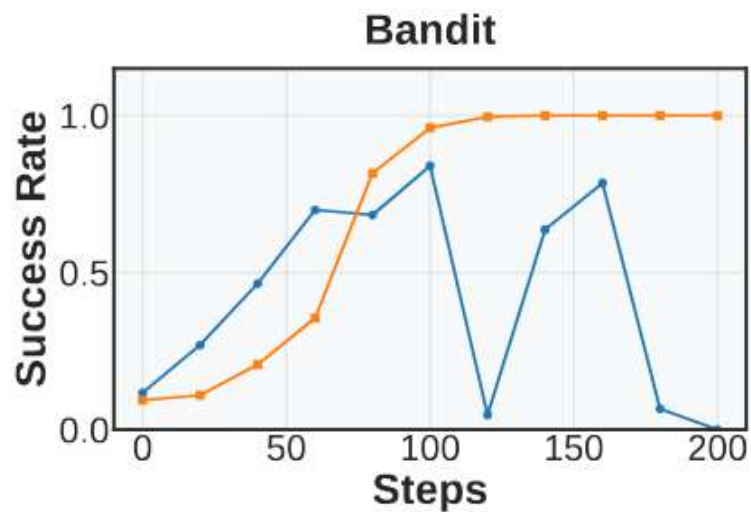
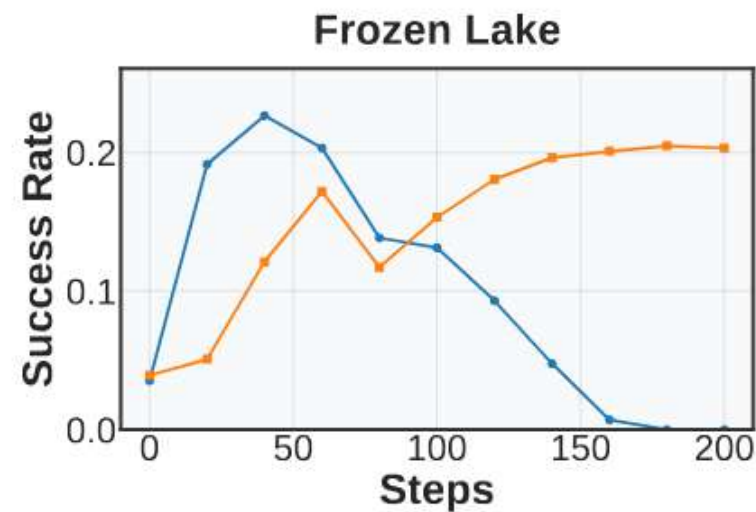
- Both methods improve peak success rate and delay collapse.
- Allows more aggressive learning from high-reward trajectories.
- More flexible gradient shaping.

StarPO-S: Stabilized StarPO



Removing KL constraints and enabling stronger positive gradient flow both improve peak performance and delay collapse in multi-turn RL.

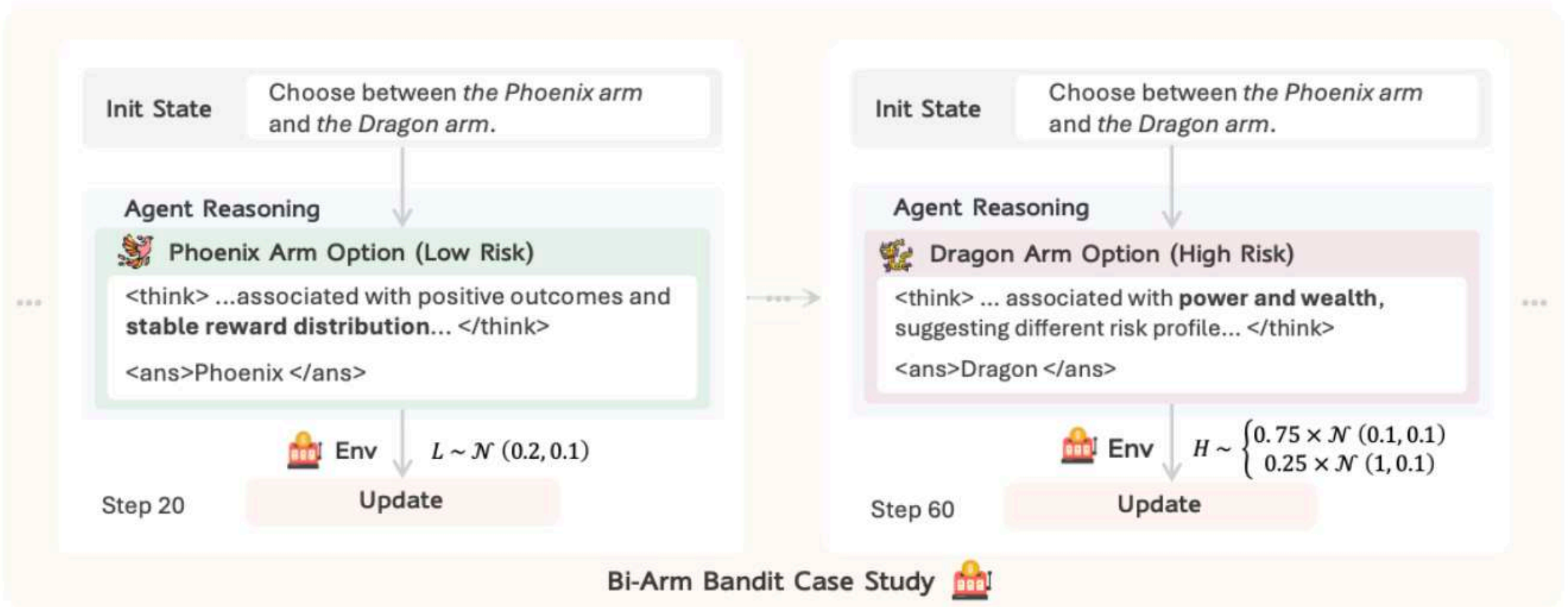
StarPO-S: Stabilized StarPO



—●— StarPO (w/ PPO) —●— StarPO-S

Finally!

RAGEN: Training Agents by Reinforcing Reasoning



Finding 6: Rollout Quality Matters



Key Dimensions: Task diversity, interaction granularity, rollout frequency



Key Dimensions of Rollout Quality:

Task Diversity

Interaction Granularity

Rollout Frequency



Task Diversity (Table 2)

Higher diversity:

More distinct prompts improves generalization

Optimal configuration:

Fewer responses per prompt (e.g., 4 per prompt)

Minimum requirement:

Multiple responses per prompt to enable contrast



Interaction Granularity (Table 3)

Optimal actions per turn:

5-6 actions enables better planning

Too many actions:

7+ actions degrades performance

Reason:

Avoids noise from overly long rollouts

Finding 7: Reasoning Needs Fine-Grained Rewards

 Problem: Simply encouraging reasoning doesn't guarantee actual reasoning

Current Reward Limitations

- Models often regress to direct action selection if no reward advantage exists.
- Shallow strategies suffice in simple MDPs.

Hallucination

When rewards only reflect success, models produce fake reasoning ("hallucinated thoughts").

Our Engineering Efforts



Structured Rollouts



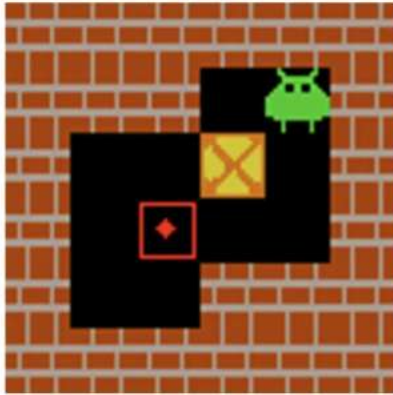
Customizable Rewards



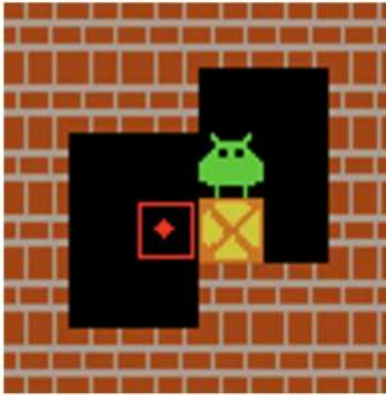
Multi-turn Environments

Easy to Plug-in Your Own Model / Environment!

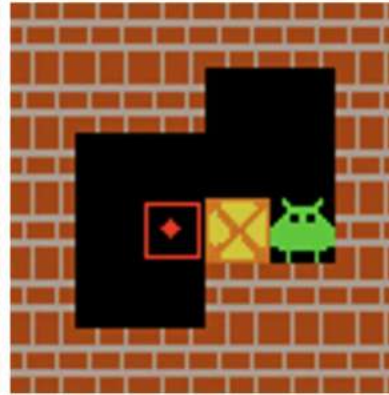
VAGEN



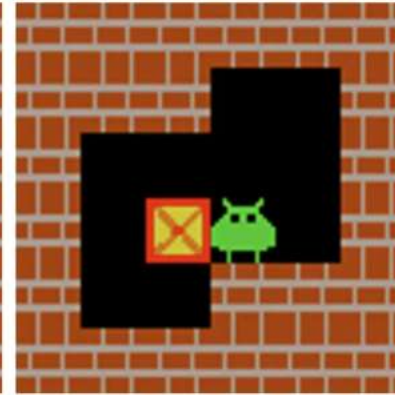
```
<think>
First, I need to push the box to
the left to clear the way to the
target. Then, I can push the box
down to the target.
</think>
<answer>
Left, Down
</answer>
```



```
<think>
Since the box is still on the left, I
need to push it to the right to the
target. Then, I can push the box
down to the target.
</think>
<answer>
Right, Down
</answer>
```



```
<think>
Since the box is still on the right,
I need to push it to the left to the
target. Then, I can push the box
down to the target.
</think>
<answer>
Left, Down
</answer>
```



Kangrui Wang*, Pingyue Zhang*, Zihan Wang*, Qineng Wang*, Linjie Li*,
Zhengyuan Yang, Hanyang Zhang, Tong Zhang, Kyunghyun Cho,
Lijuan Wang, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, Yejin Choi, Manling Li

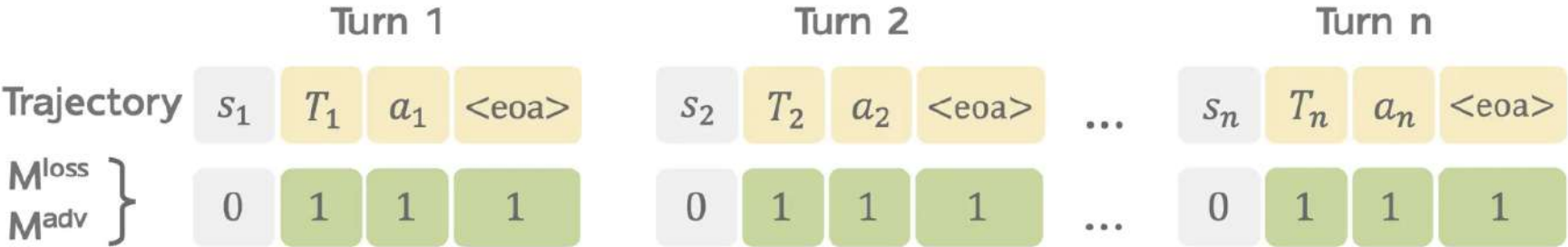
Challenge 1

States: text $\rightarrow$ visual

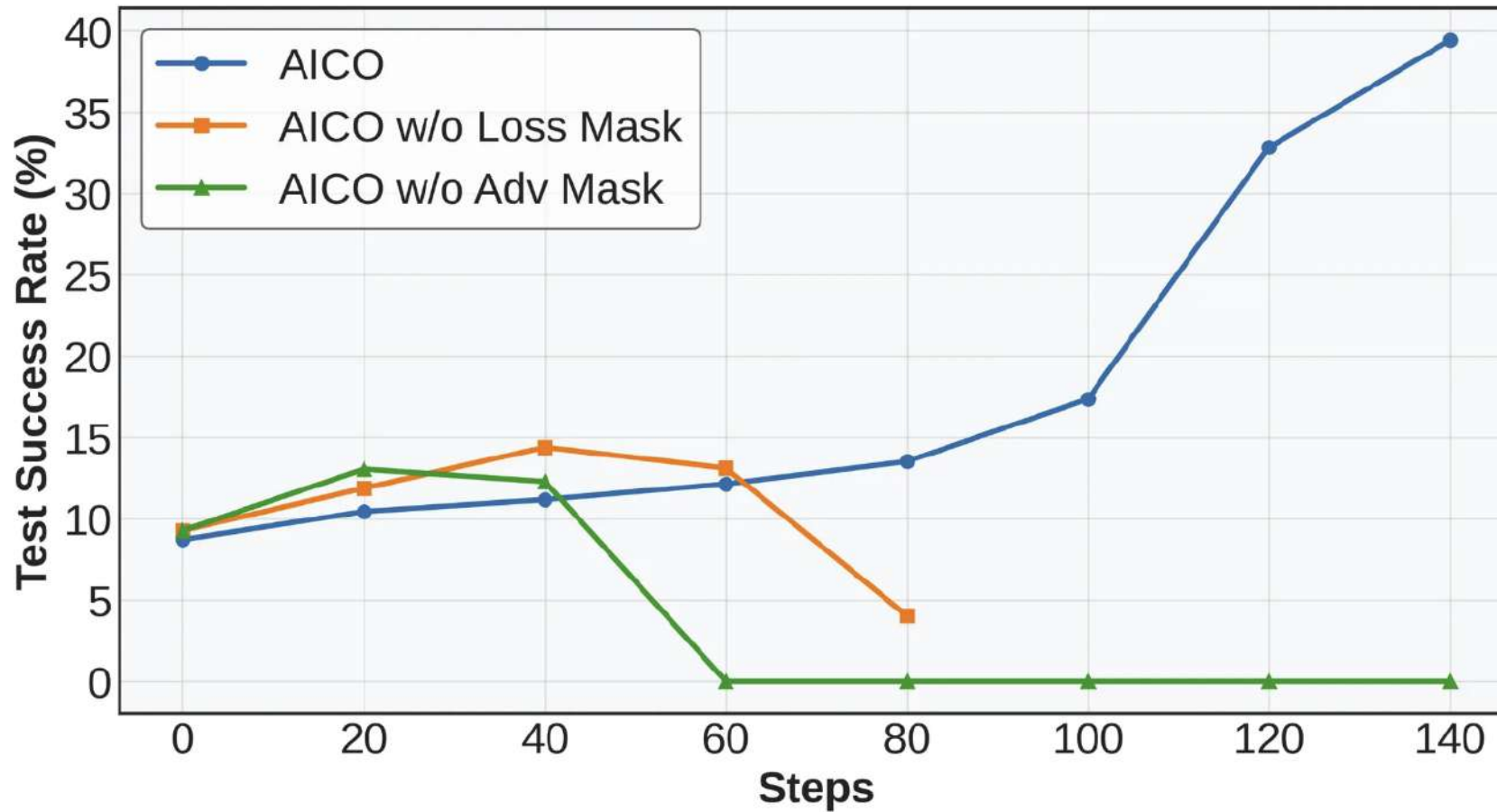
Challenge 1 States: text $\rightarrow$ visual



Selective Token Mask



Findings 1: masking image tokens during training improves performance



Comparing our method with versions that remove different masks, we find that both loss mask and GAE mask are crucial for performance. Image tokens should be ignored during both advantage estimation and loss back-propagation

Conclusion 1

Not every token is important for optimization

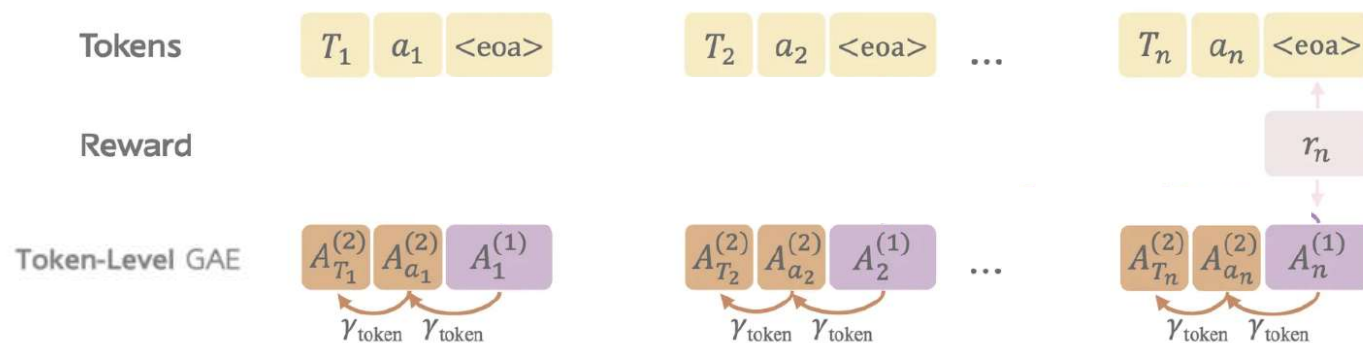
- VLMs face distribution shift when multimodal inputs are introduced (input_space!=output_space)
- Visual tasks suffer from state redundancy in long-context image and low-level details

Challenge 2

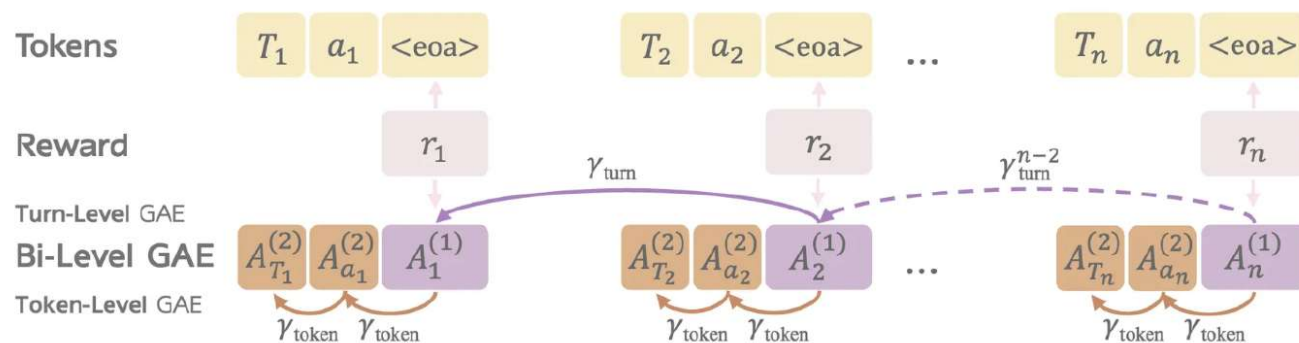
Single-turn $\rightarrow$ Multi-turn

So, stabilize the RL training via Multi-turn Advantage Estimation

Challenge 2: Single-turn $\rightarrow$ Multi-turn



Cross-Turn Credit Assignment (Turn Reward)



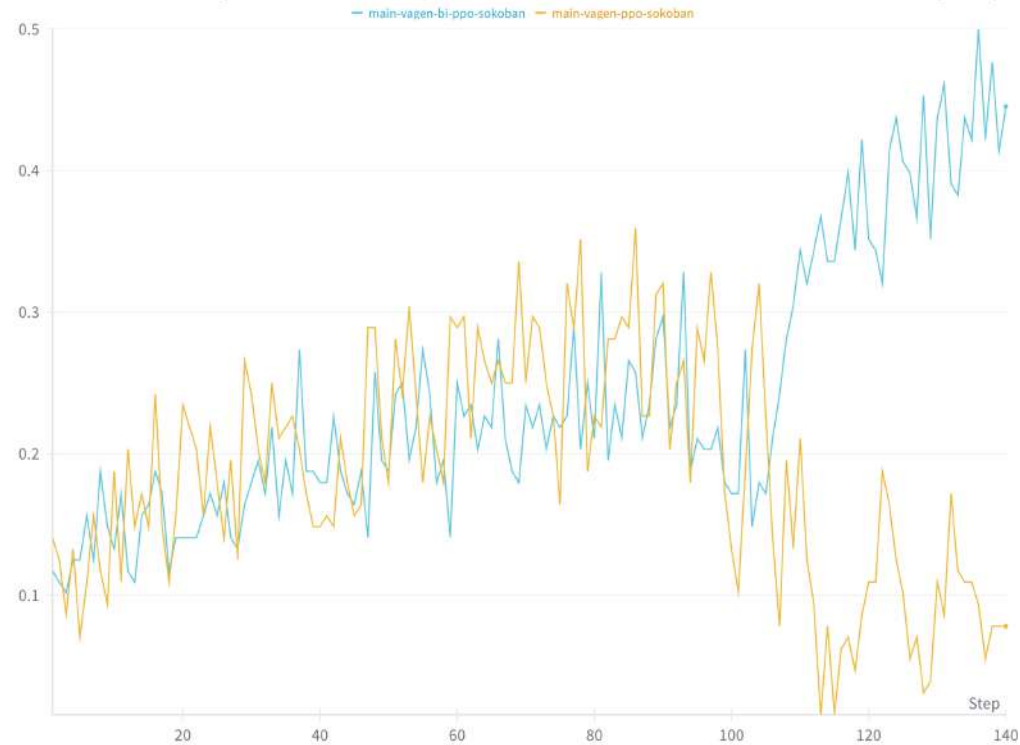
Finding 2.1: (Bi-level GAE + Turn reward) > (GAE + Traj Reward)

train/success/FrozenLakeEnvConfig(format_reward=0.5,is_slippery=False,size=4,p=0.8,render_mode=vision,max_actions_per_step=3,min_actions_to_succeed=5)

i-ppo-frozenlake train/success/FrozenLakeEnvConfig(format_reward=0.5,is_slippery=False,size=4,p=0.8,render_mode=vision,max_actions_per_step=3,min_actions_to_succeed=5)
ain-vagen-bi-ppo-frozenlake train/success/FrozenLakeEnvConfig(is_slippery=False,size=4,p=0.8,render_mode=vision,max_actions_per_step=3,min_actions_to_succeed=5)
ppo-frozenlake train/success/FrozenLakeEnvConfig(format_reward=0.5,is_slippery=False,size=4,p=0.8,render_mode=vision,max_actions_per_step=3,min_actions_to_succeed=5)
nain-vagen-ppo-frozenlake train/success/FrozenLakeEnvConfig(is_slippery=False,size=4,p=0.8,render_mode=vision,max_actions_per_step=3,min_actions_to_succeed=5)

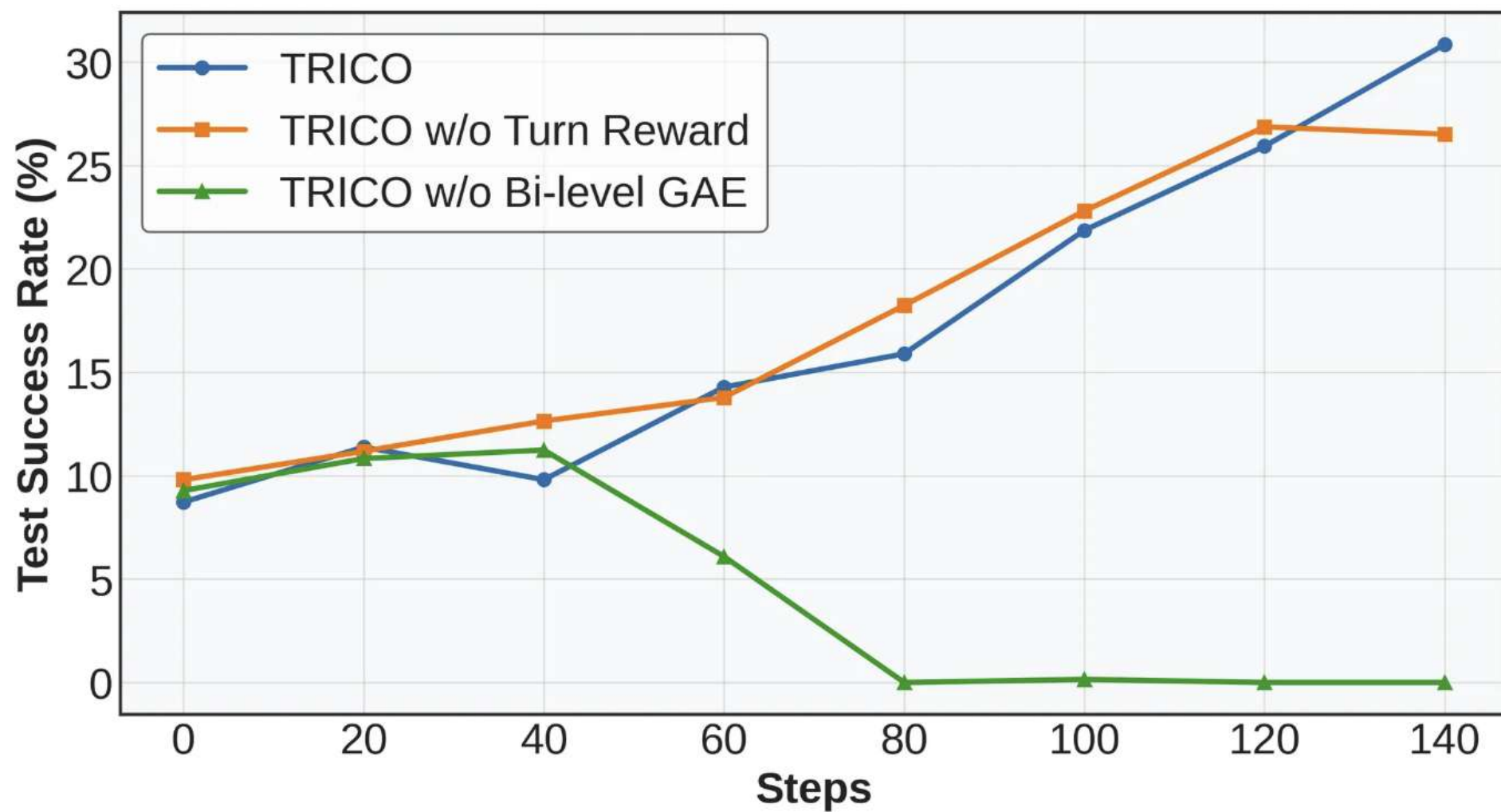


train/success/SokobanConfig(dim_room=(6, 6),max_steps=100,num_actions=10,render_mode=vision,min_actions_to_succeed=5,max_actions_per_step=3)



- (Bi-level GAE + Turn reward) is consistently better than (GAE + Traj Reward)

Finding 2.2: Turn Reward and Bi-level GAE jointly enhance cross-turn credit assignment



- Removing Turn Reward leads to a slight performance drop.
- Removing Bi-level GAE causes the model to collapse after a certain level of training.

Conclusion 2

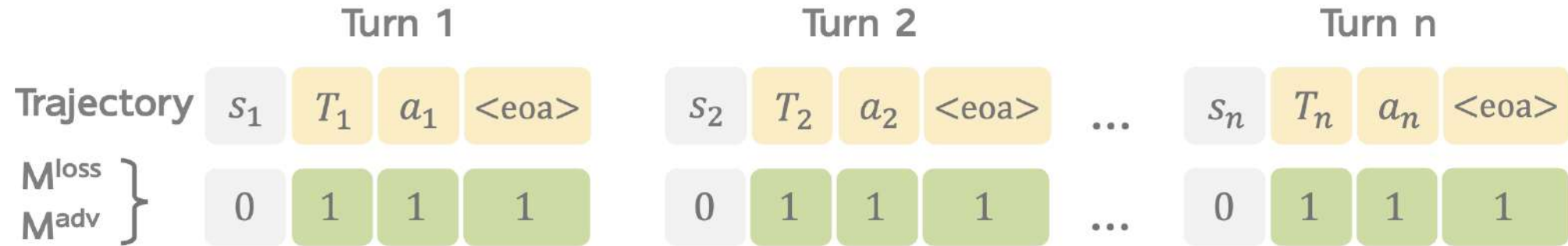
Accurate Advantage Estimation

- VLM generates an action token by token, causing two levels of decision making processes
 - Action Level: POMDP (Partially Observable Markov Decision Process)
 - Token Level: MDP (Markov Decision Process)

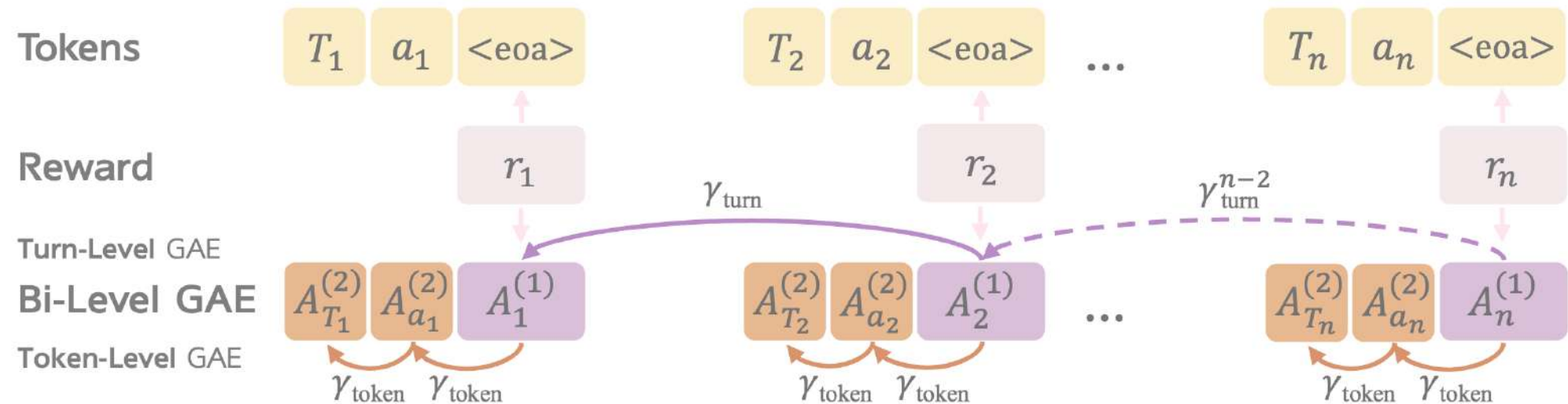
Putting them together...

TRICO: Turn-Aware Reason-Interaction Chain Optimization

Selective Token Masking



Cross-Turn Credit Assignment (Turn Reward)



Algorithm 2 TRICO: Turn-aware Reason-Interaction Chain Optimization

Input: Policy LM π_θ , Environment, Critic V_ϕ , Reference LM π_{ref}

Hyperparameters: Discount γ_{turn} , γ_{token} , GAE λ , PPO clip ε , KL penalty β

Ultimate Objective: $\max_\theta E_{s \sim \pi_\theta} [R(s)]$ where $s = (s_0, a_0^T, s_1, a_1^T, \dots, s_K, a_K^T)$

Phase 1: Trajectory Collection

- 1: Sample initial observation s_0 and generate sequence $s = (s_0, a_0^T, s_1, a_1^T, \dots, s_K, a_K^T)$ where:
 - $a_k^T = \langle \text{think} \rangle \dots \langle / \text{think} \rangle \langle \text{ans} \rangle a_k \langle / \text{ans} \rangle \langle \text{eoa} \rangle \sim \pi_\theta(\cdot | s_{<t})$
 - Only $\langle \text{ans} \rangle a_k \langle / \text{ans} \rangle$ part is sent to environment to receive s_{k+1}
 - 2: Store complete sequence with policy probabilities as π_{old}
-

Phase 2: Token-level Advantage Estimation

- 3: Define s_t as the sequence up to the t -th token in s
 - 4: Assign masks: M^{loss} (tokens to update), M^{adv} (tokens for advantage discount calculation)
 - 5: For each token t in sequence s :
 - Calculate KL penalty: $r_t^{\text{KL}} = -\beta \cdot (\pi_\theta(\cdot | s_{<t}) - \pi_{\text{ref}}(\cdot | s_{<t}))$
 - Set rewards: $r_t = r_t^{\text{KL}}$ for non-terminal tokens; $r_t = R_u(x_0, \dots, y_k) + r_t^{\text{KL}}$ at the k -th $\langle / \text{eoa} \rangle$
 - 6: Compute Bi-level Advantage using GAE:
 1. Use γ_{turn} when computing GAE across turn boundaries (at $\langle \text{eoa} \rangle$ positions)
 2. Use γ_{token} for tokens within the same turn
 - 7: Apply masking: Skip tokens with $M^{\text{adv}} = 0$ in calculations
 - 8: Normalize advantages with masked whitening
 - 9: Update returns: $\hat{R}_t = A_t^{\text{GAE}} + V_\phi(s_t)$
-

Phase 3: Policy Update with PPO

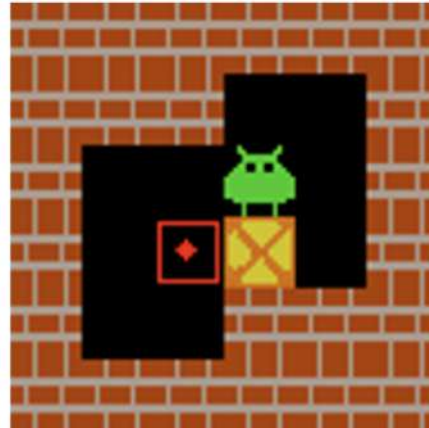
- 10: Compute probability ratios: $r_t(\theta) = \frac{\pi_\theta(s_t | s_{<t})}{\pi_{\text{old}}(s_t | s_{<t})}$
 - 11: Apply masked PPO objective: $J^{\text{PPO}}(\theta) = \frac{\sum_t M^{\text{loss}}[t] \cdot [\min(r_t(\theta) A_t^{\text{GAE}}, \text{clip}(r_t(\theta)) A_t^{\text{GAE}})]}{\sum_t M^{\text{loss}}[t]}$
 - 12: Update critic: $J^{\text{Critic}}(\phi) = \frac{\sum_t M^{\text{loss}}[t] \cdot [(V_\phi(s_t) - \hat{R}_t)^2]}{\sum_t M^{\text{loss}}[t]}$
 - 13: Return to **Phase 1** for the iterative optimization
-

Comparison with Existing Method

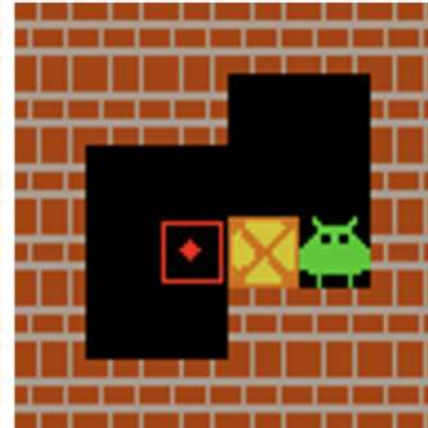
Feature	PPO	StarPO	TRICO (Ours)
Sequence Structure	Single response (y) to prompt (x)	Multiple turn interaction ($y_0, x_1, y_1, \dots$)	Multiple turn interaction ($y_0, x_1, y_1, \dots$)
LM output Representation	No special structure	<div><think>...</think></div> <div><ans>...</ans></div>	<div><think>...</think></div> <div><ans>...</ans><eo></div>
Discounting	Single discount rate γ	Single discount rate γ	Bi-level: γ_{turn} and γ_{token}
Optimization	All tokens equally	All tokens equally	Selective token optimization



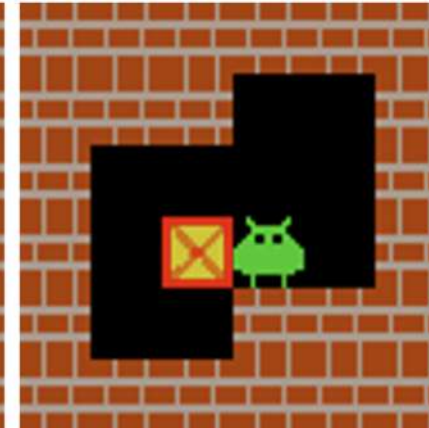
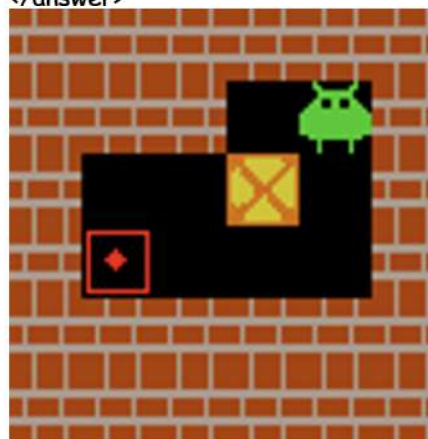
<think>
First, I need to push the box to the left to clear the way to the target. Then, I can push the box down to the target.
</think>
<answer>
Left, Down
</answer>



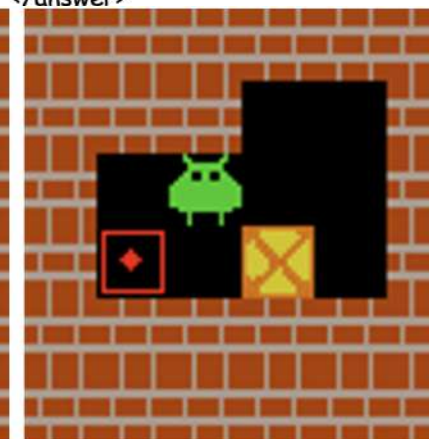
<think>
Since the box is still on the left, I need to push it to the right to the target. Then, I can push the box down to the target.
</think>
<answer>
Right, Down
</answer>



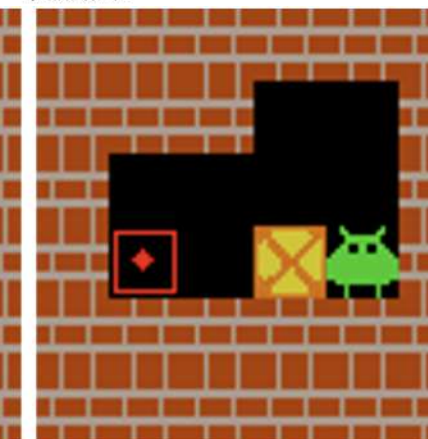
<think>
Since the box is still on the right, I need to push it to the left to the target. Then, I can push the box down to the target.
</think>
<answer>
Left, Down
</answer>



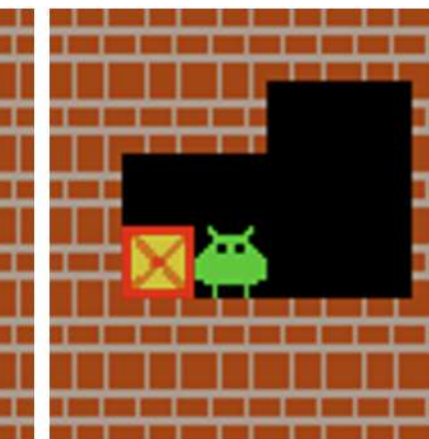
<think>
First, I need to push the box to the left to clear the way to the target. Then, I can push the box down to the target.
</think>
<answer>
Left, Down, Left
</answer>



<think>
Since the box is still on the left, I need to push it to the right to the target. Then, I can push the box down to the target.
</think>
<answer>
Right, Right, Down
</answer>



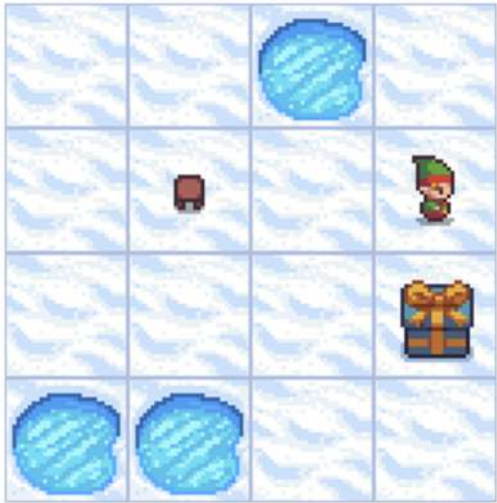
<think>
Since the box is still on the right, I need to push it to the left to the target. Then, I can push the box down to the target.
</think>
<answer>
Left, Left, Down
</answer>



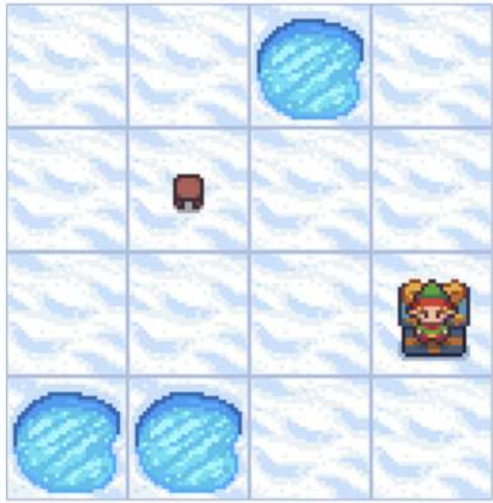
Results – Frozenlake Examples



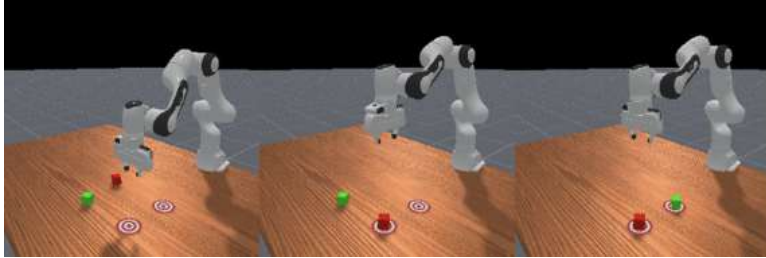
<think>
To reach the goal, I should first move towards the gift box. The gift box is located to the right of the player. Therefore, the first step should be to move right.
</think>
<answer>
Right, Right, Right
</answer>



<think>
Since moving right did not reach the goal, the next step should be to move down to see if that helps in reaching the gift box.
</think>
<answer>
Down, Down, Down
</answer>



Results – Maniskill Examples



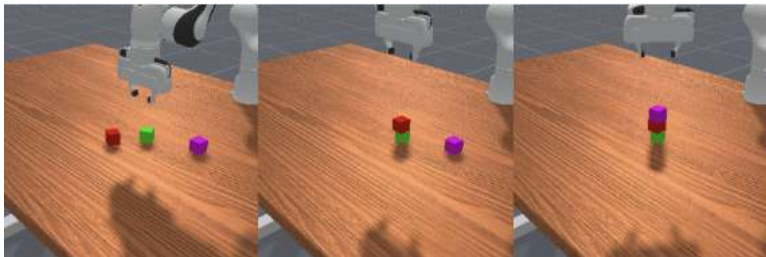
<think>First, I need to grasp cubeA at its current position and then place it at goalA.</think><answer>pick(-65, -200, 20)||place(100, 100, 0)

<think>Next, I need to grasp cubeB and place it at goalB.</think><answer>pick(142, -142, 20)||place(-100, 100, 0)</answer><|im_end|>



<think>I need to grasp the apple first to move it to the drawer.</think><answer>pick(-128, 161, 36)||place(-500, -537, 400)</answer><|im_end|>

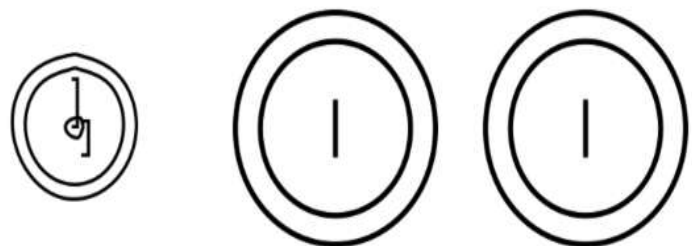
<think>I need to close the drawer to complete the task.</think><answer>push(-500, -367, 200, -500, -537, 200)</answer>



<think>I need to grasp the red cube and place it on top of the green cube first.</think><answer>pick(66, -86, 20)||place(-28, -21, 60)

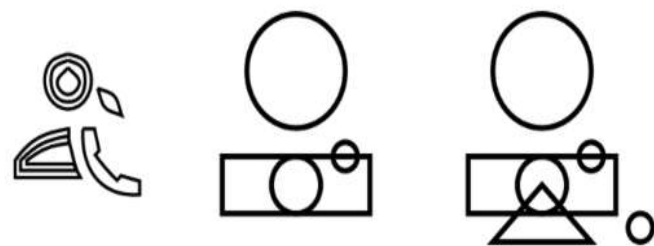
<think>I need to grasp the purple cube and place it on top of the red cube.</think><answer>pick(-96, 140, 20)||place(-20, -20, 100)</answer>

Results – SVG Code Generation Examples



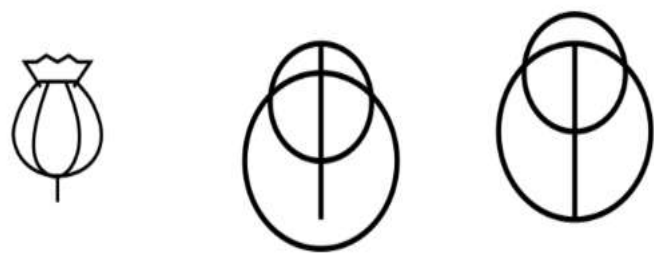
<think>
I need to generate SVG code for an icon that resembles a clock with a pendulum. The icon will have a circular shape with a pendulum inside. The pendulum should be a simple line, and the clock face should be a circle with a smaller circle inside it to represent the hour hand.
</think>

<think>
The SVG code generated initially is almost correct, but the pendulum line is not visible. I need to add the pendulum line to make it more precise and similar to the original image.
</think>



<think>
I need to generate SVG code for an icon that resembles a person with a leaf and a water droplet. The icon should be simple and clear, with a person's head, a leaf, and a water droplet as the main elements. The person's head should be circular with a small circular element in the middle, representing the eye. The leaf should be a simple, stylized shape, and the water droplet should be a small, circular element.
</think>

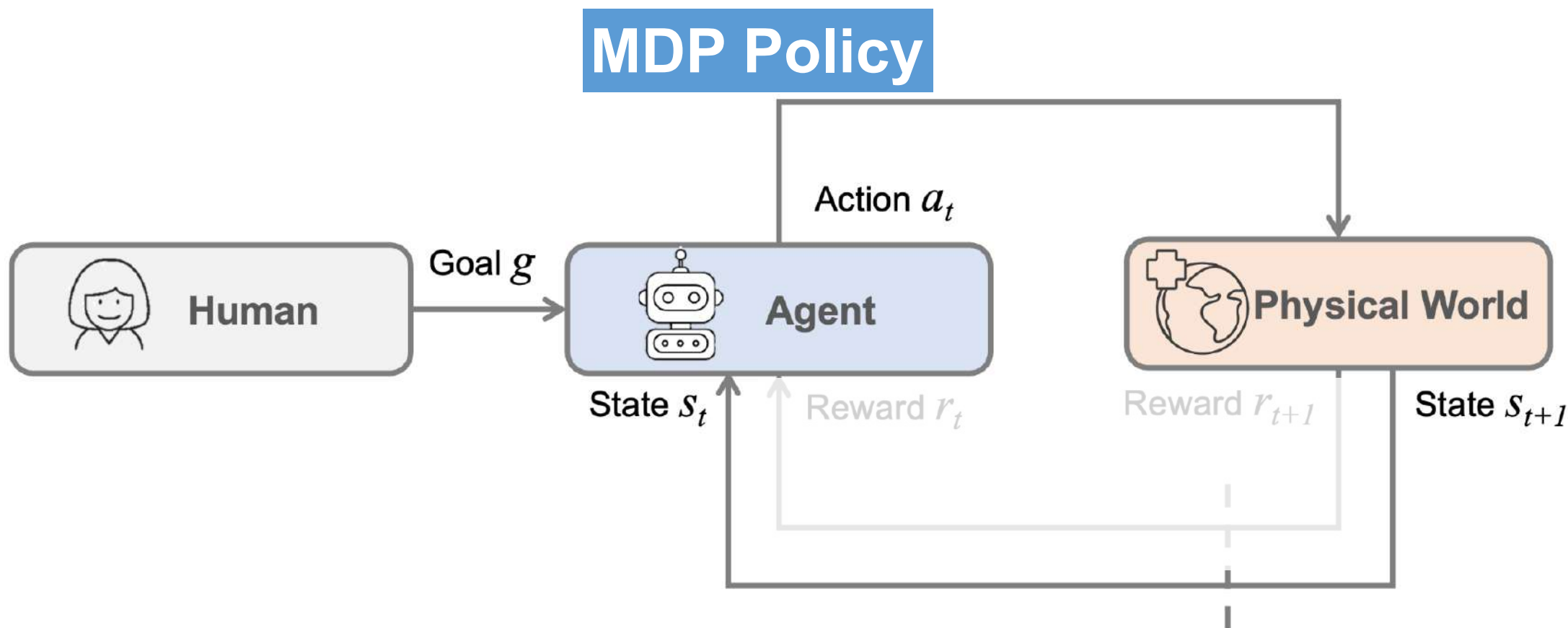
<think>
The SVG code generated earlier has a person's head with a circular element in the middle, a leaf, and a water droplet. However, the current image shows a person's head with a circular element in the middle, a rectangular base, and a small circular element on the bottom right. The leaf and water droplet are missing in this image. I need to adjust the SVG code to include these elements.
</think>



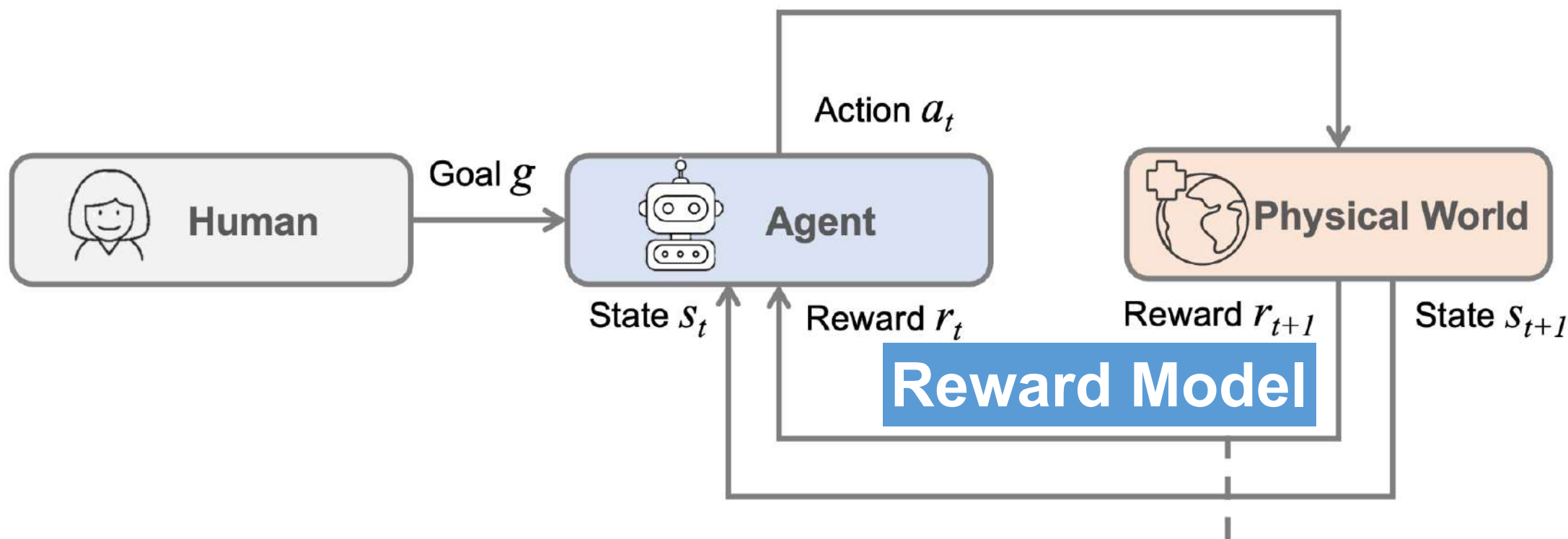
<think>
I need to generate SVG code that accurately represents the image of a flower bud. The flower bud has a round shape with a small, pointed tip at the top. The stem is a simple line extending downward from the bottom of the bud.
</think>

<think>
I need to revise the SVG code to make it more precise and similar to the original image. The original image shows a flower bud with a round shape and a small, pointed tip at the top. The stem is a simple line extending downward from the bottom of the bud. The revised SVG code should reflect these details accurately.
</think>

Let us go back to MDPs (Markov Decision Processes)



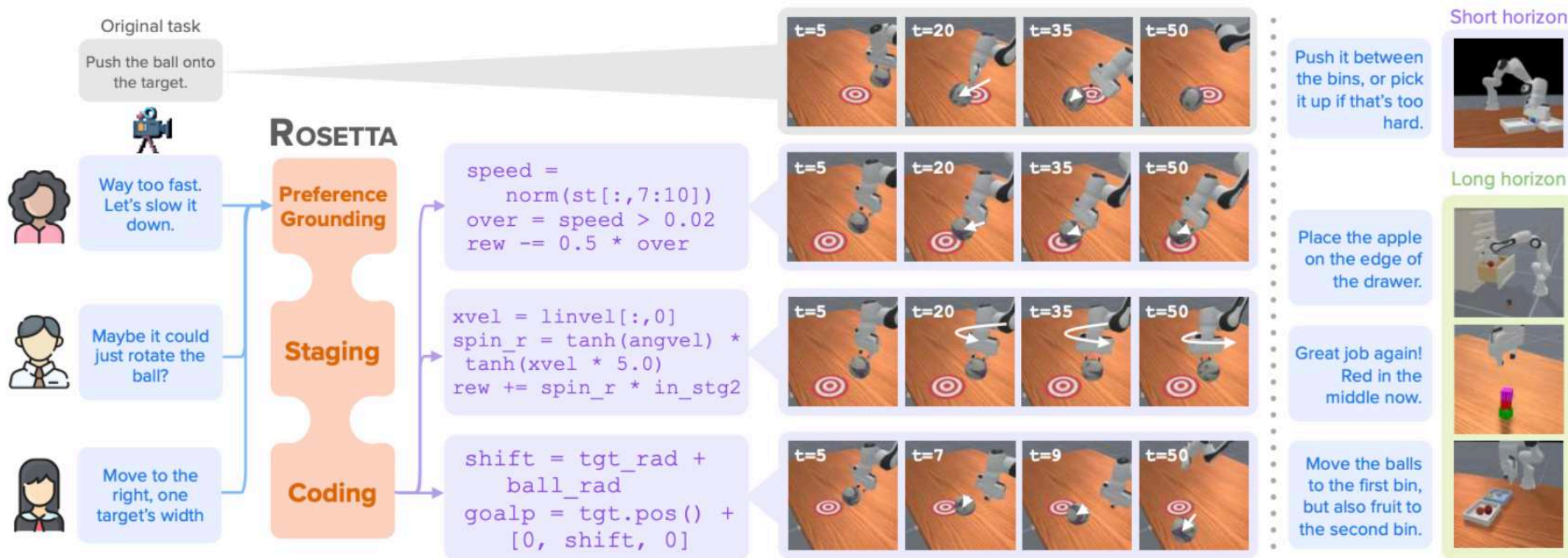
Let us go back to MDPs (Markov Decision Processes)



Reward Model

$$o, a \rightarrow r$$

Constructing Code-Based Reward from Unconstrained Language Preference



Embodied Agent Interface

Benchmarking LLMs for Embodied Decision Making



Manling Li



Shiyu Zhao*



Qineng Wang*



Kangrui Wang*



Yu Zhou*



Sanjana Srivastava



Tony Lee



Erran Li



Ruohan Zhang



Weiyu Liu



Percy Liang



Fei-Fei Li



Jiayuan Mao



Jiajun Wu



So we need **Standardization!**

LLMs

Embodied Agent Interface

VirtualHome

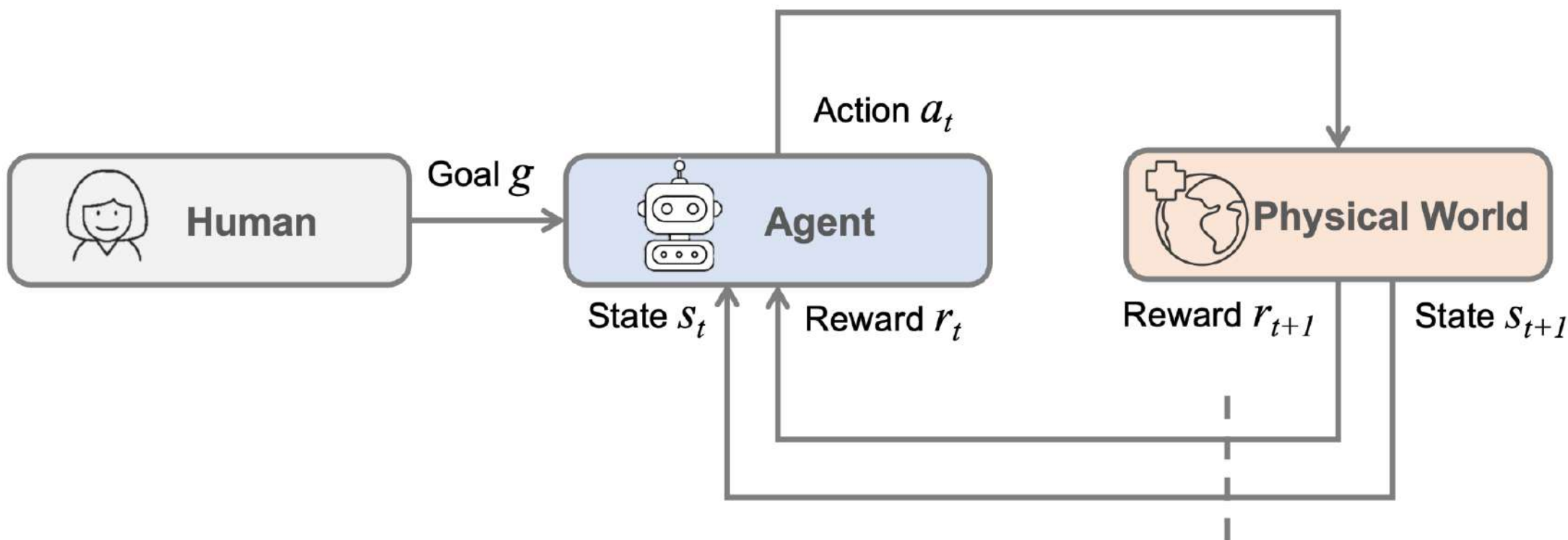
BEHAVIOR

...

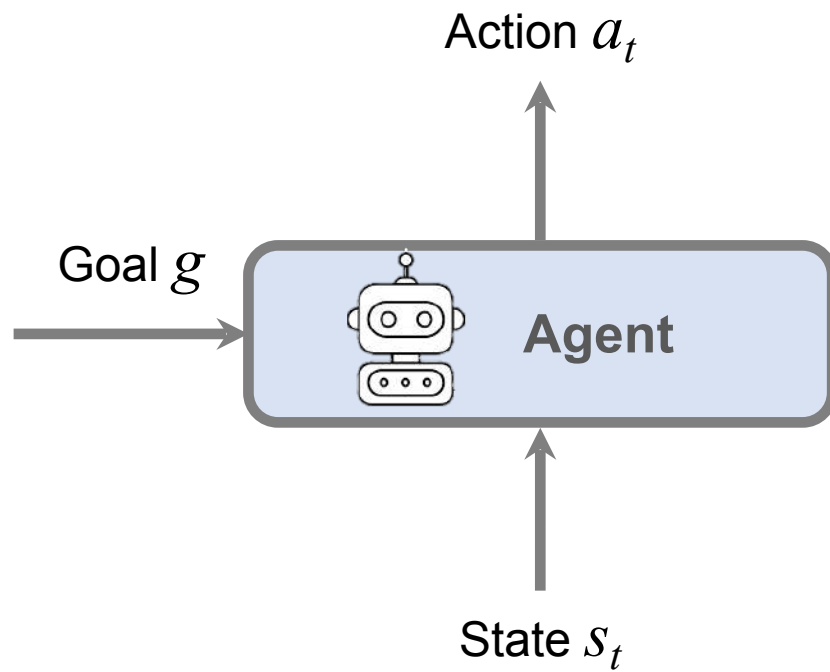
Embodied Environments

A Single Line of Code `eai-eval --all`

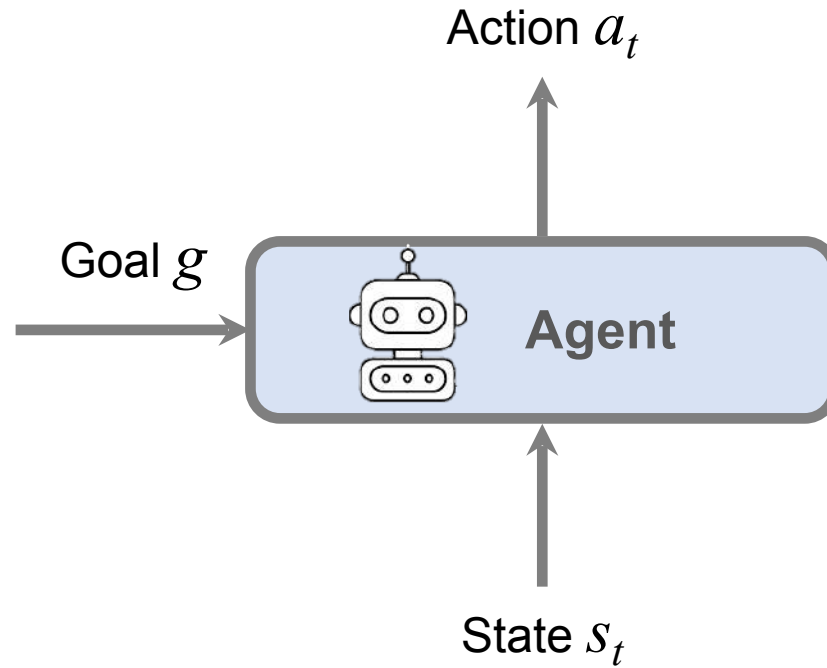
Let us go back to MDPs (Markov Decision Processes)



Markov Decision Process



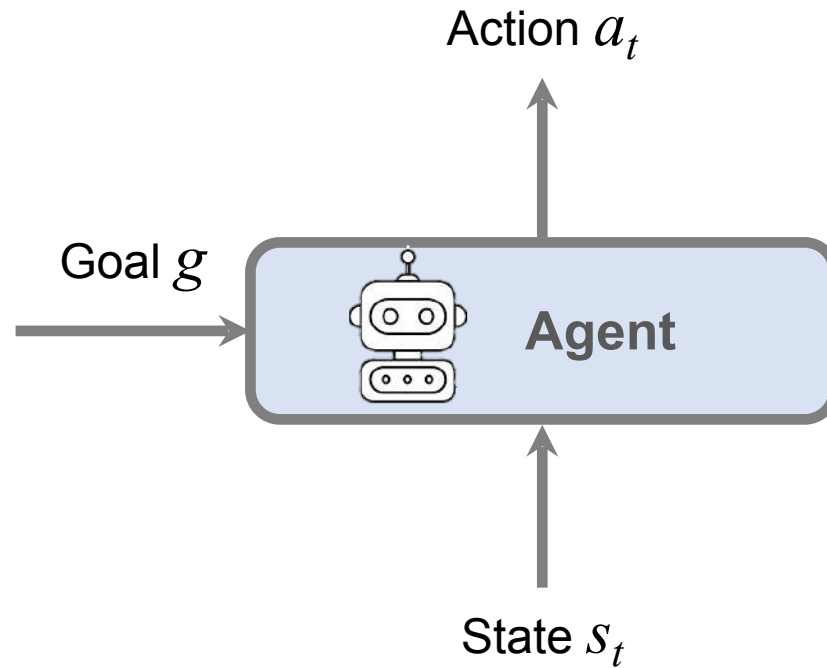
Markov Decision Process



Input
 $\langle g, s_0 \rangle$

Goal
Interpretation

Markov Decision Process



Output

$\langle a_1, a_2, \dots, a_t \rangle$

Action
Sequencing

$\langle s_1, s_2, \dots, s_t \rangle$

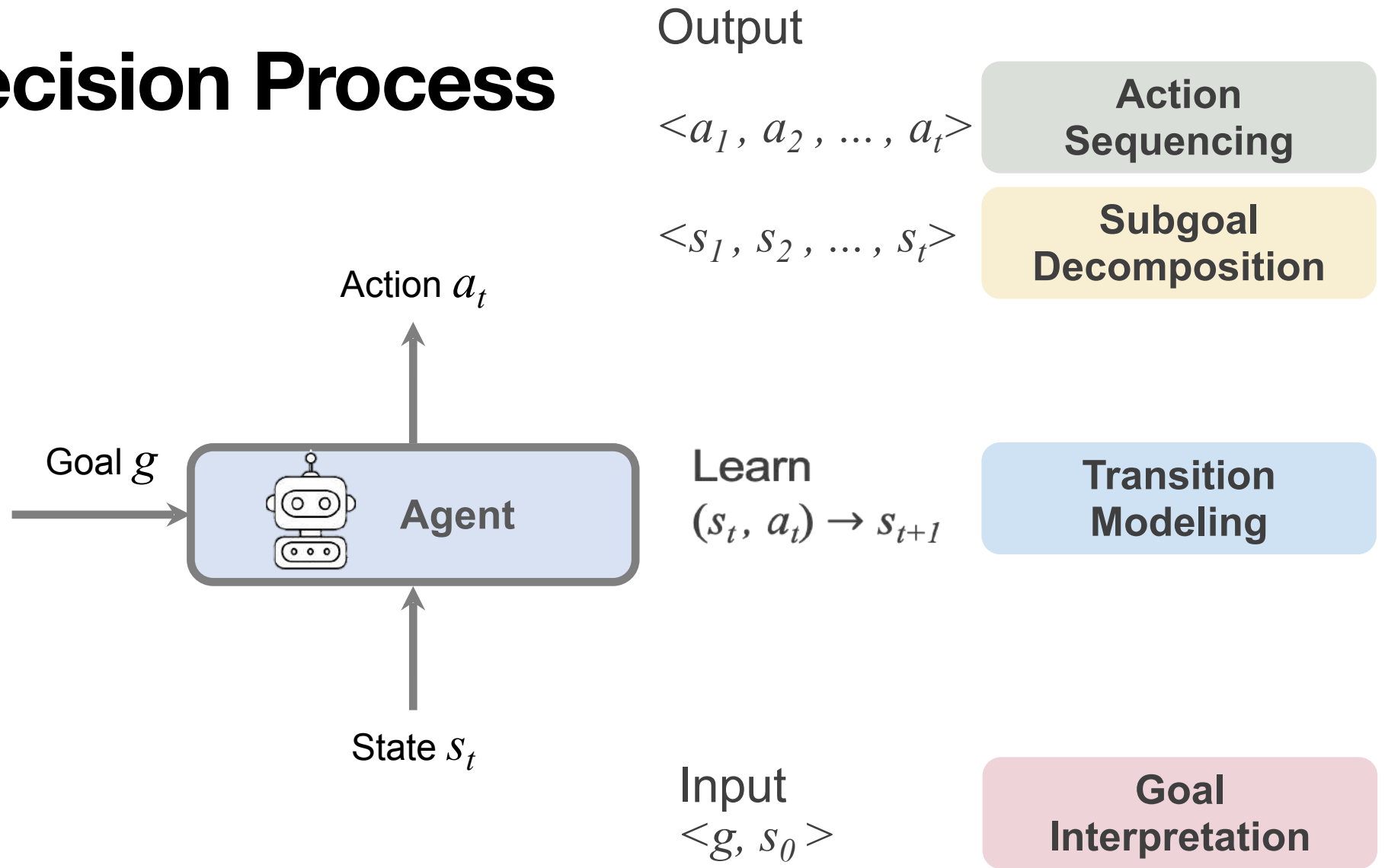
Subgoal
Decomposition

Input

$\langle g, s_0 \rangle$

Goal
Interpretation

Markov Decision Process



Embodied Agent Interface

**Action
Sequencing**

**Subgoal
Decomposition**

**Transition
Modeling**

**Goal
Interpretation**

Embodied Agent Interface

**Goal
Interpretation**

**Subgoal
Decomposition**

**Action
Sequencing**

**Transition
Modeling**

LLMs

Embodied Agent Interface

Goal
Interpretation

Subgoal
Decomposition

Action
Sequencing

Transition
Modeling

Embodied Environments:

BEHAVIOR

VirtualHome

...

Embodied Agent Interface

A Single Line of Code `eai-eval --all`



<https://embodied-agent-interface.github.io>



<https://huggingface.co/datasets/Inevitablevalor/EmbodiedAgentInterface>



<https://github.com/embodied-agent-interface/embodied-agent-interface>



<https://hub.docker.com/r/jameskrw/eai-eval>



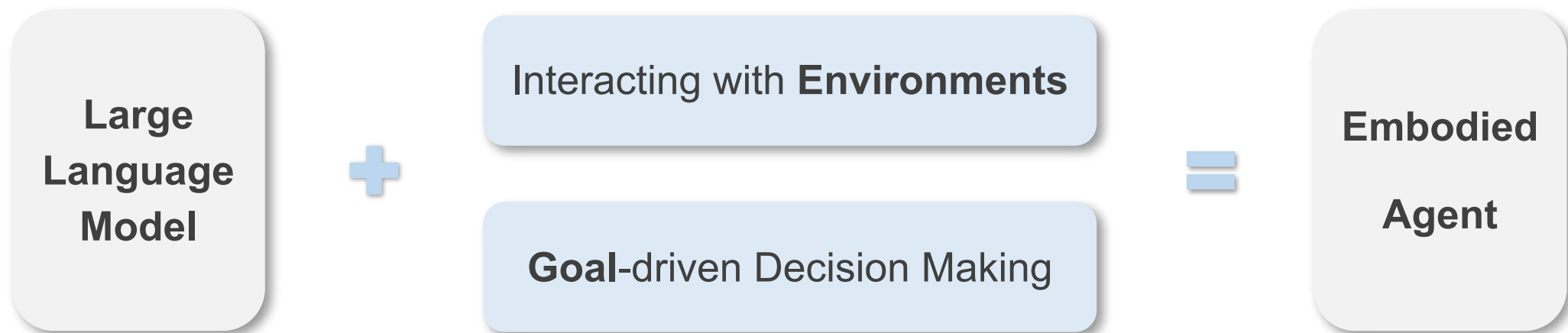
<https://pypi.org/project/eai-eval>



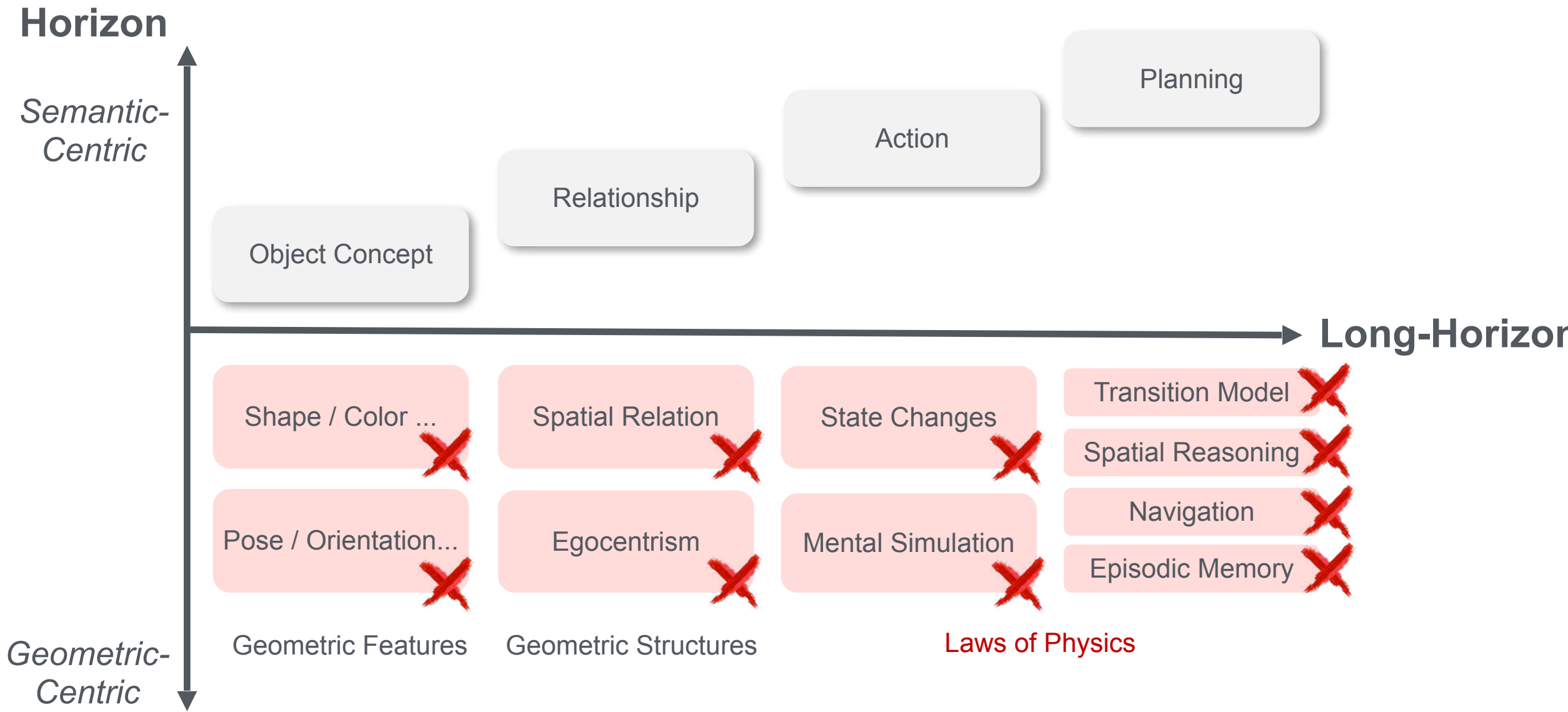
<https://embodied-agent-eval.readthedocs.io>



Future



Missing knowledge about physical world



My Goal: Physical World Knowledge

Horizon

*Semantic-
Centric*

Long-Horizon

Go to **Longer** Horizon!

Go to **Lower** Level!

*Geometric-
Centric*



Foundation Models Meet Embodied Agents

<https://foundation-models-meet-embodied-agents.github.io/>



Manling Li
Northwestern



Yunzhu Li
Columbia



Jiayuan Mao
MIT



Wenlong Huang
Stanford



Northwestern
University