CVPR JUNE 3-7, 2026 DENVER COLORADO

RENMIN UNIVERSITY OF CHINA 中国人民大学

$\mathbf{R^2VLM}$: **R**ecurrent **R**easoning with **V**ision-**L**anguage **M**odels for Estimating Long-Horizon Embodied Task Progress

Yuelin Zhang, Sijie Cheng, Chen Li, Zongzhao Li, Yuxin Huang, Yang Liu, Wenbing Huang†

CVPR 2026

Motivation

Challenges in Using VLMs to Estimate Long-Horizon Embodied Task Progress

1. **Complex Subtask Decomposition and Temporal Dependencies**
2. **Long Execution Duration** — Full video input introduces substantial computational overhead

Simple task such as "Boil water" contains multiple subtasks:

1. Locate the kettle
2. Pick up the kettle
3. Fill the kettle with water
4. Place the kettle on the stove
5. Start heating

The whole process may last several minutes or longer.

Contribution

Key Proposals of R^2VLM

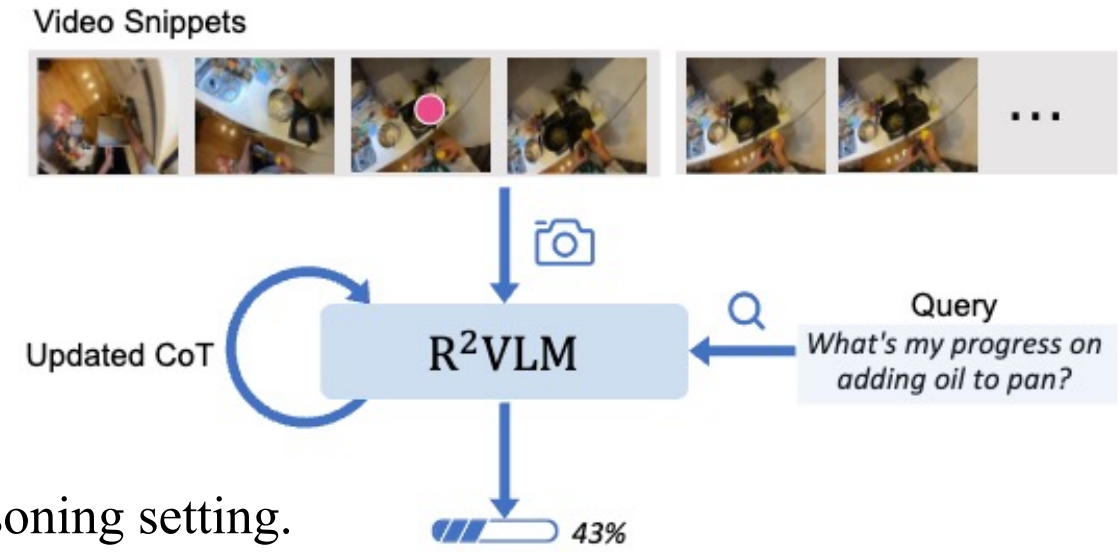
- Receive streaming video and perform periodic reasoning.

  => Improve response speed.

- Introduce chain-of-thought
  - Improve reasoning performance on complex tasks.

    => Enhance model performance.
  - Serve as memory for recurrent reasoning.

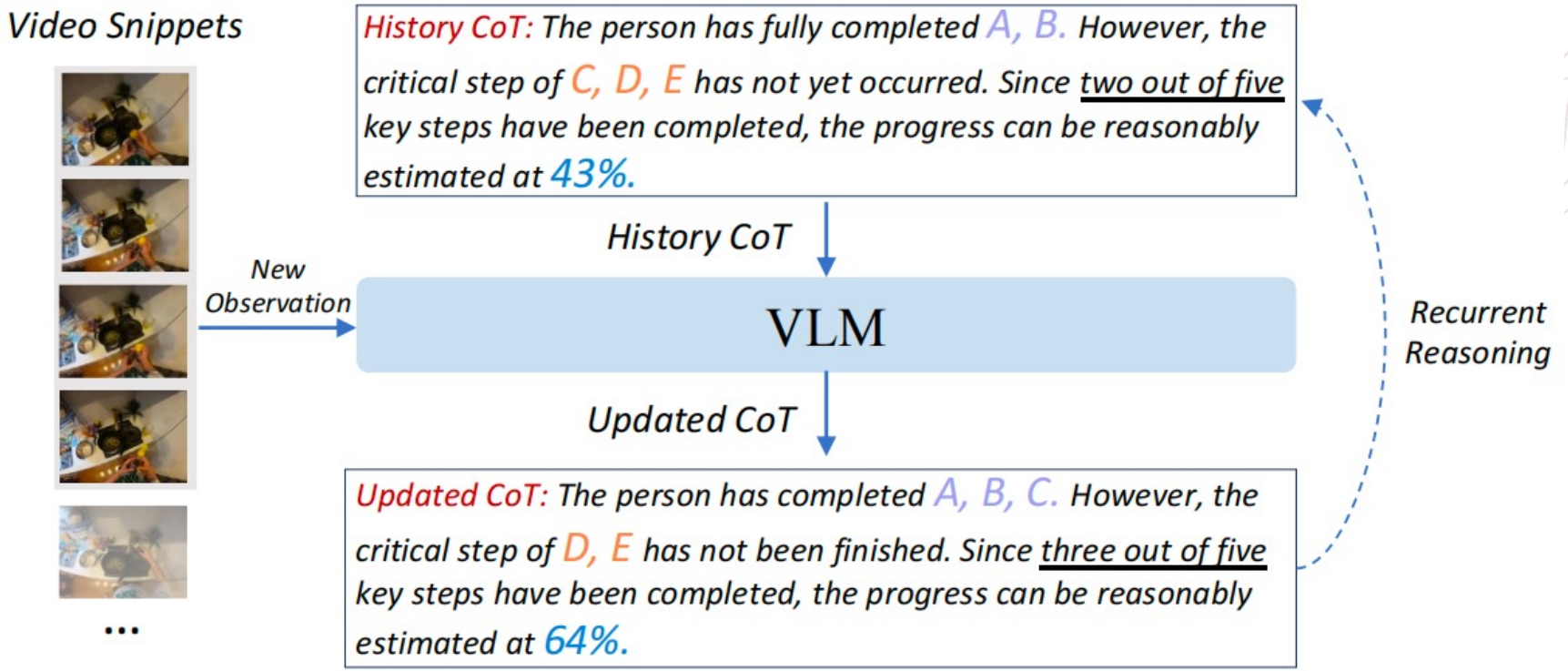
    => Preserve historical information in recurrent reasoning setting.



CoT Structure in Recurrent Reasoning

CoT — textual memory in recurrent reasoning

1. Decomposition of long-horizon tasks
2. Sub-steps categorized as: completed, in progress, and not yet completed.
3. Progress estimation based on the proportion of completed sub-steps.



Dataset

Alfred (Simulator)

Turn on the desk lamp and pick up the alarm clock.

Ego4D (Real World)

Pour water into the blender.

Training

Base Model： Qwen2.5-VL-7B-Instruct

Stage 1: Supervised Fine-Tuning (SFT) with Generated Cold-Start Data

- We used Qwen2.5-VL-72B-Instruct to generate cold-start Chain-of-Thought (CoT) data.

Stage 2: Reinforcement Learning (RL) for Enhanced Reasoning

- We adopted the GRPO algorithm to further enhance the model's reasoning capabilities.
- We carefully designed multi-dimensional outcome reward functions.

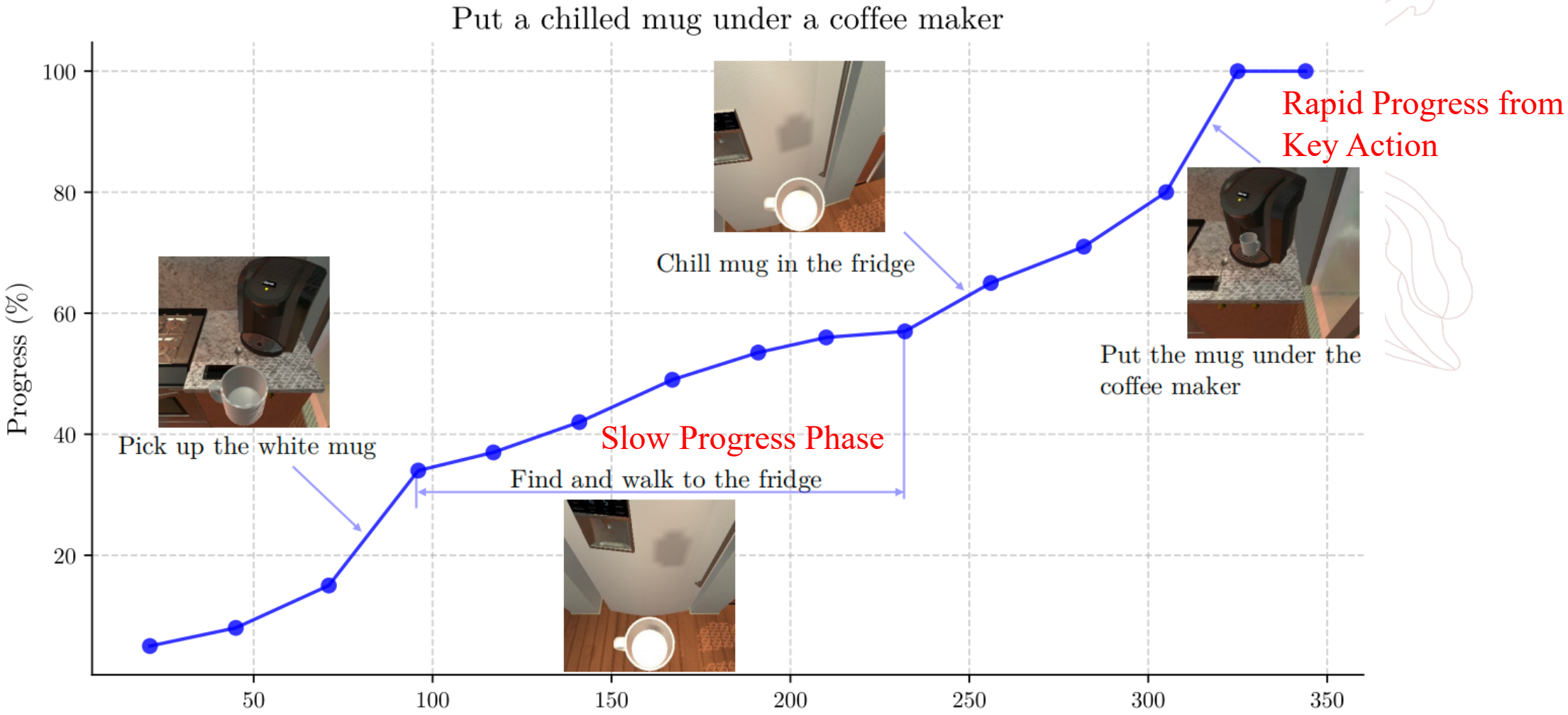
Training Setup

- Training was conducted on 8 × A100 GPUs.
- Stage 1 training time: 50 hours
- Stage 2 training time: 75 hours

Main Results

Model	Size	Alfred				Ego4D			
		p_{mae} ↓	Δp_{mae} ↓	bin ↑	acc ↑	p_{mae} ↓	Δp_{mae} ↓	bin ↑	acc ↑
API-Based Models									
GPT-5 [19]	-	18.35	15.48	0.505	0.901	25.04	15.89	0.259	0.749
Gemini-2.5-Pro [5]	-	16.27	17.37	0.481	0.830	28.22	16.71	0.217	0.713
Qwen3-VL-Plus [26]	-	27.30	15.29	0.379	0.777	28.20	10.14	0.251	0.652
Open-Source Models									
MiniCPM-V-2.6 [27]	8B	27.47	22.04	0.478	0.373	30.17	18.03	0.208	0.482
GLM-4.1V-Thinking [9]	9B	22.67	17.47	0.329	0.548	30.45	8.21	0.197	0.488
InternVL3 [30]	8B	36.09	13.56	0.311	0.621	29.43	16.80	0.187	0.511
InternVL3 [30]	78B	30.54	16.65	0.360	0.765	28.16	13.37	0.204	0.586
Qwen2.5-VL [1]	7B	27.87	10.67	0.295	0.494	28.32	10.11	0.206	0.485
Qwen2.5-VL [1]	72B	24.88	18.68	0.342	0.781	26.88	11.11	0.254	0.624
Specific Methods									
LIV [16]	<1B	35.87	32.44	0.153	0.490	34.57	24.51	0.151	0.485
ROVER [20]	-	29.43	11.52	0.308	0.575	39.83	6.48	0.174	0.548
GVL-SFT [17]	7B	6.21	8.39	0.830	0.930	26.80	13.26	0.255	0.665
R^2VLM-Zero	7B	7.21	7.78	0.742	0.490	23.58	10.97	0.256	0.676
R^2VLM-SFT	7B	2.77	2.27	0.876	0.961	23.03	3.98	0.266	0.685
R^2VLM	7B	**2.19**	**2.17**	**0.917**	**0.988**	**19.25**	**3.88**	**0.318**	**0.761**

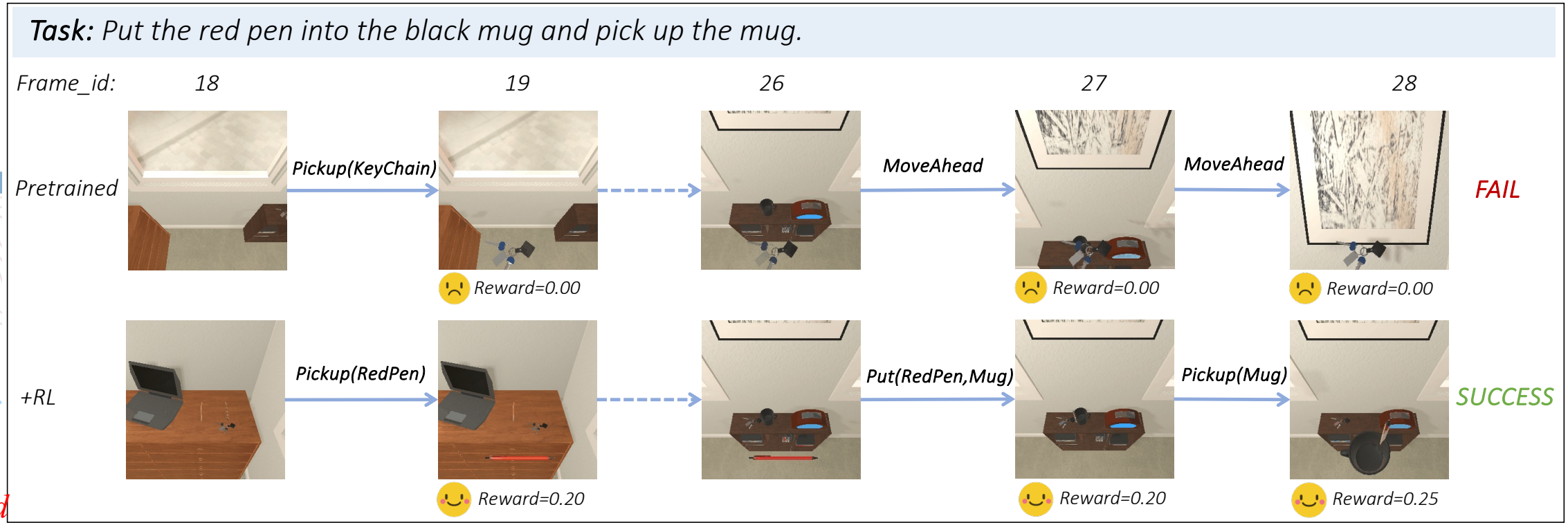
1. Progress Estimation



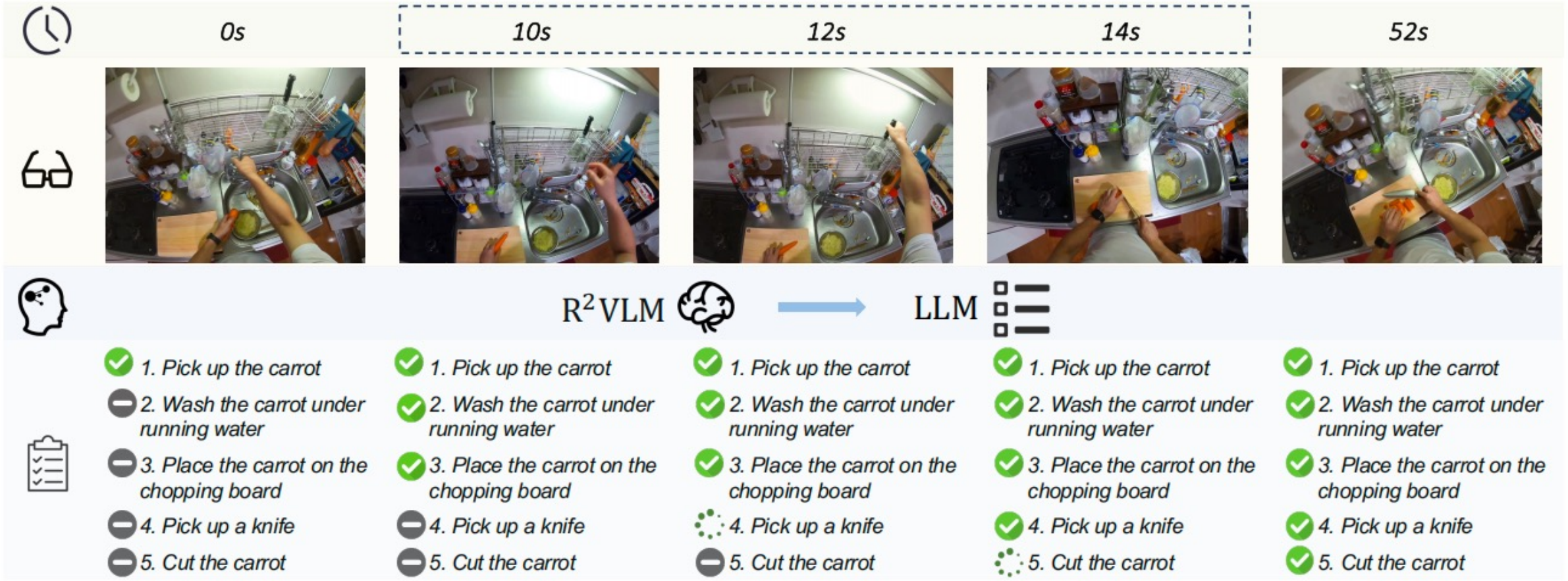
Non-linear Progress in Long-Horizon Tasks

2. Reward Modeling

- We define the progress increment induced by action as the reward signal
- Use this reward to guide the agent's reinforcement learning process



3. Proactive Assistant



Thank you!

Github

Huggingface

Contact Email:
zhangyuelin@ruc.edu.cn