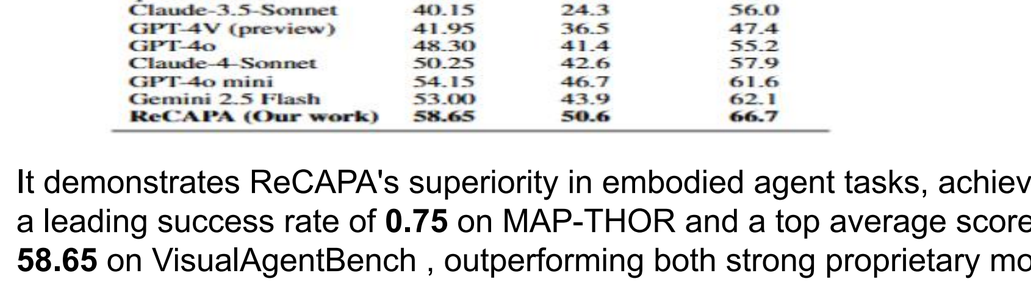
ReCAPA: Hierarchical Predictive Correction to Mitigate Cascading Failures

Xiyin Zeng*, Yuyu Sun*, Haoyang Li*, Shouqiang Liu, Hao Wang†
The Hong Kong University of Science and Technology (Guangzhou)
South China Normal University

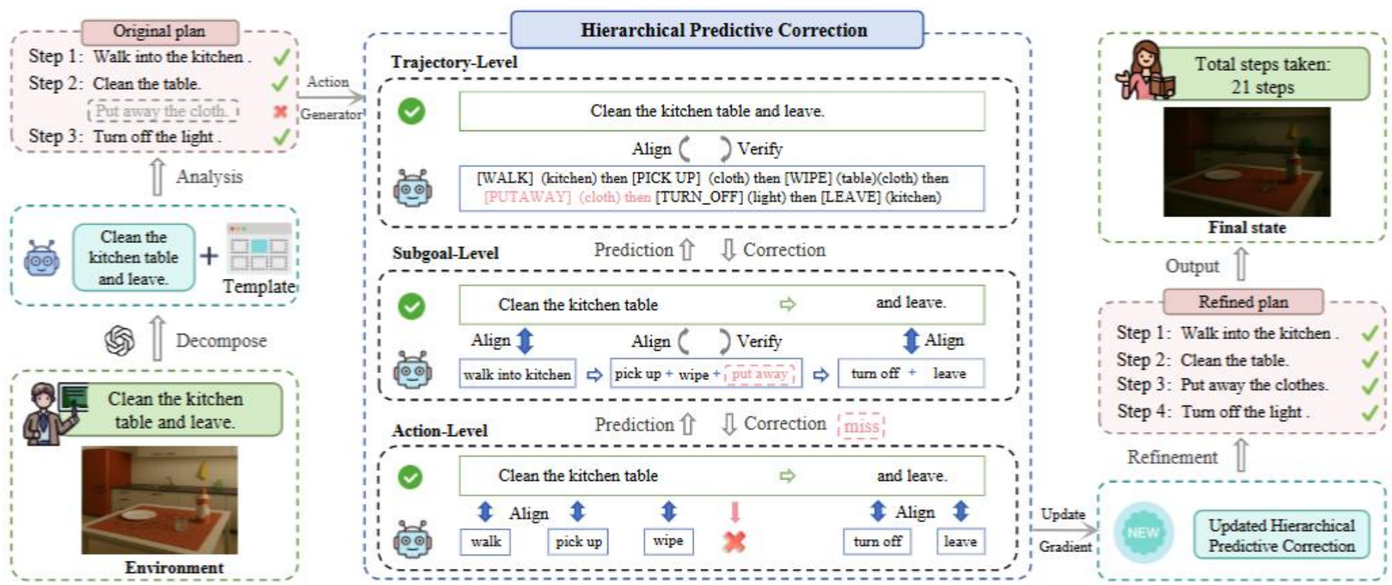
Motivation

- Local errors propagate through subsequent steps and eventually accumulate into cascading failures in long-horizon reasoning.
- Relying only on local-level alignment without global corrective signals leads to each step being optimized in isolation, thereby drifting from the overall intent easily.



Method

Overview of the ReCAPA framework



Results

Table 1: Performance on MAP-THOR across models and metrics. MAP-THOR assessed via Success Rate (SR), Transport Rate (TR), Coverage, and Balance; Coverage measures successful interactions, while Balance captures the evenness of contributions to subtasks.

Model	SR	TR	Coverage	Balance
Single-LM/Agent Baselines				
ReAct	0.34	0.72	0.92	0.67
CoT	0.14	0.59	0.87	0.62
SmartLLM	0.11	0.23	0.91	0.45
CoELA	0.25	0.46	0.76	0.73
Multi-Modal/LLM-Enhanced Baselines				
GPT-4o	0.51	0.85	0.95	0.83
LLaVA	0.54	0.84	0.91	0.75
IDEFICS-2	0.57	0.86	0.94	0.78
CogVLM	0.61	0.89	0.95	0.80
GPT-4V	0.66	0.91	**0.97**	0.82
LLaMAR	0.68	0.90	0.95	0.85
ReCAPA	**0.75**	**0.93**	0.95	**0.93**

Table 2: Performance of different models on VisualAgentBench which include OmniGibson and Minecraft. AVG. denotes the overall average score.

Model	AVG.	OmniGibson	Minecraft
Open-LMMs (Fine-tuning)			
Qwen-VL	9.90	1.7	18.1
CogVLM2	13.55	6.6	20.5
LLaVA-NeXT	16.60	9.4	23.8
GLM-4V	14.35	8.8	19.9
InternVL-2	22.20	16.0	28.4
Proprietary-LMMs (Prompting)			
qwen-vl-max	2.65	0.0	5.3
Claude-3.5-Sonnet	40.15	24.3	56.0
GPT-4V (preview)	41.95	36.5	47.4
GPT-4o	48.30	41.4	55.2
Claude-4-Sonnet	50.25	42.6	57.9
GPT-4o mini	54.15	46.7	61.6
Gemini 2.5 Flash	53.00	43.9	62.1
ReCAPA (Our work)	**58.65**	**50.6**	**66.7**

- It demonstrates ReCAPA's superiority in embodied agent tasks, achieving a leading success rate of **0.75** on MAP-THOR and a top average score of **58.65** on VisualAgentBench , outperforming both strong proprietary models like GPT-4o mini and specialized baselines like LLaMAR.

Contributions

- **We propose ReCAPA, a framework operationalizes hierarchical correction by coupling multi-level predictive representations with prompt–trajectory distributional alignment, allowing deviations to be anticipated and corrected earlier.**
- **We introduce two diagnostic metrics for error propagation in long-horizon reasoning: EPR quantifies the propagation of errors across future steps, while PAC captures the system's ability to recover by measuring how quickly post-error dissipates.**
- **ReCAPA outperforms strong LMM baselines in terms of success rate, achieving +5.65% on VisualAgentBench, +9% on MineDojo, and +7% on MAP-THOR.**

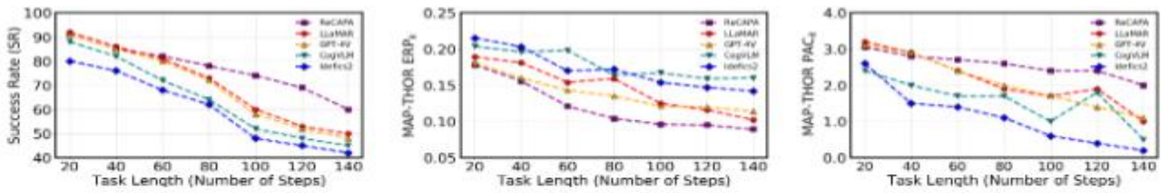


Figure 3: Left: Success rate curves across varying task lengths. Middle: EPR trends showing error propagation at different lags. Right: PAC decay rates on MAP-THOR.

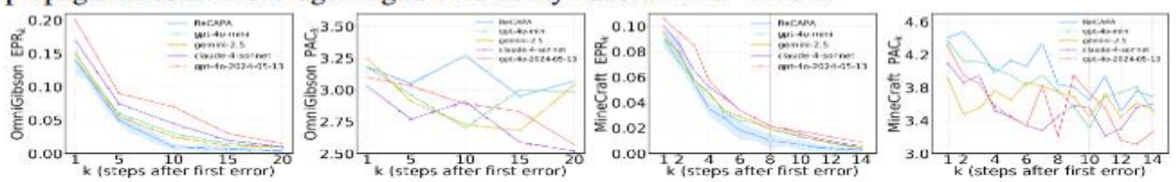


Figure 4: Results on VisualAgentBench. The left two plots show the EPR and PAC curves on OmniGibson, while the right two plots show the EPR and PAC curves on MineCraft. Shaded regions indicate 95% confidence intervals across three random seeds.

Contact

xzeng310@connect.hkust-gz.edu.cn