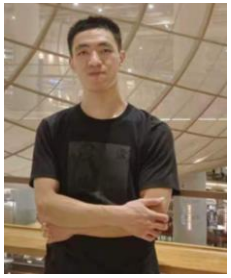
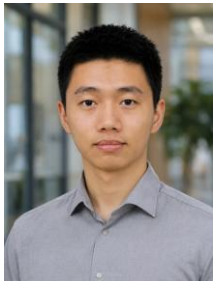
Resilience Matters for Embodied Agents System: New Metrics, Systematic Evaluation, and Optimization








Yapeng Liu[1,2,3] Yuanzhao Zhai[1,2] Xudong Gong[1,2] Dawei Feng[1,2] Bo Ding[1,†] Addison Lin Wang[3,†] Huaimin Wang[1,2]

[1]PDL Lab, College of Computer Science and Technology, National University of Defense Technology.
[2]State Key Laboratory of Complex & Critical Software Environment, Changsha, China.
[3]EmPACT Lab, Nanyang Technological University, Singapore.
[†]Corresponding authors: {dingbo@aliyun.com, linwang@ntu.edu.sg}



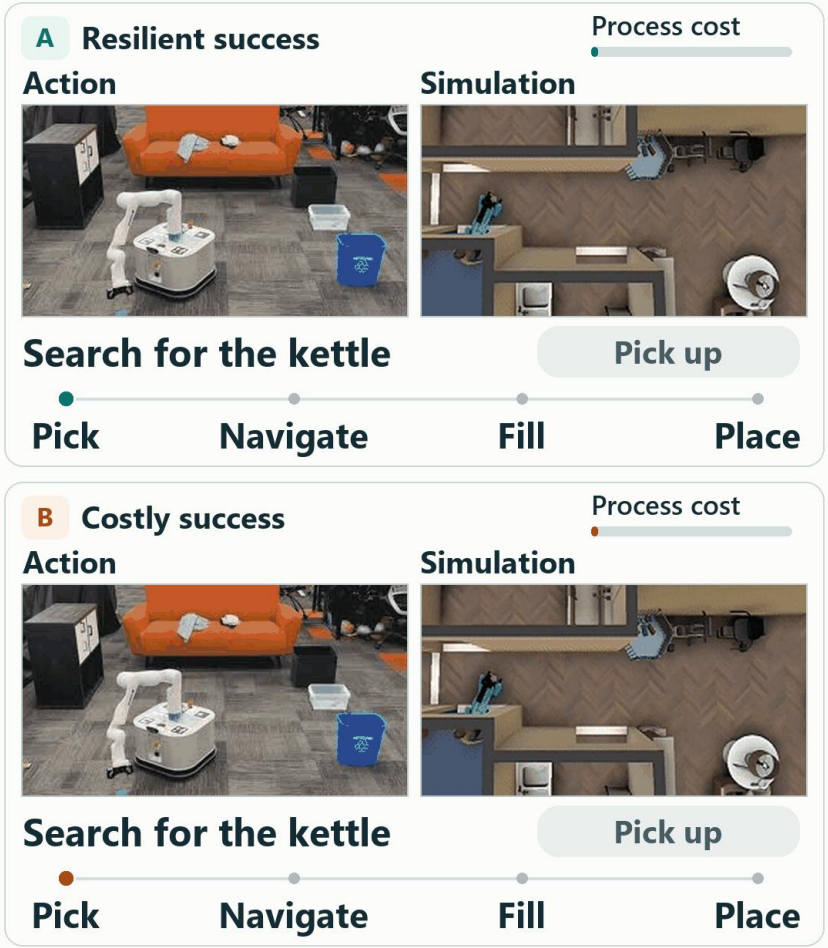
Introduction

① Embodied Agent System is increasing implemented in Open-World Physical domains, where reliability dictates deployment quality and trust [1].

② A key challenge is ***how to distinguish*** **trustworthy adaptation** from 'lucky' **brittle success** under continuous disruptions, as outcome metrics collapse execution dynamics into coarse scores.

③ The property we define as **EAS Resilience** — how EASs recover, stabilize, and extend under stress, also we prove its inherency.

Same Outcome,
Different Resilience EAS execution.



Agent A, B show the efficient adaptation and costly retry execution (***Recovery***). [2]

[1] Chang, M., Chhablani, G., Clegg, A., Cote, M. D., Desai, R., Hlavac, M., ... & Yang, T. Y. PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
[2] David D. Woods, *Four Concepts for Resilience and the Implications for the Future of Resilience Engineering* (Reliability Engineering & System Safety, 2015).

Metrics

We gain insights from **System Resilience-Engineering concepts** to EAS groundings [Woods, 1], form resilience evaluation, step as ***Metrics, Evaluation and Optimization***.

EAS → E? A? S!

- **Rebound**

  How much extra work does a recovery take?

- **Stability**

  How sensitive is execution to an episode execution disruptions?

- **Graceful Extensibility**

  How much stress can EAS in a task family tolerate?

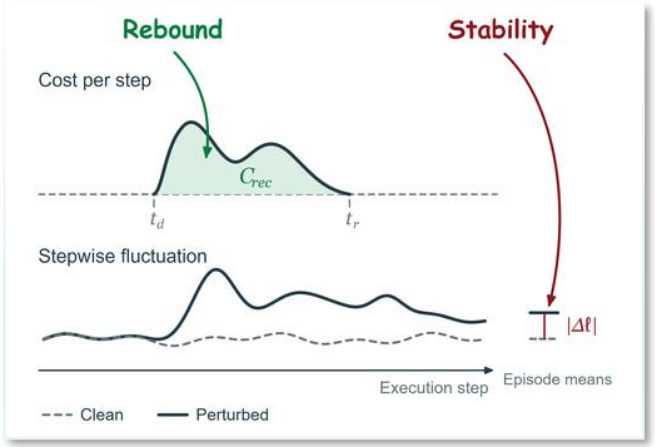
[1] David D. Woods, *Four Concepts for Resilience and the Implications for the Future of Resilience Engineering* (Reliability Engineering & System Safety, 2015).

Metrics

- **Rebound**

Unified Cognitive, Physical, State debt **recovery cost** upon statistical execution baseline.

$$C_{\text{rec}}^{(i)} = \frac{1}{W_{\text{rem}}^{*}(a_{t_d}) + \epsilon} \sum_{t=t_d}^{t_r} \frac{\Delta\tau_t}{\tau^{*}(a_t) + \epsilon} \cdot g_t^{\text{rec}}$$



(a) Recovery cost, Exec-Stability.

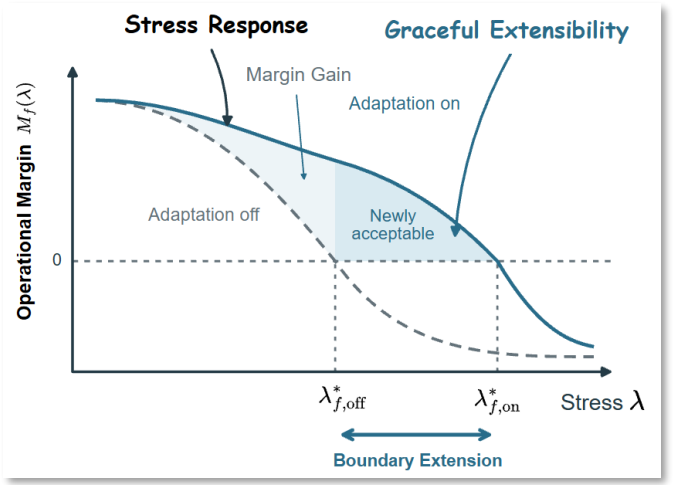
- **Stability**

Measure **sensitivity** across EAS execution-level over progress, planning cot, and LLM critic.

$$\beta_f = \sup_{z \in \mathcal{Z}_f, \rho \in \mathcal{P}_f} \frac{|\ell(\pi, \rho(z)) - \ell(\pi, z)|}{\|\rho\| + \epsilon}$$

- **Graceful Extensibility**

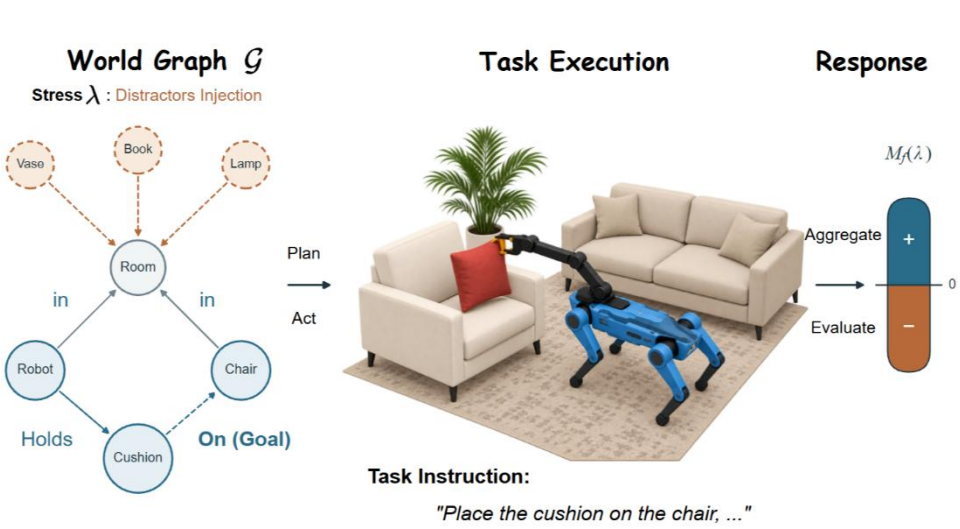
Form the stress-response margin in operational boundary and capture **graceful extensibility capacity** extension.

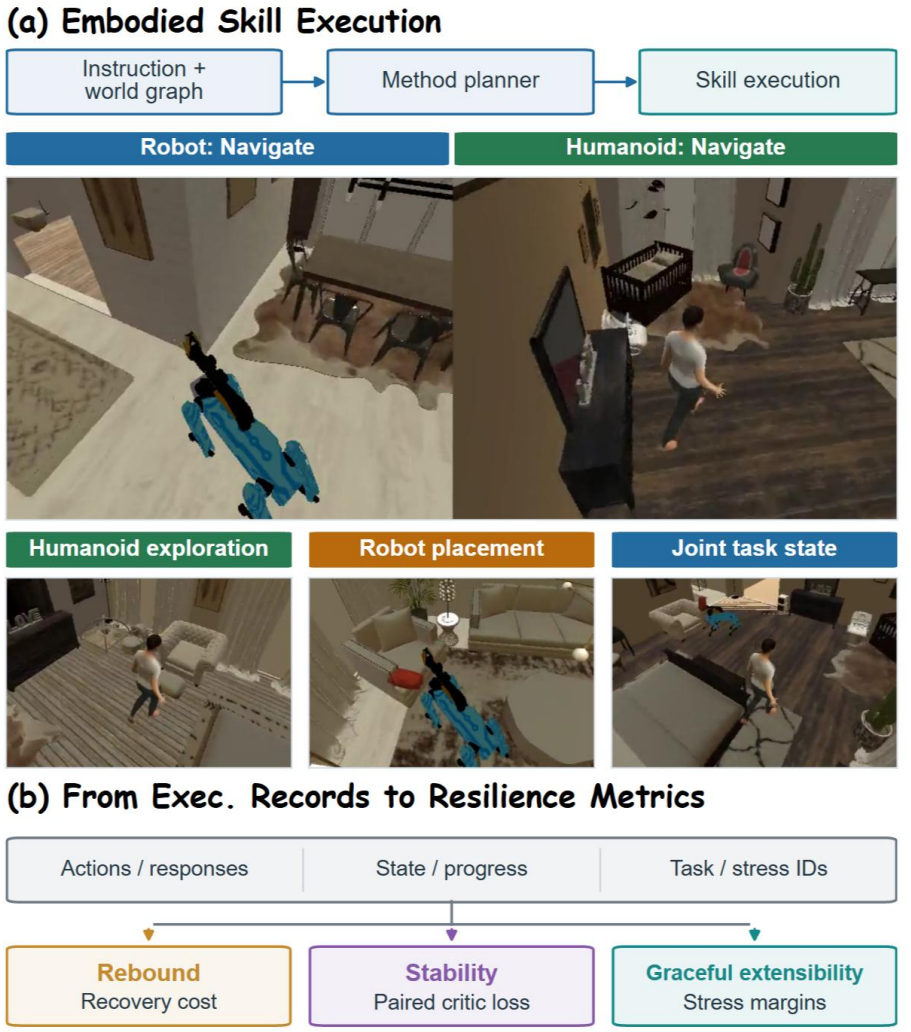


(b) Adaptation extends EAS capacity.

Evaluation

1) Integrated Extraction. 2) Modular Sim. Collect.

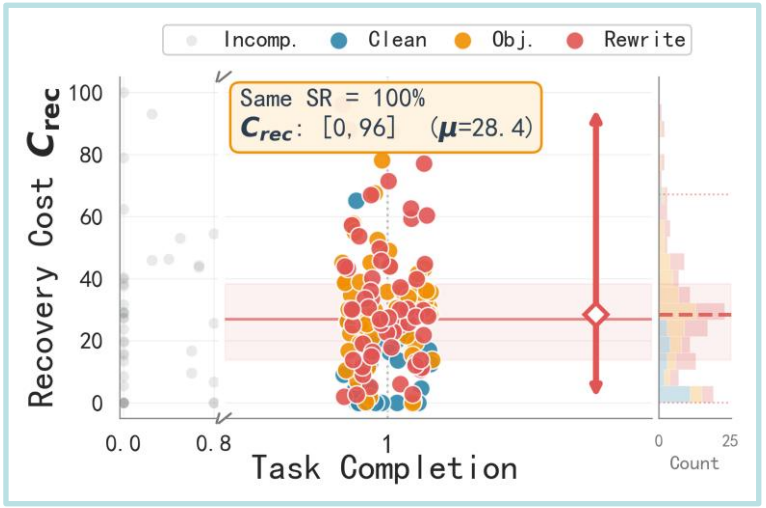




① Execution Instrumentation
② Unified Trace Logging
③ **Event Extraction**
④ Trace Structure
⑤ **Resilience Metrics Computation**

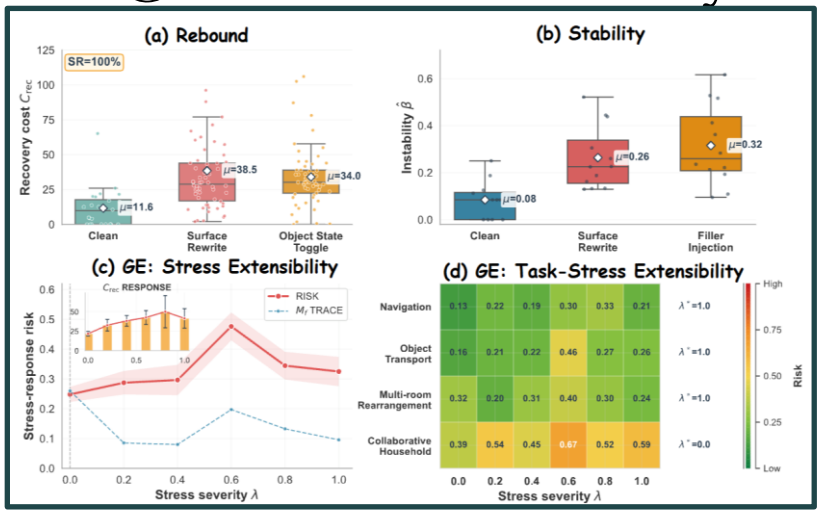
Evaluation

① Metric Validation



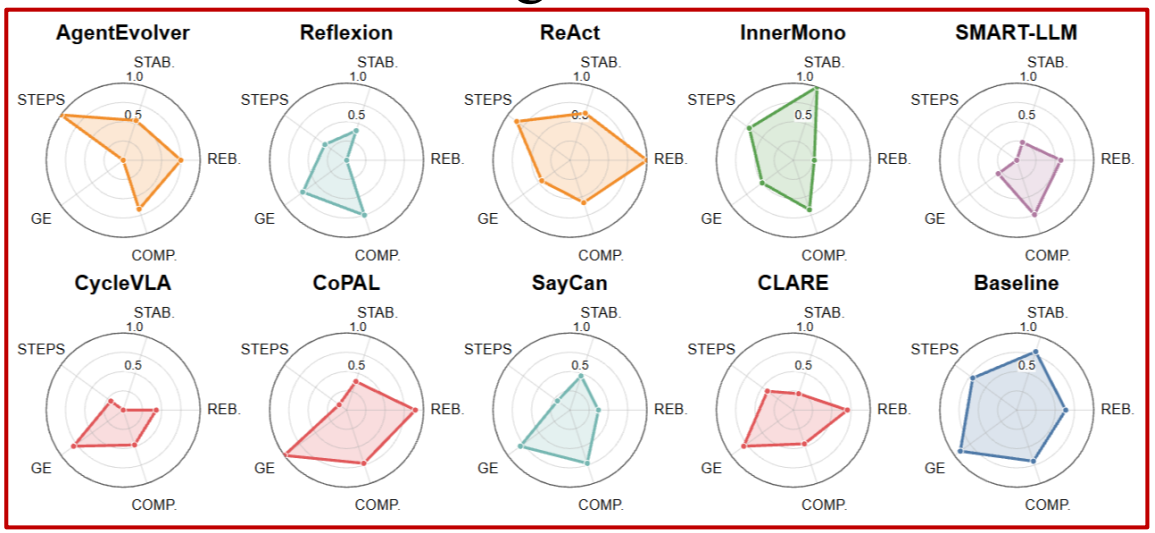
At full completion, recovery costs span **0-96**.

② Statistical Reliability

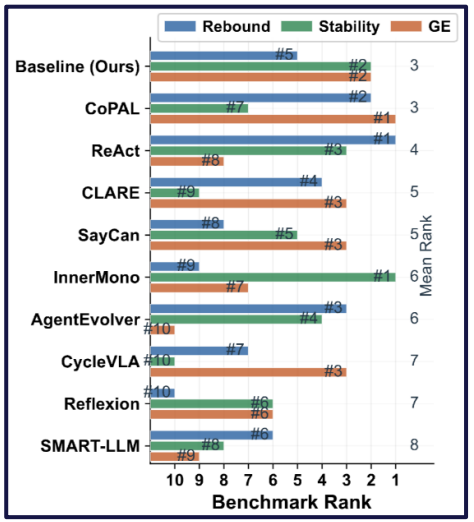


Stress execution suite shows the **discrepancy** in statistics.

③ Resilience Signatures

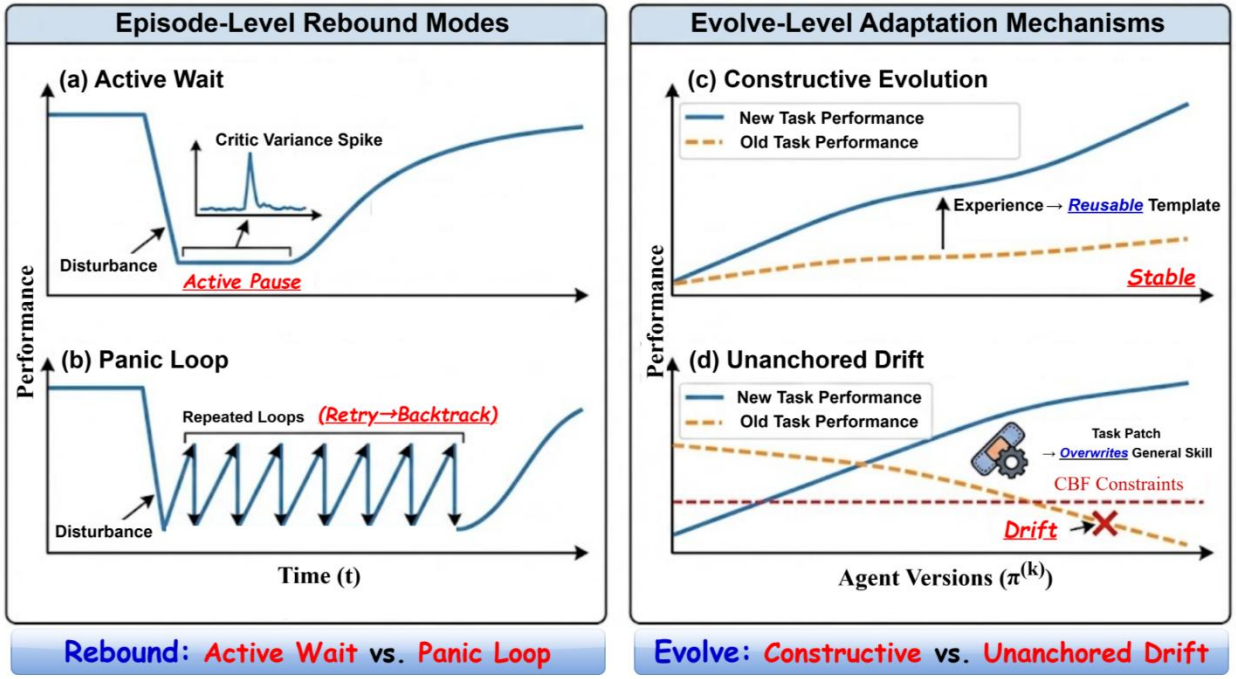


④ Benchmark Rank



Optimization

1) Resilience Diagnosis (Ep.139/141)



Rebound signature diagnosis:

(a) **Active Wait** (b) **Repeated Retries**

Evolve signature diagnosis:

(c) **Constructive Evolve**

(d) **Repeated Retries**

2) Resilience-Guided Optimization

Target	Methods	Rebound C_{rec} ↓	Rec. Win. ↓	β_{step} ↓	β ↓	GE. Comp. ↑	Formal GE
Rebound	Reflexion	59.11	3.13	0.2881	0.6834	0.887	–
	+Optimization	**33.73** ↓42.94%	**2.38** ↓23.96%	0.2079 ↓27.84%	0.6870 ↑0.53%	0.819 ↓7.67%	–
Stability	CoPAL	51.01	2.63	0.3185	0.6788	0.903	–
	+Optimization	43.01 ↓15.68%	3.00 ↑14.07%	**0.2552** ↓19.87%	**0.6772** ↓0.24%	0.750 ↓16.94%	–
GE	SMART-LLM	36.70	1.33	1.4594	0.6917	0.833	Incomplete
	+Optimization	99.59 ↑171.36%	2.25 ↑69.17%	0.7215 ↓50.56%	0.6371 ↓7.89%	**0.917** ↑10.08%	Complete†

† Formal GE is reported only for GE stress-tests, since it requires a stress-response curve over λ grid.

Thanks for watching!

Takeaway:

For embodied agents in the real world, outcome is not enough. By quantifying **Rebound, Stability, and Graceful Extensibility**, we systematically diagnose and optimize resilient EAS.