

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Automatic Cognitive Task Generation for In-Situ Evaluation of Embodied Agents

Permalink

<https://escholarship.org/uc/item/13m5d8pv>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

He, Xinyi

yang, ying

Fu, Chuanjian

et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Automatic Cognitive Task Generation for In-Situ Evaluation of Embodied Agents

Xinyi He^{1,2}, Ying Yang², Chuanjian Fu², Sihan Guo¹, Song-Chun Zhu^{1,2}, Lifeng Fan², Zhenliang Zhang² & Yujia Peng^{1,2}

¹State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China

²State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China

Abstract

As agents are poised for widespread deployment, evaluation in unseen environments has become critical. Existing benchmarks, suffering from data contamination and lacking scene specificity, are inadequate for in-situ evaluation. We propose an in-situ task generation method for unseen environments, defining tasks through graph representation and constructing a two-stage interaction-evolution task generation system for embodied agents (TEA). In the interaction stage, the agent interacts with the environment, creating a loop between task execution and generation for continuous generation. In the evolution stage, task graph modeling allows us to recombine and reuse existing tasks to generate new ones. Experiments across 10 scenes demonstrate that TEA generated 87876 tasks in two cycles. Benchmarking models against humans on in-situ tasks reveals that models, despite excelling on public benchmarks, perform poorly on basic perception tasks, lack spatial awareness and show high sensitivity in reasoning. These findings highlight the necessity of in-situ evaluation before real-world deployment.

Keywords: in-situ evaluation, embodied cognition, task generation