LLM-Enhanced Rapid-Reflex Async-Reflect Embodied Agent for Real-Time Decision-Making in Dynamically Changing Environments

Yangqing Zheng*, Shunqi Mao*, Dingxin Zhang*, Weidong Cai
School of Computer Science, The University of Sydney
yzhe6302@uni.sydney.edu.au, {smao7434, dzha2344, tom.cai}@sydney.edu.au

Abstract

In the realm of embodied intelligence, the evolution of large language models (LLMs) has markedly enhanced agent decision making. Consequently, researchers have begun exploring agent performance in dynamically changing high-risk scenarios, i.e., *fire, flood, and wind scenarios in the HAZARD benchmark. Under these extreme conditions, the delay in decision making emerges as a crucial yet insufficiently studied issue. We propose a Time Conversion Mechanism (TCM) that translates inference delays in decision-making into equivalent simulation frames, thus aligning cognitive and physical costs under a single FPS-based metric. By extending HAZARD with Respond Latency (RL) and Latency-to-Action Ratio (LAR), we deliver a fully latency-aware evaluation protocol. Moreover, we present the Rapid-Reflex Async-Reflect Agent (RRARA), which couples a lightweight LLM-guided feedback module with a rule-based agent to enable immediate reactive behaviors and asynchronous reflective refinements in situ. Experiments on HAZARD show that RRARA substantially outperforms existing baselines in latency-sensitive scenarios.*

1. Introduction

Recent advances in large language models (LLMs) have enabled promising applications in autonomous decision-making agents [2, 9, 13, 16, 19, 20]. Most embodied AI frameworks [6, 8, 10, 12, 17] focus on planning and decision quality under static conditions following a perceive-think-act paradigm. At inference time, agents pause to reason before acting, which is costly in dynamic environments where even brief delays can lead to outdated decisions. This limitation becomes especially critical in scenarios like fire rescue, as shown in Fig. 1. Such high inference latency produces stale observations and obsolete context, causing misaligned or suboptimal behaviors.

While several methods consider latency issues in embodied AI [3, 14, 15], they primarily focus on low-level control rather than high-level reasoning and planning. Meanwhile, existing benchmarks [1, 4, 5, 7, 11] adopt largely static environments, where object positions remain unchanged and the impact of inference time is minimal. The recent HAZARD benchmark [18] simulates dynamic fire, flood, and wind disaster scenarios, yet it still follows the common practice of ignoring inference latency during agent evaluation.

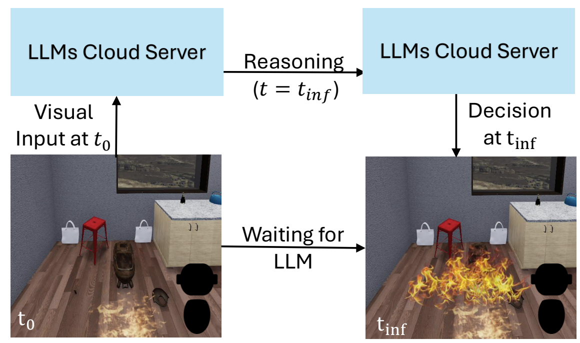

Figure 1. Environment changes during inference can lead to outdated responses, *e.g.*, object is burnt before execution completes.

To fill this gap, we introduce the Time Conversion Mechanism (TCM), which translates inference delays into equivalent simulation frames and unifies reasoning and execution costs under a single FPS-based metric. We then introduce Rapid-Reflex Async-Reflect Agent (RRARA), a hybrid agent where rapid reflexive policies trigger immediate actions while an asynchronous LLM Reflector analyzes and refines those actions in situ. Integrated with HAZARD, RRARA quantifies the cost of deliberation via TCM and demonstrates its ability to revise suboptimal choices during dynamic rescue operations.

2. Time Conversion Mechanism

In standard HAZARD, agent performance is measured solely by the number of frames spent executing actions, decoupling reasoning time from environmental progression. The Time Conversion Mechanism (TCM) remedies this by mapping inference delay into simulation frames: $F_{\text{inf}} =$

*These authors contributed equally to this work.

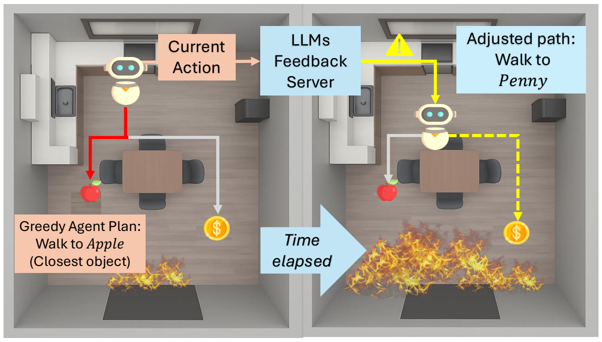

Figure 2. Illustration of the reflective process in RRARA. The low-latency policy greedily selects the closest object (Left). After receiving feedback from the LLM Reflector, it pursues a more valuable object instead (Right).

$T_{\text{inf}} \times \text{FPS}$. Here, T_{inf} denotes the agent's decision latency in seconds, and FPS is the simulation frame rate. The agent observes at T_0 and reasons on its next action A_{T_0}, incurring a latency of T_{inf} seconds. Concurrently, the environment continues evolving independently, and A_{T_0} is executed at $T_0 + T_{\text{inf}}$, increasing the risk of acting on outdated information. By integrating inference-time, TCM aligns the evaluation with real-world constraints, penalizing slow deliberation in fast-evolving environments. Consequently, agents must balance accuracy with efficiency, as extended reasoning directly reduces the time available for rescue operations.

3. Rapid-Reflex Async-Reflect Agent

To address the real-time responsiveness challenges posed by inference latency issues, as quantified by our proposed TCM, we introduce Rapid-Reflex Async-Reflect Agent (RRARA), a training-free embodied agent designed for dynamic environments. As illustrated in Fig. 2, RRARA combines a low-latency rule-based policy with an LLM-based Reflector that simultaneously reflects on the ongoing decisions and delivers feedback with in-depth reasoning.

Specifically, upon perceiving the environment, the agent executes an initial action determined by a rule-based policy with negligible latency—for example, walking toward an object in the center of the room. In parallel, the LLM-based Reflector receives details of current and prior actions along with observations of visible objects, and reflects on whether the ongoing action remains suitable for the current situation. If the Reflector validates the current action, the agent proceeds without interruption; otherwise, it interrupts the reflex and switches to the suggested alternative. The Reflector perpetually evaluates action outcomes and triggers new LLM-based reasoning immediately as each reflexive action begins. This parallel reflect-and-feedback mechanism enhances decision quality without introducing additional inference latency. By interleaving immediate reflexes with high-level reflection, RRARA achieves real-time responsiveness while integrating the high-level reasoning capabilities of the LLM, allowing for refined decision-making without sacrificing responsiveness.

Agent	VR ↑	DR↓	RL (s) ↓	LAR ↓
Rule	0.20	0.33	0.00	0.00
Greedy	0.22	0.24	0.00	0.00
MCTS	0.09	0.35	3.26	0.57
GPT-3.5	0.20	0.33	2.35	0.50
GPT-4	0.08	0.42	4.11	0.84
GPT-4.1	0.14	0.40	3.85	0.71
Llama-2-7b	0.03	0.90	15.60	0.96
RRARA (Rule)	0.25	0.29	**0.00**	**0.00**
RRARA (Greedy)	**0.29**	**0.23**	**0.00**	**0.00**

Table 1. Evaluation results of the fire hazard scenario [18] with proposed TCM. Including Value Rate (VR), Damage Ratio (DR), Respond Latency (RL), and Latency-to-Action Ratio (LAR).

4. Experiments and Discussion

We introduce two additional metrics to enhance HAZARD [18]: *Respond Latency (RL)* and *Latency-to-Action Ratio (LAR)*. RL measures the average inference time per decision step, while LAR quantifies the proportion of time spent reasoning relative to acting. As these metrics can vary with hardware, all experiments are conducted on a system with Intel Core i7-11700 and a single NVIDIA GeForce RTX 3090 GPU. We perform experiments on the fire scenario of HAZARD benchmark [18], with Rule and Greedy in [18] as the reflex policy and GPT-3.5 serving as the Reflector in RRARA. The HAZARD simulator operates at 30 FPS, and experiment results are reported in Tab. 1.

It can be observed that LLM-based and MCTS-based agents, despite their sophisticated reasoning capabilities, fail to outperform even a basic rule-based baseline. This supports our hypothesis that in highly dynamic environments, the high inference latency of complex agents results in delayed actions and outdated reasoning. By integrating TCM and deploying RRARA within HAZARD, we observe that RRARA outperforms its individual counterparts—including the Rule and Greedy poliy, and the GPT-3.5–based agent. Beyond outperforming these components in isolation, RRARA also achieves the best performance among all evaluated baselines. Empirical results show that the LLM-based evaluator intervenes in roughly 60% of action steps, steering the agent toward better planning without incurring critical latency.

In conclusion, TCM-assisted evaluation highlights the critical role of real-time responsiveness in embodied agents operating in dynamic environments. Our proposed RRARA demonstrates a simple yet effective paradigm for advancing embodied AI for real-world tasks.

References

[1] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. ARNOLD: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *ICCV*, pages 20426–20438, 2023. 1

[2] Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. Multiply: A multisensory object-centric embodied large language model in 3d world. In *CVPR*, pages 26406–26416, 2024. 1

[3] Yiyang Huang, Yuhui Hao, Bo Yu, Feng Yan, Yuxin Yang, Feng Min, Yinhe Han, Lin Ma, Shaoshan Liu, Qiang Liu, and Yiming Gan. Software-hardware co-design for embodied ai robots. *arXiv preprint arXiv:2407.04292*, 2024. 1

[4] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R. Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A benchmark for embodied AI with 1, 000 everyday activities and realistic simulation. In *CoRL*, 2022. 1

[5] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking llms for embodied decision making. In *NeurIPS*, 2024. 1

[6] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *CVPR*, pages 18061–18070, 2024. 1

[7] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul McVay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent-Pierre Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, pages 16488–16498, 2024. 1

[8] Shunqi Mao, Chaoyi Zhang, Heng Wang, and Weidong Cai. Towards generalisable audio representations for audio-visual navigation. In *CVPR-EAI*, 2022. 1

[9] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In *NeurIPS*, 2023. 1

[10] Yiran Qin, Enshen Zhou, Qichang Liu, Zhenfei Yin, Lu Sheng, Ruimao Zhang, Yu Qiao, and Jing Shao. Mp5: A multi-modal open-ended embodied system in minecraft via active perception. In *CVPR*, pages 16307–16316, 2024. 1

[11] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, pages 10737–10746, 2020. 1

[12] Chan Hee Song, Brian M. Sadler, Jiaman Wu, Wei-Lun Chao, Clayton Washington, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *ICCV*, pages 2986–2997, 2023. 1

[13] Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, R. Devon Hjelm, Zhe Gan, Zsolt Kira, and Alexander Toshev. From multimodal llms to generalist embodied agents: Methods and lessons. *arXiv preprint arXiv:2412.08442*, 2024. 1

[14] Shagun Uppal, Ananye Agarwal, Haoyu Xiong, Kenneth Shaw, and Deepak Pathak. SPIN: simultaneous perception, interaction and navigation. In *CVPR*, pages 18133–18142, 2024. 1

[15] Zishen Wan, Jiayi Qian, Yuhang Du, Jason Jabbour, Yilun Du, Yang Katie Zhao, Arijit Raychowdhury, Tushar Krishna, and Vijay Janapa Reddi. Generative ai in embodied systems: System-level analysis of performance, efficiency and scalability. *arXiv preprint arXiv:2504.18945*, 2025. 1

[16] Qi Zhao, Haotian Fu, Chen Sun, and George Konidaris. EPO: hierarchical LLM agents with environment preference optimization. In *EMNLP*, pages 6401–6415, 2024. 1

[17] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *CVPR*, pages 13624–13634, 2024. 1

[18] Qinhong Zhou, Sunli Chen, Yisong Wang, Haozhe Xu, Weihua Du, Hongxin Zhang, Yilun Du, Joshua B. Tenenbaum, and Chuang Gan. HAZARD challenge: Embodied decision making in dynamically changing environments. In *ICLR*, 2024. 1, 2

[19] Filippo Ziliotto, Tommaso Campari, Luciano Serafini, and Lamberto Ballan. TANGO: training-free embodied AI agents for open-world tasks. *arXiv preprint arXiv:2412.10402*, 2024. 1

[20] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 1