A Visualized Framework for Event Cooperation with Generative Agents

Yuyang Tian [1,2*], **Shunqiang Mao** [1,3*], **Wenchang Gao** [1], **Lanlan Qiu** [1], **Tianxing He** [4, 1, 5†]

[1]Shanghai Qi Zhi Institute
[2]University of Science and Technology of China
[3]Sun Yat-sen University
[4]Tsinghua University
[5]Xiongan AI Institute

tianyuyang@mail.ustc.edu.cn, maoshq@mail2.sysu.edu.cn, gaowenchang@sqz.ac.cn, qiulanlan@sqz.ac.cn, hetianxing@mail.tsinghua.edu.cn

Abstract

Large Language Models (LLMs) have revolutionized the simulation of agent societies, enabling autonomous planning, memory formation, and social interactions. However, existing frameworks often overlook systematic evaluations for event organization and lack visualized integration with physically grounded environments, limiting agents' ability to navigate spaces and interact with items realistically. We develop MiniAgentPro, a visualization platform featuring an intuitive map editor for customizing environments and a simulation player with smooth animations. Based on this tool, we introduce a comprehensive test set comprising eight diverse event scenarios with basic and hard variants to assess agents' ability. Evaluations using GPT-4o demonstrate strong performance in basic settings but highlight coordination challenges in hard variants.

Code — https://github.com/Just-A-Pie/MiniAgentStudio

Introduction

The landscape of artificial intelligence has been profoundly reshaped by recent breakthroughs in Large Language Models (LLMs) (Hong et al. 2024; Singhal et al. 2023). These powerful models have demonstrated an unprecedented ability to achieve human-level performance across a multitude of tasks, ranging from complex reasoning to creative content generation (Chen et al. 2025; Zhou et al. 2023). This capability has, in turn, infused new vitality into the concept of building agent societies. Generative Agent (GA) (Park et al. 2023) demonstrates how LLM-driven agents could autonomously plan, form memories, and engage in believable interactions within a sandbox environment. Subsequent research has focused on enriching the generative agent paradigm by integrating more sophisticated and human-centric elements (Li et al. 2024; Wang, Chiu, and Chiu 2023). Additionally, efforts have extended to downstream applications, including policy simulation (Hou et al. 2025; Park et al. 2024), sociological theory modeling (Yang et al. 2025), and the design of non-player characters (NPCs) in games (Christiansen et al. 2024; Cox and Ooi 2023).

However, there's an absence of systematic test sets to evaluate how well these agents perform in organizing various events, such as a birthday party, in addition to their routine daily plans. Furthermore, current simulations often lack physical grounding, which inherently limits the evaluation of agents' ability to exhibit exquisite interactions with the simulation environment.

Our work makes the following contributions: 1) We develop a visualization platform for simulation play, paired with a map editor, to streamline human evaluation and enable easy customization of environments for future research. 2) We enhance the prior GA framework with on-the-fly planning, basic physics, and item interactions for more realistic simulations. We also introduce a test set and evaluation protocol across 8 diverse task settings to assess agents' event coordination abilities.

Visualization Tool

In this section, we introduce MiniAgentPro, a visualization tool to improve the whole pipeline of agent simulation. Our tool contains two parts, the map editor and the simulation player as shown in Figure 1.

Map Editor The map editor allows users to customize the simulation environment. It offers an intuitive interface supporting a range of basic functions, including placing, moving, and deleting buildings and items. We facilitate the management of complex relationships of items through our container and property system, where items can be nested within others, and users can assign specific properties to elements, reflecting real-world scenarios.

The editor supports environments with up to 15 agents and 100 unique types of items, providing flexibility for complex and dynamic scenarios. The editor saves the environment configuration in CSV and JSON formats, ensuring compatibility with backend simulation.

Simulation Player We incorporate a simulation player, which allows for detailed inspection of each agent's current state, activity history, and interactions with other agents and the map within the environment. The platform is designed to

*These authors contributed equally.
†Corresponding Author

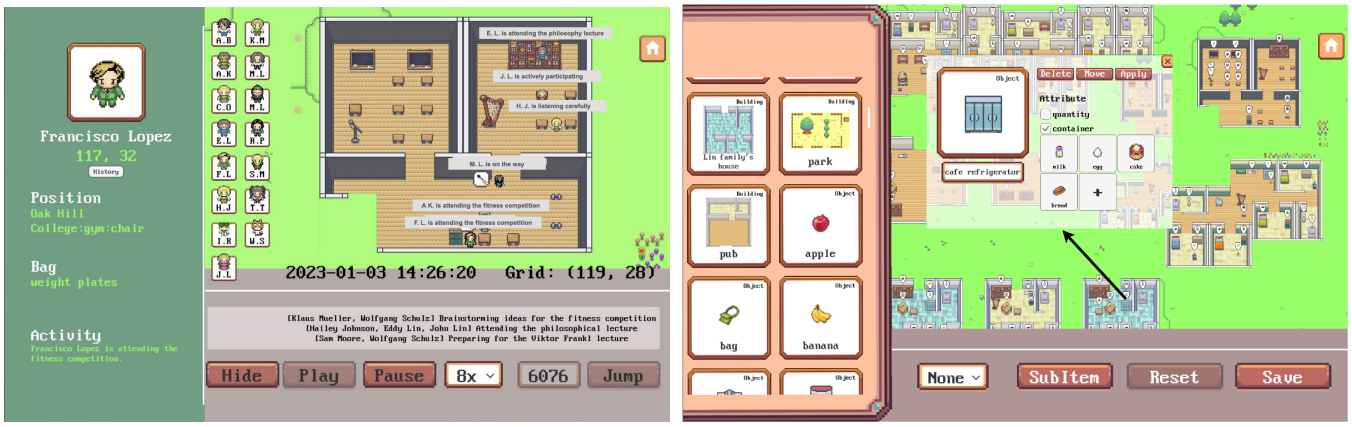

Figure 1: User interfaces of MiniAgentPro. (left) Simulation Player for observing agent interactions and activities. (right) Map Editor for creating and editing the environment.

be user-friendly, with intuitive controls for playback speed, jumping to specific steps, and inspecting agent and item details. It supports smooth animation of agent movements. By leveraging Unity, we provide a visually appealing and interactive way to observe and analyze the results, making it a useful tool for researchers, educators, and developers in the field of LLM-based social simulations.

Agent Framework

Our agent framework builds on GA (Park et al. 2023), and incorporates an on-the-fly adaptive planning mechanism and physically grounded constraints.

Activity Planning Following GA, we decompose an agent's daily activities into two hierarchical levels: high-level plans and low-level actions. In contrast to standard GA pipelines, where high-level plans are often predetermined at the beginning of the simulation, our system dynamically generates high-level plans and low-level actions on-the-fly as the simulation proceeds. By leveraging each agent's personality descriptions, history of past high-level plans, and the retrieved memories, their behaviors evolve in response to both internal states and external event requirements.

Additionally, we introduce physical constraints to enhance realism. Agents must navigate the environment with a constrained movement speed, perceiving and reacting to events along the way, which introduces temporal costs and increases complexity. Furthermore, our item interaction mechanism requires agents to collect specific items associated with each low-level action before proceeding to the designated location.

Dialogue Our dialogue system, inspired by prior work on (Wang, Chiu, and Chiu 2023), enables rich, context-aware communication that mirrors real-world interactions. Agents integrate contextual information, including personality traits, core characteristics, current daily activities, life progress statements, the activities of other agents, and event-related context, into dialogue prompts. This information determines how a dialogue is initiated and shapes its topic, ensuring conversations are task-relevant. Then, the two agents start to generate dialogue in turn until one decides to stop.

Evaluation

We develop a test set comprising eight diverse event settings: a fitness competition, a friends' dinner, a Lin's family party, a music jam session, a mixology workshop, an open mic comedy night, a philosophy lecture, and a writing workshop. These scenarios involve between 3 and 6 agents each, and are designed to evaluate the agents' abilities in event organization, coordination, and interaction within physically grounded environments. Each setting incorporates specific tasks, such as gathering required items, navigating to designated locations, and engaging in meaningful dialogues. We also design basic and hard versions: The basic version of each setting directly gives all participants the event information, while the hard version only informs the host about the event and lets the host invite all others.

Settings	RF	LA	BIR	DRC	IQ	Avg.
basic	7.04	8.69	6.71	7.87	6.91	7.44
hard	3.21	4.50	5.14	8.07	4.50	5.08

Table 1: The evaluation results of the average score from all event settings.

We test the criteria by self-designed metrics: Role Fulfillment (RF), Location Adherence (LA), Bag Item Relevance (BIR), Daily Requirement Consistency (DRC), and Interaction Quality (IQ). The score is given by GPT-4o mini in a range of 1 to 10, and the LLM used for simulation is GPT-4o. The results are shown in Table 1. The evaluation reveals strong performance in basic settings, with an overall average of 7.44. In hard variants, performance drops to 5.08, with Role Fulfillment plummeting to 3.21 due to coordination failures in invitation and participation. These findings show our framework's potential for straightforward tasks but emphasize the need for advanced mechanisms to foster emergent collaboration in complex, invitation-based scenarios.

References

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; and Wang, B. 2025. Towards Medical Complex Reasoning with LLMs through Medical Verifiable Problems. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 14552–14573. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Christiansen, F. R.; Hollensberg, L. N.; Jensen, N. B.; Julsgaard, K.; Jespersen, K. N.; and Nikolov, I. 2024. Exploring Presence in Interactions with LLM-Driven NPCs: A Comparative Study of Speech Recognition and Dialogue Options. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*, VRST '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705359.

Cox, S. R.; and Ooi, W. T. 2023. Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback. In *Chatbot Research and Design: 7th International Workshop, CONVERSATIONS 2023, Oslo, Norway, November 22–23, 2023, Revised Selected Papers*, 167–184. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-54974-8.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; Ran, C.; Xiao, L.; Wu, C.; and Schmidhuber, J. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*.

Hou, A. B.; Du, H.; Wang, Y.; Zhang, J.; Wang, Z.; Liang, P. P.; Khashabi, D.; Gardner, L.; and He, T. 2025. Can A Society of Generative Agents Simulate Human Behavior and Inform Public Health Policy? A Case Study on Vaccine Hesitancy. arXiv:2503.09639.

Li, J.; Li, J.; Chen, J.; Li, Y.; Wang, S.; Zhou, H.; Ye, M.; and Su, Y. 2024. Evolving Agents: Interactive Simulation of Dynamic and Diverse Human Personalities. arXiv:2404.02718.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P.; Seneviratne, M.; Gamble, P.; Kelly, C.; Babiker, A.; Schärli, N.; Chowdhery, A.; Mansfield, P.; Demner-Fushman, D.; Agüera y Arcas, B.; Webster, D.; Corrado, G. S.; Matias, Y.; Chou, K.; Gottweis, J.; Tomasev, N.; Liu, Y.; Rajkomar, A.; Barral, J.; Semturs, C.; Karthikesalingam, A.; and Natarajan, V. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Wang, Z.; Chiu, Y. Y.; and Chiu, Y. C. 2023. Humanoid Agents: Platform for Simulating Human-like Generative Agents. In Feng, Y.; and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 167–176. Singapore: Association for Computational Linguistics.

Yang, Y.; Wen, Y.; Wang, J.; and Zhang, W. 2025. Agent Exchange: Shaping the Future of AI Agent Economics. arXiv:2507.03904.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*.