Benchmarking Arbitrary Natural Language Tasks in 3D Open Worlds

Sonny George
Brandeis University

Chris Sypherd
University of Edinburgh

Rocco Ahching
University of Georgia

Dylan Cashman
Brandeis University

Abstract

3D-embodied autonomy toward arbitrary task outcomes is a long-standing goal in AI and Robotics. However, programmatically verifying arbitrary outcomes in open worlds is a challenge. This work proposes: (1) giving Minecraft agents the ability to capture screenshots as evidence for task completion and (2) having vision-language models (VLMs) evaluate these screenshots. We also present SemanticSteve, a high-level Minecraft skill library that includes a "take screenshot" skill. We use an expert-annotated dataset of tricky task-screenshot pairs to evaluate the capabilities of GPT-4.1 in our proposed screenshot-evaluation role and find that it is indeed fit for the task. We make both the SemanticSteve library as well as the code and data for our experiments publicly available at https://github.com/sonnygeorge/semantic-steve.

1. Introduction

While low-level control is considerably simpler in Minecraft than in the physical world, the space of high-level semantic outcomes remains similarly open-ended [1].

Nevertheless, evaluations of task completion in Minecraft have mostly focused on closed sets of tasks with easily verifiable success criteria—such as locating objects or acquiring items [1, 2, 6, 9, 12, 14–16]. Such task sets are often highly Minecraft-specific, limiting the extent to which the required decision-making reflects real-world problem-solving.

Two notable exceptions to this trend, however, include MineClip [1] and Clip4MC [6], reward models which score the semantic similarity of video snippets against open-ended language descriptions. Nevertheless, these models are not intended for assessing the *final* cumulative outcomes of long-horizon tasks.

Contributions. We propose a novel benchmarking technique that enables any describable and visible Minecraft outcome to be programmatically scored. Specifically, we propose (1) giving agents the ability to capture screenshots as evidence of task completion and (2) using vision-language models (VLMs) to score task completion based on the content of the screenshots.

Additionally, we present SemanticSteve, a novel interface and skill library for observing and controlling a player in the 3D world of Minecraft. Crucially, the SemanticSteve library includes—among other skills that unlock the bulk of the game—the *"take a screenshot"* skill, which enables capturing screenshots of arbitrary Minecraft things.

Furthermore, we conduct basic preliminary experimentation with GPT-4.1 as initial validation of VLMs in the proposed screenshot-evaluation role.

These contributions have a particular significance for the study of open-ended high-level planning over extremely long horizons since, together, they represent the first open-source, ready-to-use system that:

1. Abstracts away low-level motor control of a human-analogous 3D embodiment with semantic primitives (removing this as a variable that can confound a study focused solely on high-level planning)
2. Enables the scoring of an open-ended number of natural language tasks, *regardless* of their time horizon

2. Experiments

Our experiments aim to assess whether VLMs can evaluate screenshots as evidence of open-ended task completion in Minecraft. To do this, we compare the judgments of a state-of-the-art VLM (GPT-4.1) with those of expert annotators.

Dataset. For data, our goal was to focus on tricky evaluation scenarios requiring scene comprehension that is analogous to or otherwise meaningful for a range of useful open-language tasks. In our final curated dataset of 52 screenshot-task pairs, we included both: (1) task-outcome screenshots captured autonomously by language models (LMs) using the SemanticSteve skill library and (2) manually captured screenshots of interesting scenarios that our work-in-progress planning systems currently struggle to achieve autonomously[1].

[1]Figure 2 shows both cases, where the screenshot on the left was captured autonomously by an LM using SemanticSteve and the screenshot on the right was captured manually.

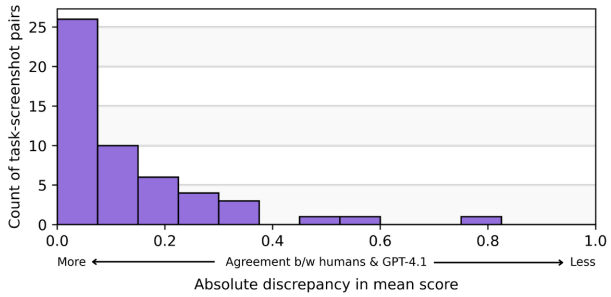

Figure 1. Histogram depicting the distribution of absolute discrepancies between the mean GPT-4.1-assigned score and the mean annotator-assigned score ($|\bar{x}_{\text{GPT-4.1}} - \bar{x}_{\text{human}}|$).

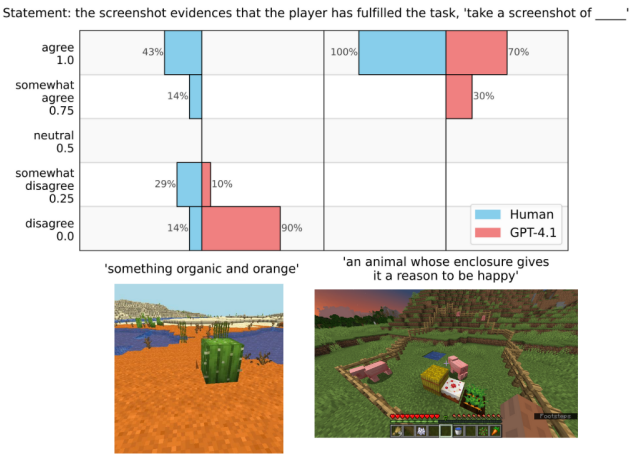

Figure 2. Stepped half-violin plots contrasting the different distributions in scores assigned by human annotators (blue) and GPT-4.1 (red) for two example task-screenshot pairs.

Annotators. To gather annotations, we recruited 10 university students who self-identified as "experienced Minecraft players." From these, only 7 passed our initial screening test of Minecraft knowledge.

Experiment Verbiage. Although more thorough prompting could help VLMs better understand the context of their evaluation role, we phrased the experiments minimalistically in order to reduce cognitive load for annotators. We phrased all tasks as *"take a screenshot of* _____(thing/outcome)*"* and asked participants to rate their agreement with the statement, *"the screenshot evidences that the player has fulfilled the task,"* using a Likert scale mapping to a score between 0 and 1 (see Figure 2).

GPT-4.1 Sampling. For each task-screenshot pair, we sampled 10 responses from `gpt-4.1-2025-04-14` using a temperature of 0.6.

Results. Overall, the results obtained were very promising. The distribution of discrepancies in mean score between GPT-4.1 and our human annotators is shown in Figure 1. GPT-4.1 gave very similar scores to the human annotators in all but a few cases. I.e., there was a strong positive correlation between GPT-4.1 and human annotator scores (Pearson's $r = 0.88$).

Manual analysis of the discrepancies revealed that there was only one case where GPT-4.1 misinterpreted the screenshot with respect to the task. All other discrepancies could be attributed to either: (1) differing interpretations of something subjective (e.g., the degree to which a 'bee' can be described as dangerous) or (2) the annotators being less semantically precise than GPT-4.1 (e.g., in the *'something organic and orange'* example in Figure 2, the annotators often felt the task was fulfilled, despite the screenshot not depicting anything that was *both* organic and orange).

3. Conclusion

We propose a novel technique for scoring an open-ended number of Minecraft outcomes and present SemanticSteve, a skill library designed to enable skill planners to achieve such outcomes. We view this as an initial step toward our planned future work to research 3D-embodied skill-planning systems that are general-purpose—that is, able to integrate with and use any lower-level skill set (open[2] [3, 4, 13] or closed [11]) to pursue open-ended outcomes.

Discussion. Open-ended high-level planning stands to be highly pertinent for the indefinite future, since, regardless of the low(er)-level controller[2], there will always be a higher-level time horizon over which lower-level actions can be sequenced.

By expanding the set of scorable Minecraft outcomes to include anything that a VLM can reliably evaluate with a screenshot, not only do we enable the scoring of more *general*—i.e., less game-specific—outcomes, our method greatly expands the number of *extremely* long-horizon tasks that can be scored. Thus, we open the door to a more thorough investigation of open-ended high-level planning over extremely long horizons.

For example, are fixed-depth hierarchical planners [5, 8, 10] sufficient for scaling up to many-hour tasks (e.g., starting with no resources, *"build a museum with displays for at least five minerals"* or *"fashion a dining arrangement and prepare a dinner party for four"*)? Or, at what point would algorithms that recursively decompose plans into arbitrarily nested subtask hierarchies[3] lead to more stable and coherent long-horizon behavior?

[2] We assert that arbitrary-language-conditioned control systems can be thought of as exposing open-ended sets of natural-language-invoked skills that can be sequenced to achieve even longer-horizon outcomes than otherwise achievable.

[3] E.g., Hierarchical Task and Motion Planning in the Now (HPN) [7], if adapted to use open-ended representations for subtask nodes (e.g., natural language expressions)

References

[1] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge, 2022. 1

[2] Yicheng Feng, Yuxuan Wang, Jiazheng Liu, Sipeng Zheng, and Zongqing Lu. Llama rider: Spurring large language models to explore the open world, 2023. 1

[3] David Ha and Jürgen Schmidhuber. World models, 2018. 2

[4] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. 2

[5] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, 2022. 2

[6] Haobin Jiang, Junpeng Yue, Hao Luo, Ziluo Ding, and Zongqing Lu. Reinforcement learning friendly vision-language model for minecraft, 2024. 1

[7] Leslie Pack Kaelbling and Tomás Lozano-Pérez. Hierarchical task and motion planning in the now. In *2011 IEEE International Conference on Robotics and Automation*, pages 1470–1477, 2011. 2

[8] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: from natural language instructions to feasible plans. *Autonomous Robots*, 47 (8):1345–1365, 2023. 2

[9] 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, 2024. 1

[10] Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. Planning with large language models via corrective re-prompting. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. 2

[11] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning, 2021. 2

[12] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. 1

[13] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making, 2024. 2

[14] Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. Skill reinforcement learning and planning for open-world long-horizon tasks, 2023. 1

[15] Zhonghan Zhao, Wenhao Chai, Xuan Wang, Li Boyi, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. See and think: Embodied agent in virtual environment, 2024.

[16] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory, 2023. 1