WorldModel Gym: A Long-Horizon Planning Benchmark for Imagination-Based Agents

Birajit Saikia[1]

[1]Student Researcher
[1]Address correspondence to: birajitsaikia@email.com

Abstract

World models; learned simulators that predict how an environment evolves under actions; are increasingly positioned as a path toward more generally capable agents, in part because they enable training and evaluation across diverse simulated curricula rather than a fixed set of tasks. However, many widely used reinforcement learning (RL) benchmarks emphasize short-horizon control, dense rewards, or limited stress-testing of planning under partial observability, where compounding model error and uncertainty become central obstacles. This paper proposes *WorldModel Gym*, an open long-horizon benchmark and evaluation platform designed to quantify progress on imagination-based agents: agents that learn dynamics and then plan in imagination using classical planners (e.g., model-predictive control and Monte Carlo tree search) or hybrid methods. The benchmark targets three gaps: sparse-reward partially observed tasks where naive model-free RL struggles; standardized measurement of the interaction between world-model fidelity, uncertainty, and planning compute; and a continual-learning track that tests adaptation under nonstationarity while monitoring catastrophic forgetting.

1 Introduction

WorldModel Gym is motivated by two developments. First, world simulation is increasingly treated as a practical route toward agents that can acquire reusable knowledge, support curriculum-based training, and enable planning beyond short reactive horizons. Second, the community has repeatedly observed that what is measured shapes what is built: well-specified benchmarks and centralized evaluation protocols can accelerate progress by standardizing comparisons, surfacing systematic failure modes, and improving reproducibility.

The benchmark is intended to push evaluation toward capabilities that plausibly matter in long-horizon decision making: delayed credit assignment, robust planning under uncertainty, memory in partially observed environments, and continual adaptation. Formally, the benchmark emphasizes settings that are better captured by partially observable Markov decision processes (POMDPs), where an agent must act under incomplete information and uncertain sensing rather than assuming access to a fully observed Markov state [1].

1.1 Motivation and Research Questions

WorldModel Gym is organized around four research questions. First, how do different planners; including tree search, sampling-based MPC, and trajectory optimization; trade off

performance against compute when planning over learned dynamics, and how sensitive are they to model bias and compounding rollout error [2]? Second, which uncertainty representations (e.g., ensembles or stochastic latent dynamics) most effectively reduce catastrophic failures induced by model error over long imagined horizons [**moerland2020mbrlsurvey**]? Third, do imagination-based agents generalize more reliably than model-free baselines under procedural distribution shift in layouts, goals, and dynamics parameters [3]? Fourth, in sequential-task settings, can online world models mitigate catastrophic forgetting, and which metrics best capture forward transfer, backward transfer, and retention [**pan2025crlsurvey**, **liu2025crlplanning**]?

1.2 Related Work

World Models and Imagination-Based Control

Modern world-model framing is often traced to work on learning compact latent representations and generative dynamics for control [4]. Subsequent model-based RL systems built latent dynamics models and performed planning in learned spaces, including PlaNet [5] and Dreamer [6]. Search-with-learned-model approaches, notably MuZero, use tree search over a learned predictive model that does not require direct access to the true environment dynamics [7]. In continuous control, latent trajectory optimization and implicit world models have become strong baselines for robustness and scalability [8]. In robotics and embodied domains, recent work combines action-conditioned visual world models with MCTS and MPC-style execution while explicitly addressing hallucination control during planning [**khorrambakht2025worldplanner**].

Planning Under Partial Observability

Long-horizon difficulty is amplified when the agent cannot observe a Markov state directly. POMDPs provide a principled framework for decision making under uncertain sensing and hidden state [1]. Online POMDP planning methods such as POMCP show how Monte Carlo tree search with particle beliefs can scale without explicit probability tables by relying on a generative simulator [9]. DESPOT and related approaches offer complementary online planning frameworks with regularization and scenario-based approximations [10]. Broader survey material on MCTS clarifies how tree-search planners behave under compute budgets and why explicit compute reporting is essential for fair comparison [2].

Benchmarks for Long-Horizon Ability and Generalization

Benchmark design for WorldModel Gym draws from procedural generalization suites and long-horizon environments. Procgen popularized procedural generation to measure both sample efficiency and generalization under distribution shift [3]. Crafter evaluates a spectrum of capabilities within a single open-world setting using achievement-style signals that require exploration and long-term reasoning [11]. The NetHack Learning Environment (NLE) provides a fast-simulation long-horizon benchmark and reinforces the importance of supporting hybrid and non-neural baselines [12].

Continual Reinforcement Learning

Continual RL addresses sequential learning under nonstationarity, where agents must adapt while avoiding catastrophic forgetting. Recent survey work synthesizes settings, metrics, and common failure modes, highlighting that evaluation must capture both transfer and retention [**pan2025crlsurvey**]. Planning with online world models provides one

promising direction for continual adaptation and motivates explicit benchmarking protocols [**liu2025crlplanning**].

2 Materials and Methods

WorldModel Gym is designed as a suite of long-horizon POMDP tasks together with an evaluation protocol that attributes performance to both (a) the learned predictive model and (b) the planner's use of the model. The benchmark is intended to support neural and non-neural baselines in a unified interface, enabling comparisons between purely learned approaches, classical online planning, and hybrid systems.

2.1 Task Principles

Tasks are constructed to isolate long-horizon planning behavior under partial observability. Rewards are intentionally sparse or delayed, requiring multi-step sequences and discouraging purely reactive heuristics. Observations are restricted (e.g., limited field-of-view, sensor noise, or latent variables such as keys, hazards, or goals), so that memory and belief tracking become necessary. To measure generalization rather than memorization, train and test splits are defined over procedural seeds and parameter ranges, aligning with established procedural benchmarking practice [3]. Finally, an optional continual-learning track introduces structured nonstationarity, such as changes in layout distributions or dynamics parameters, to quantify forgetting and transfer [**pan2025crlsurvey**].

2.2 Proposed Task Families

WorldModel Gym is proposed as a multi-tier benchmark to support rapid iteration and rigorous validation. The *Fast 2D Memory-and-Planning* track emphasizes high-throughput evaluation in partially observed environments with stochastic transitions and locked subgoals, enabling controlled comparisons to online POMDP planners such as POMCP and DESPOT [9, 10]. The *Open-World Long-Horizon* track evaluates exploration and compositional behaviors using achievement-oriented measurement, following the motivation that a single environment can reveal multiple capabilities when instrumented with semantically meaningful objectives [11]. A higher-complexity *3D Partial-Observation* track serves as a capstone for perception-driven long-horizon planning. An optional *Continual WorldModel* track introduces sequential distribution shifts and adopts standardized continual-learning reporting practices [**pan2025crlsurvey**, **liu2025crlplanning**].

2.3 Standardized Interfaces to Enable Fair Comparison

To isolate sources of improvement, WorldModel Gym proposes two standardized interfaces. The first is an environment interface (reset/step) with structured logging of episode events and achieved milestones. The second is a world-model interface that supports imagined rollouts for planner queries, enabling evaluation of how model fidelity and uncertainty affect planning outcomes. This separation follows common decompositions in model-based RL, where model learning and planning integration are analyzed jointly and via controlled ablations [**moerland2020mbrlsurvey**].

2.4 Evaluation Protocol and Baselines

Benchmark evaluation reports task performance, procedural generalization, planning cost, and diagnostic measures linking outcomes to model quality. Performance is summarized by episodic return and success rate, complemented by achievement completion in multi-capability environments [11]. Generalization is evaluated on held-out procedural seeds and parameter regimes [3]. Planning

cost is reported explicitly using wall-clock planning time per environment step, number of imagined transitions, and peak memory usage; this is essential because planners such as MCTS are anytime algorithms whose performance depends on compute budgets [2]. World-model fidelity is assessed using multi-step predictive diagnostics (e.g., reward prediction error or rollout consistency proxies) and is analyzed against performance degradation as imagined horizons increase, reflecting known concerns about compounding model errors [**moerland2020mbrlsurvey**]. For continual learning, the protocol reports forward transfer and backward transfer and quantifies forgetting as the performance drop on earlier tasks after learning new tasks [**pan2025crlsurvey**].

Baseline families include model-free RL baselines as reference points; imagination-based agents that learn latent dynamics and plan in imagination (e.g., PlaNet/Dreamer-style) [5, 6]; search-with-learned-model agents (MuZero-style) [7]; latent trajectory optimization agents in continuous control [8]; and classical online POMDP planners (POMCP/DESPOT-like) as non-neural comparators or hybrid components [9, 10]. Oracle planning controls using the true simulator can establish a ceiling and help disentangle model error from planner error [9].

2.5 Reproducibility and Analysis

To support reproducible reporting, the benchmark is designed around deterministic evaluation seeds and versioned task specifications. A reference harness records both performance and compute-normalized planning statistics. In addition, the benchmark supports controlled ablations that are commonly required for publishable evaluation: holding the planner fixed while varying uncertainty modeling; holding the model fixed while comparing planners under matched compute; varying the severity of partial observability; and varying the type and schedule of nonstationarity in the continual track.

3 Results

WorldModel Gym defines benchmark outputs along three axes: performance in sparse-reward, partially observed tasks; generalization across held-out procedural regimes; and diagnostic measures linking long-horizon success to world-model fidelity, uncertainty handling, and planning compute. The evaluation harness is designed to enable controlled comparisons in which either the planner or the world model is held fixed, allowing practitioners to attribute gains to algorithmic changes rather than uncontrolled increases in compute or implicit tuning. In continual settings, sequential shifts expose trade-offs between rapid adaptation and retention, providing a standardized way to measure forgetting dynamics.

As a benchmark proposal, the primary result is the specification of tasks, interfaces, and metrics that render these comparisons measurable and reproducible. Empirical reporting is expected to emphasize regimes where model-free baselines struggle with long-horizon credit assignment, where planning performance is sensitive to uncertainty modeling and rollout depth, and where continual adaptation exposes forgetting that can potentially be mitigated by online model updating and planning.

4 Discussion

WorldModel Gym is designed to measure long-horizon planning under partial observability, but several limitations merit explicit discussion. First, any benchmark can be over-optimized; procedural generation reduces memorization but does not eliminate the possibility of overfitting to a particular procedural distribution [3]. Second, compute incentives must be handled carefully: planners can trade compute for performance, and larger models can im-

prove results without necessarily improving algorithmic efficiency, so compute reporting and matched-budget comparisons are essential [2, 8]. Third, POMDP planning is computationally difficult in general, and practical solvers rely on approximations; benchmark difficulty should reflect meaningful long-horizon structure rather than adversarial edge cases [1, 9]. Finally, continual-learning evaluation remains an evolving area, so the continual track should report multiple complementary metrics and disclose shift schedules clearly to support robust interpretation [**pan2025crlsurvey**].

Acknowledgments

General

None.

Author Contributions

B. Saikia conceived and wrote the manuscript.

Funding

None.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Data Availability

Not applicable.

Supplementary Materials

None.

References

1. Spaan MTJ. Partially Observable Markov Decision Processes. Lecture notes / chapter. 2012. URL: https://www.st.ewi.tudelft.nl/mtjspaan/pub/Spaan12pomdp.pdf.
2. Browne CB, Powley E, Whitehouse D, et al. A Survey of Monte Carlo Tree Search Methods. IEEE Transactions on Computational Intelligence and AI in Games 2012.
3. Cobbe K, Klimov O, Hesse C, Kim T, and Schulman J. Leveraging Procedural Generation to Benchmark Reinforcement Learning. In: *International Conference on Machine Learning*. 2020. URL: https://arxiv.org/abs/1912.01588.
4. Ha D and Schmidhuber J. World Models. arXiv preprint arXiv:1803.10122 2018.
5. Hafner D, Lillicrap T, Ba J, and Norouzi M. Learning Latent Dynamics for Planning from Pixels. arXiv preprint arXiv:1811.04551 2018.
6. Hafner D, Lillicrap T, Norouzi M, and Ba J. Dream to Control: Learning Behaviors by Latent Imagination. arXiv preprint arXiv:1912.01603 2019.
7. Schrittwieser J, Antonoglou I, Hubert T, et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. arXiv preprint arXiv:1911.08265 2019.
8. Hansen N et al. TD-MPC2: Scalable, Robust World Models for Continuous Control. arXiv preprint arXiv:2310.16828 2023.
9. Silver D and Veness J. Monte-Carlo Planning in Large POMDPs. In: *Advances in Neural Information Processing Systems*. 2010. URL: https://papers.nips.cc/paper/4031-monte-carlo-planning-in-large-pomdps.

10. Somani A, Ye N, Hsu D, and Lee WS. DESPOT: Online POMDP Planning with Regularization. arXiv preprint arXiv:1609.03250 2016.

11. Hafner D et al. Benchmarking the Spectrum of Agent Capabilities. arXiv preprint arXiv:2109.06780 2021.

12. Küttler H, Nardelli N, Miller A, et al. The NetHack Learning Environment. arXiv preprint arXiv:2006.13760 2020.