ChatGPT

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

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. [1] 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. [2]

This paper proposes **WorldModel Gym**, an open long-horizon benchmark and evaluation platform designed to quantify progress on “imagination agents”: agents that **learn dynamics** and then **plan in imagination** using classical planners (e.g., MPC/MCTS) or hybrid methods. [3] The benchmark targets three gaps: (i) sparse-reward, partially observed tasks where naive model-free RL struggles; (ii) standardized measurement of the *interaction* between world-model fidelity, uncertainty, and planning compute; and (iii) a continual-learning track that tests adaptation under nonstationarity while monitoring catastrophic forgetting. [4]

Motivation and research questions

Two core trends motivate WorldModel Gym. First, frontier research labs increasingly treat *world simulation* as a stepping stone to more general agents—explicitly framing world models as enabling richer curricula of interactive environments and supporting broader generality than single-task agents. [1] Second, the research community has repeatedly learned that *what we measure shapes what we build*: benchmarks and competitions (e.g., centralized evaluation setups) can accelerate progress by standardizing protocols, surfacing failure modes, and enabling reproducible comparisons across approaches and compute budgets. [5]

WorldModel Gym is intended to push evaluation toward capabilities that plausibly matter in “2030-style” agents: long-horizon credit assignment, robust planning under uncertainty, memory in partially observed worlds, and continual adaptation. Formally, the benchmark emphasizes settings better captured by **partially observable Markov decision processes (POMDPs)**, where the agent must act with incomplete information and uncertainty in sensing, rather than assuming full state access. [6]

The project is guided by the following research questions: - **Planning–model interaction:** How do different planners (tree search, sampling-based MPC, trajectory optimization) trade off performance vs. compute when planning over learned dynamics, and how sensitive are they to model bias/compounding error? [7]
- **Uncertainty and robustness:** Which uncertainty representations (ensembles, stochastic latent states, epistemic uncertainty proxies) most effectively reduce catastrophic failures caused by model error over long

imagined rollouts? [8]
- **Generalization across procedurally generated tasks:** Do imagination agents generalize better than model-free baselines under distribution shift in layouts, goals, and dynamics parameters? [9]
- **Continual reinforcement learning:** Can planning with online world models mitigate catastrophic forgetting in sequential-task settings, and how should benchmarks quantify forward/backward transfer? [10]

Related work

World models and imagination-based control

Modern "world model" framing is often traced to work by David Ha [11] and Jürgen Schmidhuber [12], who studied learning generative models of environments and training policies using compact learned representations. [13] Subsequent model-based RL systems built latent dynamics models and performed planning in that learned space, including PlaNet (latent dynamics + online planning) and Dreamer (latent imagination + value gradients), both associated with Danijar Hafner [14]. [15] Separately, MuZero introduced a planning approach that uses a learned model sufficient for predicting reward/value/policy and performs tree search without access to true environment dynamics, led by Julian Schrittwieser [16]. [17] More recent work (e.g., TD-MPC2) studies scaling implicit world models and latent trajectory optimization across many continuous-control tasks. [18]

In robotics and embodied domains, recent approaches explicitly combine **visual world models** with **MCTS and MPC-style controllers** to plan long action sequences, highlighting the ongoing convergence of learned dynamics with classical planning/control. [19]

Planning under partial observability

Much of long-horizon difficulty emerges when the agent does not observe a Markov state directly. POMDPs provide a principled formalism for decision making with uncertain sensing and hidden state, and they motivate planners that reason over histories or belief states. [6] For online planning in large POMDPs, POMCP (introduced by David Silver [20] and colleagues) demonstrated how Monte Carlo tree search with particle beliefs can scale without explicit probability tables, relying only on a simulator. [21] DESPOT and related methods provide alternative online POMDP planning frameworks with theoretical guarantees under sampling assumptions. [22]

Benchmarks for long-horizon ability and generalization

Several environments motivate design choices for WorldModel Gym. DeepMind Lab is a 3D platform explicitly intended for research on agents in large, partially observed, visually diverse worlds. [23] Procgen popularized procedural generation as a way to measure both sample efficiency and generalization, and it inspired competition-style standardized evaluation infrastructure. [9] Crafter (introduced in work published at ICLR [24]) evaluates a spectrum of abilities within a single open-world environment and emphasizes achievements that require exploration and long-term reasoning. [25]

Finally, the NetHack [26] Learning Environment (NLE) is a notable long-horizon benchmark with fast simulation; reporting on the NeurIPS [27] 2021 challenge emphasized that symbolic and hybrid systems can outperform deep RL baselines in this domain, reinforcing the value of benchmarks that remain open to non-neural methods. [28]

Continual reinforcement learning

Continual RL addresses the real-world constraint that tasks and dynamics shift over time, and that agents must learn sequentially without catastrophic forgetting. Recent surveys synthesize settings, metrics, and methods (including world-model-based approaches) and highlight nonstationarity as a core unsolved challenge. [29] Empirical work in robotics/lifelong RL similarly discusses practical strategies for preserving and combining knowledge over long horizons. [30]

Benchmark design

WorldModel Gym is designed as a **suite of long-horizon POMDP tasks** plus an **evaluation protocol** that attributes success/failure to both (a) the learned model and (b) the planner's use of it. The benchmark intentionally supports **non-neural baselines** (e.g., online POMDP planners) alongside neural world models by standardizing interfaces and providing simulator access where appropriate. [31]

Task principles

Each task family is constructed to stress one or more of the following, all of which are repeatedly highlighted as hard in the literature on world models and partial observability:

- **Sparse reward and delayed credit assignment:** rewards are infrequent and require multi-step sequences; this increases the value of planning and imagination while surfacing compounding model error. [32]
- **Partial observability requiring memory/belief tracking:** hidden variables (goal location, keys, hazards, NPC intent) cannot be inferred from a single observation, pushing agents toward belief-state reasoning. [33]
- **Procedural generalization:** train/test splits are defined over environment seeds and parameter ranges to prevent memorization and to measure generalization, aligning with procedural benchmarking practice. [9]
- **Nonstationarity for continual RL:** sequential task regimes alter dynamics parameters, map distributions, sensor noise, or reward structures to quantify forgetting and transfer. [34]

Proposed task families

WorldModel Gym is proposed as a multi-tier benchmark so researchers can iterate quickly on fast environments while still validating claims in richer visual domains:

Fast 2D “Memory & Planning” worlds (high-throughput). Grid-based or top-down continuous variants with partial observability (limited field-of-view), stochastic transitions, and locked subgoals. These tasks encourage comparisons to online POMDP planners (POMCP/DESPOT-like) and allow stress tests of belief tracking vs. recurrent-memory policies. [35]

Open-world long-horizon survival/crafting tasks (mid-throughput). Inspired by the observation that a single environment can evaluate many capabilities if instrumented by semantically meaningful achievements, similar to Crafter's "achievement unlock" framing. [25] The WorldModel Gym variant would explicitly parameterize difficulty (resource scarcity, hazard stochasticity, partial observability) to expose planner sensitivity to model bias and uncertainty. [36]

3D partial-observation navigation and interaction (lower-throughput, higher realism). Environments in the spirit of DeepMind Lab—first-person visual observations with larger state spaces—are included as "capstone" tasks to validate that improvements transfer beyond 2D grids. [23]

Optional "Continual WorldModel" track. Task distributions shift over time (layout families, dynamics regimes, sensor noise, or goal semantics). The protocol reports forward transfer, backward transfer, and forgetting measures aligned with continual RL survey recommendations and recent continual-planning directions. [37]

Standardized interfaces to enable fair comparison

To help isolate where performance comes from, WorldModel Gym proposes two standard APIs:

- **Environment API:** a Gym-style step/reset loop plus standardized logging of episode events and achieved subgoals (for achievement-style scoring). This mirrors how benchmarks like NLE and procedural suites define task interfaces and evaluation hooks. [38]
- **World-model API:** a callable generative model interface used by planners for imagined rollouts. This is exactly the interface boundary emphasized by model-based RL surveys that separate "model learning" from "planning/learning integration." [39]

Evaluation protocol and baselines

Benchmark papers succeed when they (i) define a protocol that prevents "benchmark hacking" and (ii) make it easy to reproduce results. WorldModel Gym's evaluation is therefore defined along three axes: **task performance**, **generalization**, and **planning/model diagnostics**.

Primary metrics

Task performance. Report episodic return and success rate, but also achievement completion (for multi-capability tasks), following the motivation that achievements can represent semantically meaningful milestones within an episode. [25]

Sample efficiency. Report learning curves by environment steps, consistent with procedural and competition benchmarking practice. [9]

Generalization. Split evaluation across held-out procedural seeds/configurations; Procgen-style design explicitly targets measuring generalization and sample efficiency. [9]

Planning cost. Because planners can trade compute for performance, report: (a) wall-clock planning time per environment step, (b) number of imagined rollouts/simulated transitions, and (c) peak memory. This reflects longstanding observations that MCTS is an anytime algorithm where additional compute can improve decision quality. [40]

World-model fidelity. Track k-step prediction metrics (e.g., reward prediction error, latent rollout divergence proxies, or likelihood where available) and relate them to performance degradation under longer planning horizons, a known concern in model-based RL due to compounding errors and generalization gaps between learned and true dynamics. [41]

Continual learning outcomes. If the continual track is used, report forgetting/transfer metrics as adopted in continual learning/RL surveys, including performance drop on prior tasks after training on new tasks. [29]

Baseline families

WorldModel Gym's baseline suite is proposed to include both neural and non-neural families to encourage hybrid approaches and to avoid overfitting the benchmark to a single paradigm:

- **Model-free RL baselines** as a reference point for "naive RL" on sparse, partially observed tasks, motivated by the fact that many RL methods assume Markov state access and can struggle without memory/belief structure. [6]
- **Latent world-model agents** that explicitly plan in learned latent dynamics (PlaNet/Dreamer-style) as canonical imagination agents. [15]
- **Search-with-learned-model agents** (MuZero-style), leveraging MCTS over a learned predictive model. [42]
- **Latent trajectory optimization agents** (TD-MPC2-style) as strong baselines in continuous control with implicit world models. [18]
- **Classical online POMDP planners** (POMCP/DESPOT-like) as non-neural comparators and as oracle components when combined with learned observation models. [43]
- **Oracle planning controls**: planners given true simulator access establish ceiling performance and help disentangle "model error" from "planner error." This is aligned with the simulator-based emphasis in POMCP-style approaches. [21]

Benchmark ablations designed to be publishable

To make the paper evaluative (not merely descriptive), WorldModel Gym proposes pre-registered ablations:

- Fix the planner (e.g., MCTS) and vary world-model uncertainty modeling; measure performance as a function of planning horizon and model rollout depth. [44]
- Fix the world model and compare planners (sampling MPC vs. tree search vs. trajectory optimization) under identical compute budgets. [45]
- Vary partial observability intensity (field-of-view radius, sensor noise) and measure whether improvements come from better belief tracking or better imagination. [33]
- In the continual track, vary shift type (layout vs. dynamics parameters) and quantify forgetting and transfer. [29]

Platform design and reproducibility

WorldModel Gym is intended to be more than a set of environments: it is proposed as a **benchmarking platform** with an evaluation harness and researcher-facing tooling. This is motivated by evidence that centralized benchmark infrastructures can scale end-to-end evaluation across many submitted codebases, enabling fair comparisons and accelerating iteration (as seen in Procgen-style competition infrastructure). [46]

Web-based evaluation and analysis tools

The platform proposal includes: - **A task designer** for parameterized environment generation (procedural seeds, difficulty knobs, nonstationarity schedules), aligned with procedural benchmarking motivations. [9]
- **A leaderboard** with standardized submission format (containerized agents), similar in spirit to competition and challenge designs for RL environments. [47]
- **Agent trace visualization** showing (i) real trajectory; (ii) imagined rollouts from the world model; and (iii) planner decision traces (tree expansions, sampled trajectories). These diagnostics are directly motivated by the fact that planning algorithms such as MCTS construct an explicit search tree and derive actions from exploratory simulation. [40]

Reproducibility commitments

To be “paper-grade,” the benchmark should ship with: - Deterministic evaluation seeds and versioned task specifications, matching best practices implied by standardized benchmark designs. [48]
- A reference evaluation harness that measures both agent performance and planning compute (time, simulated transitions), consistent with the need to compare anytime planners under controlled budgets. [49]
- Clearly separated tracks (fast 2D, open-world, 3D, continual) so researchers can report results without hiding failures in expensive regimes. [50]

Limitations, risks, and ethics

WorldModel Gym’s design inherits several known limitations and risks in model-based RL and benchmarking.

Benchmark validity and overfitting. Any fixed benchmark can be over-optimized. Procedural generation is proposed partly to mitigate memorization and improve generalization measurement, but procedural distributions themselves can become a new target of overfitting if not carefully diversified. [9]

Compute incentives. Planners like MCTS can trade compute for performance, and larger world models can improve results, as suggested by scaling studies in model-based agents. [51] Without explicit compute reporting and budgets, a leaderboard could unintentionally reward brute force rather than algorithmic insight.

Partial observability complexity. POMDP planning is computationally hard in general, and practical solvers often rely on sampling or approximate belief representations. [35] The benchmark must therefore remain careful that “difficulty” comes from meaningful long-horizon structure rather than pathological edge cases.

Nonstationarity and forgetting. Continual RL is still an area with unsettled best practices and many interacting definitions of tasks and shifts. The continual track should disclose shift schedules and report multiple metrics (forgetting + transfer), consistent with survey recommendations, to avoid misleading conclusions. [29]

Dual-use considerations. More capable planning under uncertainty can be beneficial (robotics, control, simulation-based training), but capability advances can also be dual-use. A responsible benchmark should avoid directly packaging tasks that facilitate harmful real-world planning and should document intended use and limitations, mirroring the broader emphasis on responsible development in world-model deployments. [52]

References

The paper's core positioning is grounded in primary sources that (i) motivate world models as a path toward more general agents, (ii) establish canonical imagination-based methods and planning approaches, and (iii) demonstrate how benchmark infrastructure and procedural generation can standardize evaluation: - World-model motivation and "AGI stepping stone" framing from Google DeepMind [53] announcements and explainers. [1]
- Canonical imagination agents and search-with-learned-model systems (PlaNet, Dreamer, MuZero) and scaling model-based agents (TD-MPC2). [54]
- POMDP foundations and online planning algorithms (POMDP chapter, POMCP, DESPOT) plus MCTS survey material. [55]
- Benchmark precedents for partial observability and generalization (DeepMind Lab, Procgen, Crafter, NLE/NetHack challenge). [56]
- Continual RL framing and metrics guidance. [34]

---

[1] Genie 3: A new frontier for world models — Google DeepMind
https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/

[2] [23] [50] [56] [1612.03801] DeepMind Lab
https://arxiv.org/abs/1612.03801?utm_source=chatgpt.com

[3] [15] [32] [54] Learning Latent Dynamics for Planning from Pixels
https://arxiv.org/abs/1811.04551?utm_source=chatgpt.com

[4] [6] [20] [31] [33] [35] [55] pomdpChapterRLBook10Web.dvi
https://www.st.ewi.tudelft.nl/mtjspaan/pub/Spaan12pomdp.pdf

[5] [12] [46] [47] [48] Measuring Sample Efficiency and Generalization in Reinforcement Learning Benchmarks: NeurIPS 2020 Procgen Benchmark
https://arxiv.org/abs/2103.15332?utm_source=chatgpt.com

[7] [8] [39] Model-based Reinforcement Learning: A Survey
https://arxiv.org/abs/2006.16712?utm_source=chatgpt.com

[9] [14] [27] Leveraging Procedural Generation to Benchmark ...
https://arxiv.org/abs/1912.01588?utm_source=chatgpt.com

[10] [53] Continual Reinforcement Learning by Planning with Online ...
https://icml.cc/virtual/2025/poster/44151?utm_source=chatgpt.com

[11] [29] [34] [37] A Survey of Continual Reinforcement Learning
https://arxiv.org/html/2506.21872v1

[13] [1803.10122] World Models
https://arxiv.org/abs/1803.10122?utm_source=chatgpt.com

[16] [40] [49] [51] repository.essex.ac.uk
https://repository.essex.ac.uk/4117/1/MCTS-Survey.pdf

[17] [42] Mastering Atari, Go, Chess and Shogi by Planning with a ...
https://arxiv.org/abs/1911.08265?utm_source=chatgpt.com

[18] TD-MPC2: Scalable, Robust World Models for Continuous Control
https://arxiv.org/abs/2310.16828?utm_source=chatgpt.com

[19] [45] WorldPlanner: Monte Carlo Tree Search and MPC with Action-Conditioned Visual World Models
https://arxiv.org/abs/2511.03077?utm_source=chatgpt.com

[21] [43] Monte-Carlo Planning in Large POMDPs - NIPS
https://papers.nips.cc/paper/4031-monte-carlo-planning-in-large-pomdps?utm_source=chatgpt.com

[22] DESPOT: Online POMDP Planning with Regularization
https://arxiv.org/abs/1609.03250?utm_source=chatgpt.com

[24] [28] The NetHack Learning Environment
https://arxiv.org/abs/2006.13760?utm_source=chatgpt.com

[25] [2109.06780] Benchmarking the Spectrum of Agent Capabilities
https://arxiv.org/abs/2109.06780

[26] [38] proceedings.nips.cc
https://proceedings.nips.cc/paper/2020/file/569ff987c643b4bedf504efda8f786c2-Paper.pdf

[30] Preserving and combining knowledge in robotic lifelong ...
https://www.nature.com/articles/s42256-025-00983-2?utm_source=chatgpt.com

[36] [41] [44] Model-based Reinforcement Learning: A Survey
https://liacs.leidenuniv.nl/~plaata1/papers/model_based_rl_survey_fnt.pdf?utm_source=chatgpt.com

[52] Project Genie: AI world model now available for Ultra users in U.S.
https://blog.google/innovation-and-ai/models-and-research/google-deepmind/project-genie/