ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Reward-Free Continual Adaptation for Resilient Space Robots

Andrej Orsula[1], Miguel Olivares-Mendez[1], Carol Martinez[1]

Abstract*— Space robots operate in extreme environments where hardware degradation can critically compromise traditional control strategies. While continual reinforcement learning offers a promising mechanism for online adaptation, it inherently requires access to a reward signal during deployment. However, precise reward computation in space is often infeasible due to the lack of external tracking systems and the overall complexity of the environment. To address the challenge of unobservable rewards, we introduce a reward-free continual learning framework that leverages latent-state world models. By pre-training a model-based agent across diverse simulations, the world model learns a robust predictor of the reward structure within its latent space. Upon deployment to an environment with severe hardware degradation, we freeze the observation encoder and reward predictor to update only the transition dynamics of the world model through unsupervised rollouts. By training the policy entirely on imagined trajectories generated by this updated world model, the agent adapts to altered dynamics without receiving new rewards. We demonstrate our approach across simulated planetary traversal, orbital navigation, and precision assembly tasks subjected to severe morphological failures. *The source code is available at github.com/AndrejOrsula/space_robotics_bench.

I. Introduction

Long-duration space exploration endeavors, such as the deployment of Martian rovers, lunar habitat construction, and orbital servicing missions, heavily depend on robotic systems that can operate reliably and adapt to extreme conditions with minimal human supervision. While data-driven approaches like reinforcement learning (RL) offer a powerful paradigm for acquiring complex adaptive behaviors, a fundamental obstacle to their long-term autonomy is hardware degradation. Phenomena such as micrometeoroid impacts, severe thermal cycling, thruster failures, or accumulated actuator wear can introduce significant morphological discrepancies that catastrophically break pre-trained control policies. When an agent encounters such severe out-of-distribution changes, zero-shot transfer fails, and retraining from scratch is prohibitively expensive in terms of time, energy, and communication bandwidth.

Continual RL provides a compelling alternative by allowing an agent to learn and adapt its policy continuously during deployment. Yet, the primary barrier to applying continual RL in space robotics is its strict reliance on accurate and dense reward signals. In simulation, computing an arbitrarily complex reward function is trivial via access to the full and perfect state of the environment. In contrast, real physical robots in space cannot rely on privileged simulation states to calculate such metrics. Due to intermittent telemetry, harsh visual conditions that degrade sensor performance, and the absence of external tracking infrastructure such as GPS or motion capture, estimating rewards onboard is often practically impossible. This fundamental challenge of unobservable rewards poses a critical bottleneck to deploying adaptive systems in high-stakes extraterrestrial domains.

Fig. 1: Our continual learning framework enables reward-free adaptation to severe changes in dynamics by leveraging the latent reward landscape of a pre-trained world model.

To address this gap, we hypothesize that latent-state world models, pre-trained across diverse and highly randomized simulations, inherently encode a robust reward landscape that generalizes well beyond the training distribution. Subsequent online adaptation can be achieved by updating only the transition dynamics of the world model through unsupervised environmental rollouts while deliberately freezing the observation encoder and the reward predictor. This methodology enables the active policy to adapt to changing physical dynamics purely via synthetic imagination, effectively recovering task performance without ever receiving new external rewards from the physical environment.

II. Methodology

Our framework for reward-free continual adaptation builds upon the sample-efficient DreamerV3 architecture [1]. The workflow is divided into two distinct phases: a computationally intensive pre-training phase in simulation, followed by a resource-constrained adaptation phase upon deployment.

During the world model pre-training phase, the Recurrent State-Space Model (RSSM) jointly optimizes a deterministic sequence model, a stochastic forward dynamics model, an observation encoder, a decoder, and both reward and continuity predictors to compress high-dimensional inputs into compact latent states. Concurrently, an actor-critic policy is trained entirely within the synthesized latent trajectories generated by the RSSM. The critic is optimized using the predicted rewards from the reward head, encouraging the world model to learn a representation that captures the underlying structure of the task. To ensure these learned latent representations and reward mappings are robust, we employ extensive domain randomization by varying mass distributions, friction coefficients, sensor noise, and actuator strengths. This encourages the encoder to learn a highly

[1]University of Luxembourg

generalizable mapping into the latent space and ensures the reward head accurately reflects an invariant true objective.

During subsequent reward-free adaptation, the robot is deployed and inevitably encounters novel unmodeled degradation. As the true transition dynamics of the physical system diverge from the pre-trained simulation model, the zero-shot policy fails because the actions no longer yield the expected outcomes. To recover without observable rewards, we leverage the robust pre-trained reward predictor. By explicitly freezing the reward predictor, the encoder, and the decoder, the actor-critic continues to receive meaningful reward signals based on its updated latent representations, even as the physical morphology of the robot changes.

We update the sequence model and forward dynamics using a balanced Kullback-Leibler (KL) divergence loss between the posterior representations and the prior predictions. To prevent representation collapse, we incorporate the stop-gradient operator, $\mathrm{sg}(\cdot)$, which scales the gradients flowing into the prior and posterior differently:

$$\mathcal{L}_{dyn} \doteq \alpha \mathrm{KL}(\mathrm{sg}(q_\phi) \parallel p_\phi) + (1-\alpha)\mathrm{KL}(q_\phi \parallel \mathrm{sg}(p_\phi))$$

where $q_\phi \doteq q_\phi(z_t|h_t, x_t)$ and $p_\phi \doteq p_\phi(\hat{z}_t|h_t)$. To mitigate catastrophic forgetting of the foundational dynamics learned during pre-training, we reduce the world model's learning rate by an order of magnitude and inject small Gaussian exploration noise into the normalized action space.

III. Experiments and Results

We evaluate our framework across three simulated space robotics domains built upon NVIDIA Isaac Lab, each paired with an unmodeled morphological failure:

(1) Planetary Traversal: Navigation with a 12-DOF rover across procedurally generated uneven terrain, where a locked front-right wheel joint induces severe asymmetric drag.

(2) Orbital Navigation: 6-DOF dynamic waypoint tracking in microgravity, degraded by the complete failure of three co-located off-axis thrusters.

(3) Screwdriving Assembly: High-precision robotic manipulation task requiring the insertion of a bolt under tight tolerances, where a 15° bend in the tool flange significantly distorts the end-effector kinematics.

To accurately mimic the severe data-collection bottlenecks of physical extraterrestrial deployments, the agents are pre-trained for 20 million environment steps across highly randomized parallel workers. However, the online adaptation phase is strictly constrained to a single environment instance and a maximum 60-minute interaction window. We track normalized task progress metrics and evaluate four agents: a zero-shot baseline, an agent retrained from scratch to serve as an asymptotic upper bound, an adaptive agent with privileged real-time reward access, and our proposed reward-free adaptive agent.

The experimental results validate our central hypothesis. The zero-shot policies fail catastrophically across all three domains due to the unmodeled shifts in transition dynamics. While the retrained-from-scratch baseline confirms that the tasks remain physically solvable under degradation, it highlights the extreme sample inefficiency of standard RL by requiring millions of steps to recover. The agent with privileged reward access demonstrates rapid and stable recovery.

Crucially, our reward-free agent successfully utilizes the encoded latent reward landscape to achieve rapid initial policy recovery, proving that world models can guide short-term adaptation in the complete absence of external rewards. In the early stages of the adaptation window, the reward-free agent achieves performance levels analogous to the privileged baseline. However, learning profiles reveal distinct limitations over extended periods. Our agent exhibits significant volatility and late-stage performance decay, particularly in the highly dynamic orbital and high-precision assembly tasks. This indicates that continuously updating the core transition dynamics on physically degraded morphologies causes the RSSM's internal representations to gradually drift away from the original latent space. As the representations shift, the frozen reward predictor's accuracy inherently degrades, leading to suboptimal policy updates.

IV. Conclusion

This work demonstrates that latent-state world models pre-trained across diverse simulated environments encode a robust reward landscape capable of facilitating reward-free continual adaptation. By isolating dynamics updates from reward prediction, we directly address the critical bottleneck of unobservable rewards in space robotics, enabling agents to recover from severe hardware degradation using only unsupervised rollouts and synthetic imagination.

While the framework yields short-term recovery, the observed late-stage representation drift highlights clear avenues for future research. Future work will investigate the integration of localized latent-space residual adapters to strictly bound the transition dynamics updates. This would allow the agent to learn a corrective delta for the novel physical constraints without overwriting or drifting away from the original world model. Furthermore, we aim to extend this framework to high-dimensional visuomotor sim-to-real deployments on representative physical platforms, paving the way for stable and long-term autonomy in the harsh environments of space exploration.

References

[1] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering Diverse Control Tasks through World Models," *Nature*, vol. 640, 2025.