ARROW: Towards Memory-Efficient Continual Reinforcement Learning with World Models

Abdulaziz Alyahya*, Abdallah Al Siyabi†, Markus R. Ernst‡, Luke Yang †, Levin Kuhlmann † and Gideon Kowadlo§
* IMSIU † Monash University ‡ UNSW Sydney § Cerenaut

***Abstract*—We present ARROW (Augmented Replay for RObust World models), a bio-inspired, memory-efficient, continual RL algorithm that extends DreamerV3 with a distribution-matching replay buffer. On tasks without shared structure, ARROW reduces catastrophic forgetting relative to DreamerV3 while retaining competitive forward transfer; on tasks with shared structure, ARROW achieves near-zero maximum forgetting and improved stability-plasticity trade-off across almost all settings tested. The corresponding full paper is published in Transactions on Machine Learning Research (TMLR) and can be accessed through Open Review.**

I. Introduction

Many real-world robotics settings require adaptation across sequential tasks, without catastrophically forgetting old ones. One potential formalism to study this problem is *continual reinforcement learning* (CRL). Here, the dominant approach to continually acquire new skills is model-free RL with large replay buffers, which face severe scalability challenges as the number of tasks grows.

World models offer a natural alternative that support off-policy learning and can be directly linked to neuroscience [1]. Yet, existing world model approaches to CRL (e.g., Continual-Dreamer [2]) still rely on buffers with millions of high-dimensional samples. The central question ARROW addresses is: *can strategic, memory-efficient replay to a world model enable robust continual learning without task identifiers?*

II. Method

ARROW extends DreamerV3 [3] with an augmented replay buffer consisting of two complementary stores:

- A short-term FIFO buffer ($\mathcal{D}_1$) retains recent experience, preserving plasticity on the current task.
- A long-term distribution-matching buffer ($\mathcal{D}_2$) preserves a representative sample of past experience via reservoir sampling, mitigating recency bias and forgetting.

Both buffers are sampled uniformly in each mini-batch. The design is directly motivated by Complementary Learning Systems (CLS) theory [1], where the hippocampus (fast, episodic) replays to the neocortex (slow, statistical). Accordingly, ARROW replays to the predictive world model rather than directly to the policy.

We evaluate ARROW on two continual RL regimes across three task orders (default, reversed, two-cycle):

- Atari (6 games): tasks without shared visual or reward structure.
- Procgen CoinRun (6 variants): tasks with shared dynamics and visual features, enabling knowledge transfer.

As baselines we chose DreamerV3 (same architecture, standard FIFO buffer) and TES-SAC (strongest model-free agent at the same memory budget), both matched at 2^{19} observations in their respective buffers. We report average performance (ACC), average minimum accuracy (min-ACC), worst-case accuracy (WC-ACC), forward transfer, maximum forgetting (Max-F), and recovery.

III. Key Results

Atari (no shared structure). ARROW nearly eliminates forgetting compared to DreamerV3 regardless of task order. In the two-cycle setting ARROW is the only method to achieve a positive WC-ACC, and its Max-F is an order of magnitude below DreamerV3.

CoinRun (shared structure). Forgetting is lower across all methods when tasks share representations. ARROW achieves near-zero Max-F in two-cycle training and attains the highest WC-ACC in every configuration. All methods achieve recovery >1.0 (revisiting tasks is beneficial). ARROW shows the strongest overall stability-plasticity balance.

IV. Relevance to the Workshop

ARROW directly addresses some of the workshop's core questions: (i) *when learning inside a world model works in practice*: our results show that strategic replay to the world model suffices for robust continual behavior at modest memory cost, (ii) *how world models can generalize across tasks and environments*: distribution-matching replay preserves task diversity in the world model's training distribution, enabling graceful generalization to previously seen tasks, and (iii) *how world models should be evaluated beyond prediction accuracy*: the CRL formalism and revisiting past tasks offer compelling metrics like *recovery* that world models can be tested against. We see ARROW as a step toward lifelong robot world models that continuously acquire and refine skills without forgetting.

References

[1] Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).

[2] Kessler, S. *et al.* The effectiveness of world models for continual reinforcement learning. In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, vol. 232 (2023).

[3] Hafner, D., Pasukonis, J., Ba, J. & Lillicrap, T. Mastering diverse control tasks through world models. *Nature* **640**, 647–653 (2025).