MONASH University

ARROW

Augmented Replay for RObust World models*

*Abdulaziz Alyahya, Abdallah Al Siyabi, Markus R. Ernst, Luke Yang, Levin Kuhlmann, and Gideon Kowadlo. ARROW: Augmented Replay for Robust World Models. Transactions on Machine Learning Research, 2026. ISSN 2835-8856.

* Accepted as a poster at NeurIPS 2026.

Introduction

01 STATIONARY RL

Reinforcement learning has achieved strong results in single, stationary tasks (Sutton & Barto, 1998; Mnih et al., 2015; Silver et al., 2017).

02 REAL-WORLD ADAPTATION

Real agents need to adapt across sequentially changing tasks, often without task labels (Khetarpal et al., 2022).

03 CONTINUAL RL

This creates continual reinforcement learning - Main challenge: balancing stability, and plasticity (Parisi et al., 2019).

04 THE BOTTLENECK

Catastrophic forgetting remains a core obstacle (McCloskey & Cohen, 1989; French, 1999).

Problem Overview - Catastrophic forgetting

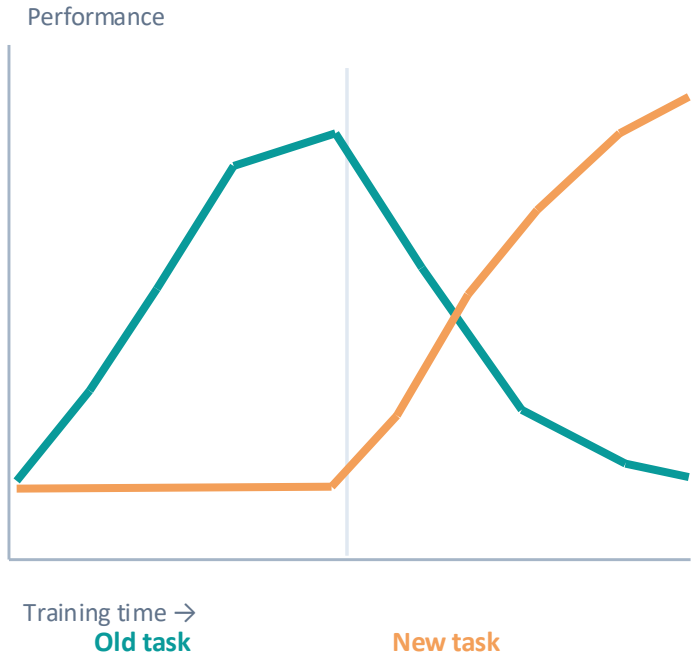
Intelligence requires learning new skills without losing old ones.

Neural networks suffer catastrophic forgetting: new tasks overwrite old capabilities (McCloskey & Cohen, 1989; French, 1999)

We want agents that learn over a sequence of tasks — often with no task labels.

This is Continual Reinforcement Learning (CRL).

BEFORE / AFTER TASK SWITCH



Problem Overview - Stability–plasticity dilemma

S

STABILITY

Stability — retain performance on past tasks

P

PLASTICITY

Plasticity — learn new tasks quickly

T

TRANSFER

Transfer — reuse old knowledge to help new tasks (and vice versa)

> These goals are in tension the stability–plasticity dilemma (Parisi et al., 2019; Mermillod et al., 2013)

> RL is harder than supervised CL: data is streamed, non-stationary, can't just reshuffle a dataset (French, 1999).

Bio-inspired replay

Complementary Learning Systems (CLS) theory

The brain has two memory systems:

FAST MEMORY

Fast system (hippocampus) — captures recent episodes

SLOW MEMORY

Slow system (neocortex) — builds structured, long-term knowledge

EXPERIENCE REPLAY →

Hippocampus replays experiences to the neocortex which reduces forgetting(Hassabis et al., 2017; Khetarpal et al., 2022).

Key insight: the neocortex behaves like a predictive World Model (Mathis, 2023).

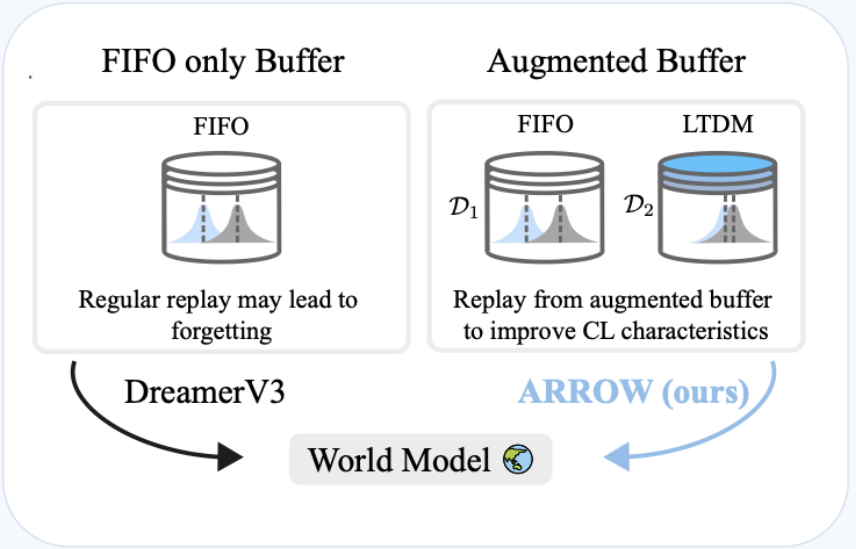
In RL, replay usually trains a policy — we instead replay to a World Model

Our contribution

ARROW = DreamerV3 (Hafner et al., 2025) + a dual-buffer replay mechanism

Two complementary buffers under a fixed memory budget:

- Short-term FIFO — recent experience (the "hippocampus")
- Long-term distribution-matching — preserves task diversity (the "neocortex")



Contribution is the working recipe, not a new sampling trick:

operationalizing the CLS split inside a state-of-the-art world model is enough to turn a recency-biased agent into a strong continual learner

The Augmented Replay Buffer

Total budget: 2^19 ≈ 524k observations.

SHORT-TERM / FIFO

Short-term FIFO buffer (~262K): all incoming experience, recency bias: fast learning of current task.

LONG-TERM / LTDM

Long-term distribution-matching (LTDM) buffer (~262K), inspired by selective replay (Isele & Cosgun, 2018):

Reservoir sampling (random priority keys) keeps a uniform random subset across all tasks.

preserves task diversity: fights forgetting.

Both sampled uniformly for each minibatch.

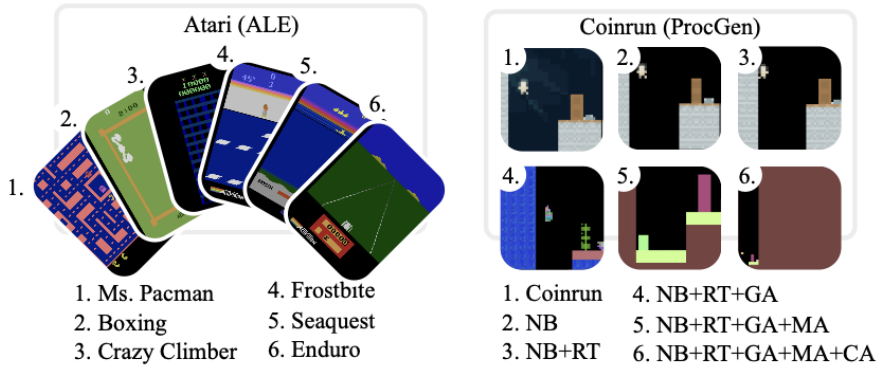
Spliced rollouts: store fixed-length chunks (512 steps), not whole episodes: more unique trajectories, better coverage.

Experimental Setup

Two task suites:

Atari / ALE (6 games) — no shared structure (Bellemare et al., 2013; Machado et al., 2018)

Procgen CoinRun (6 variants) — shared structure (Cobbe et al., 2020)



Three training settings: Default order · Reversed order · Two-cycle (revisit tasks)

Baselines, all at the same memory budget:

DreamerV3 — model-based, same architecture (Hafner et al., 2025)

TES-SAC — model-free SAC variant with target-entropy scheduling (Xu et al., 2021)

5 seeds; report median with interquartile ranges

How we measure success

Forgetting

Forgetting — how much old-task performance is lost by end of training (Kessler et al., 2023)

FORWARD TRANSFER

Forward transfer — does old knowledge speed up new learning? (Kessler et al., 2023)

STABILITY × PLASTICITY

Stability–plasticity metrics: ACC, min-ACC, WC-ACC (De Lange et al., 2023)
WC-ACC is the key summary: balances current-task learning and worst retained performance

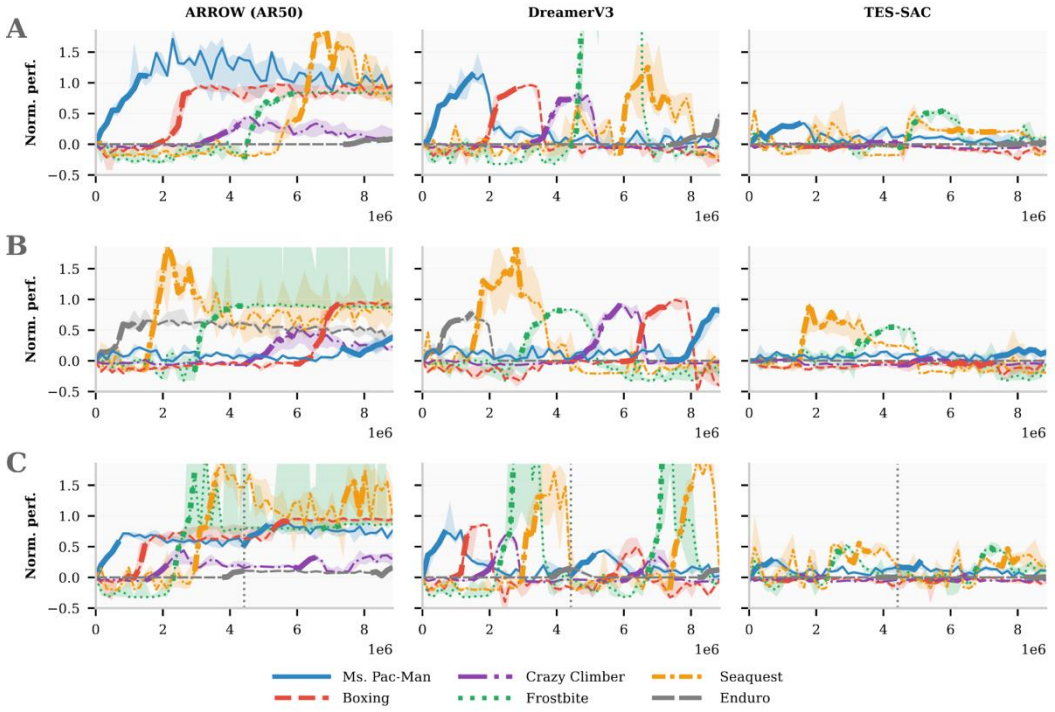
TWO-CYCLE Setting

Two-cycle metrics: Max-F (max forgetting between exposures) & Recovery

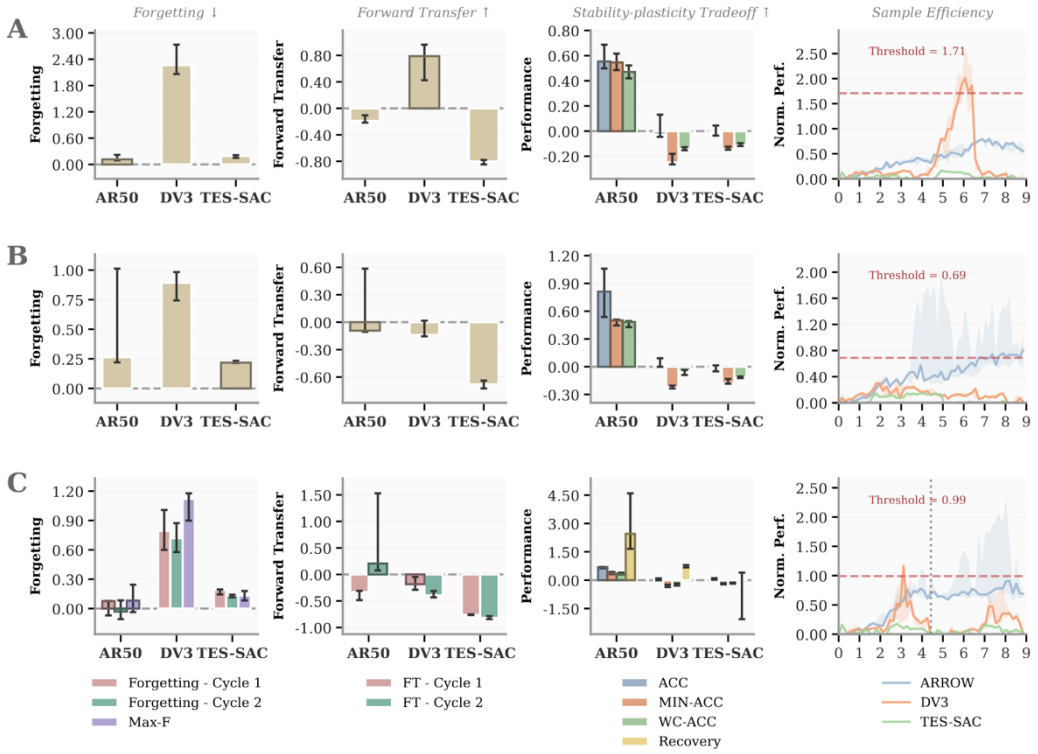
WC-ACC IS THE KEY SUMMARY

Results: Atari (no shared structure)

Performance

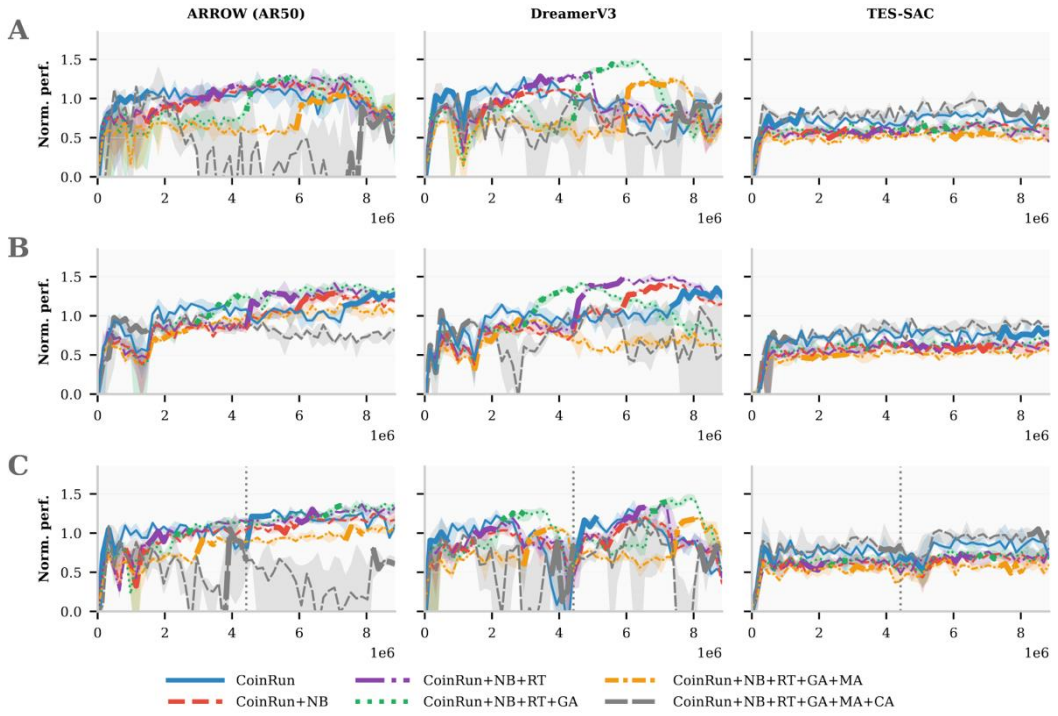


SUMMARY METRICS

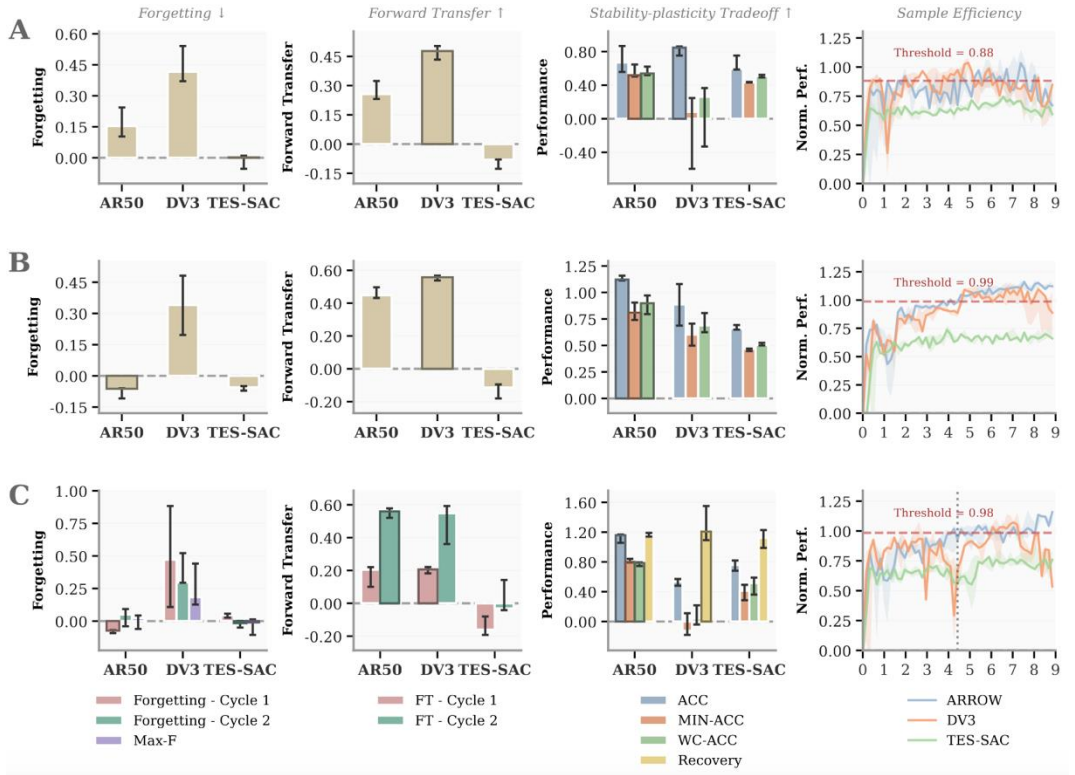


Results: CoinRun (shared structure)

Performance

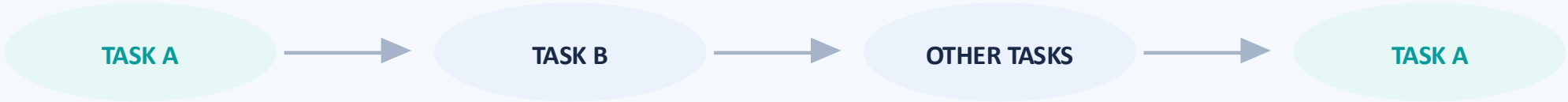


SUMMARY METRICS



Results: Two-cycle Training

Two-cycle = revisit tasks -> closest to real deployment



ARROW's maximum forgetting (Max-F) is ~an order of magnitude below DreamerV3

Atari: small Max-F; CoinRun: essentially zero

ARROW is the only method with positive WC-ACC in the two-cycle Atari setting

Strong Recovery — ARROW recovers and even surpasses prior performance when tasks return

We view this setting as the most practically informative of the three

Discussion & limitations

ARROW = strong foundation for continual RL: model-based + bio-inspired replay

Learns tasks with wildly different reward scales, no task IDs — hard on its own

Complementary to existing methods: EWC, P&C, CLEAR (Kirkpatrick et al., 2017; Schwarz et al., 2018; Rolnick et al., 2019)

Limitations:

Fixed 50/50 split -> dynamic, task-aware allocation is future work

Any finite buffer eventually loses old samples if tasks are never revisited

Only two task orderings tested

Thank you

Questions & discussion

PAPER

Paper: https://openreview.net/forum?id=3FK2tFwNwK

CODE

Code: https://github.com/Cerenaut/ARROW