AWM: ACCUMULATIVE WORLD MODELING FOR CONTINUAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual model-based reinforcement learning requires agents to adapt to non-stationary environments while preserving predictive dynamics from earlier tasks. A fundamental challenge lies in balancing the plasticity required to capture diverse, non-stationary dynamics against the stability needed to shield historical representations from backward interference. To resolve this fundamental problem, we introduce Accumulative World Modeling (AWM), a continual world-modeling framework that separates the capacity required to acquire new tasks from that needed to preserve past dynamics, while maintaining an evolving, trainable shared world model. Specifically, Composable Refinement Mechanisms (CRM) seamlessly adapt the shared world model by composing newly acquired adapters with those from prior tasks. Return-Constrained Compression (RCC) then compresses the adapters into a compact set of atoms, preventing parameter explosion. For task-agnostic inference, Confidence-Guided Expert Routing (CGER) uses cumulative observation-reconstruction error to identify the appropriate expert without task identities. Across two public benchmarks, AWM outperforms the strongest baselines, improving normalized final average performance (ACC) by 212.3% on Atari and 15.5% on CoinRun. *Code will be publicly available upon acceptance.*

1 INTRODUCTION

World models learn to predict how environments evolve in response to an agent's actions, enabling policy learning through imagined experience (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2022; Wu et al., 2023; Hafner et al., 2025). As environments change, however, previously learned world models may no longer provide accurate predictions. Continual world modeling aims to support such adaptation while preserving knowledge of earlier environments (Kessler et al., 2023; Yang et al., 2024; Alyahya et al., 2026; Pan et al., 2026). A key challenge is catastrophic forgetting, in which learning new tasks may erode knowledge (Kessler et al., 2023).

Traditional continual learning methods mitigate forgetting through regularization, replay, or architectural expansion (Kirkpatrick et al., 2017; Zenke et al., 2017; Rolnick et al., 2019; Rusu et al., 2016). In deep visual continual world modeling, experience replay is a practical solution to tackle forgetting (Kessler et al., 2023; Yang et al., 2024; Alyahya et al., 2026; Chen et al., 2022). Although experience replay can mitigate forgetting by revisiting examples from past tasks, a memory budget limits the stored data, and updates for a new task may conflict with previous task objectives (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019). In world models, changes to shared representations and dynamics can affect both state inference from old observations and imagined rollouts for policy learning (Hafner et al., 2020). The question is therefore whether revisiting past experience reliably preserves these predictive dynamics.

To examine this question, we evaluate DreamerV3 equipped with an unbounded replay buffer across two sequential Atari tasks (Hafner et al., 2025). As shown in Figure 1, evaluating on a held-out sequence from Seaquest reveals a sharp degradation in prediction quality after learning Enduro (left). Furthermore, internal responses shift substantially across world-model components, with the largest normalized output change observed in the actor–critic (right). These results demonstrate that even with idealized, unbounded replay, a continual world model fails to prevent catastrophic forgetting.

This diagnostic points to representational interference in the shared world model as a possible source of forgetting, even with an unbounded replay buffer. Fully isolating tasks in independent models

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

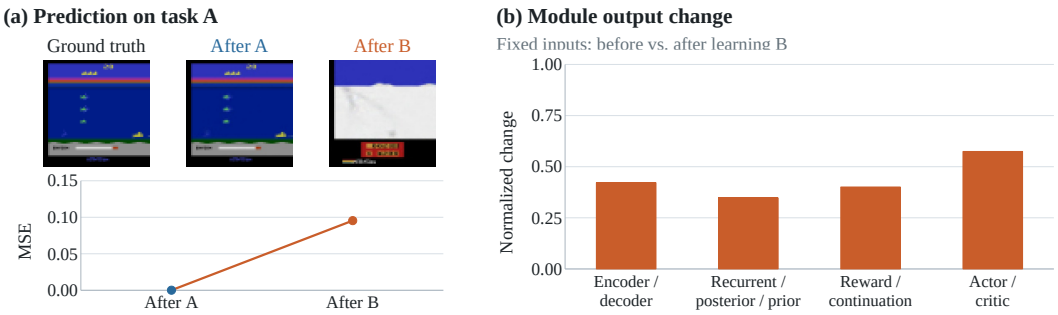

Figure 1: **Sequential-task forgetting diagnostic.** We assess task-A retention in DREAMERV3 before and after learning task B by comparing (a) predictions and MSE on a fixed task-A sequence and (b) module outputs under fixed inputs. A and B denote Seaquest and Enduro, respectively.

could reduce this interference but limit cross-task reuse (Xu et al., 2023; Georgiev et al., 2025) and increase parameter count. Continual world models must therefore decouple the accumulation of shared knowledge from the preservation of task-specific dynamics.

Guided by this principle, we propose **A**ccumulative **W**orld **M**odeling (AWM), a framework that resolves the plasticity–stability dilemma by coupling an evolving shared world model with modular, task-specific experts. Rather than freezing historical parameters or relying on monolithic shared dynamics, AWM allows the central world model to continuously learn across tasks to compound transferable environment representations. Simultaneously, each expert retains dedicated dynamics adapters alongside an isolated actor–critic pair, effectively insulating task-specific transitions and behavioral policies from catastrophic interference. At deployment, the agent autonomously identifies the relevant expert by evaluating observation consistency, achieving adaptive control across diverse environments without requiring ground-truth task identifiers.

Specifically, AWM integrates three components. Composable Refinement Mechanisms (CRM) learn full-width dynamics adapters at the recurrent, posterior, and prior interfaces and compose them with historical adapter atoms from earlier tasks. Return-Constrained Compression (RCC) prunes acquired adapter networks, retaining the candidate that satisfies the completed-task return constraint to reduce model parameter capacity. Finally, Confidence-Guided Expert Routing (CGER) accumulates expert reconstruction errors during inference, updating the selected expert and executing its actor without an extra task identity (Aljundi et al., 2017). On Atari and CoinRun, AWM improves normalized final average performance (ACC) over the strongest baseline by 212.3% and 15.5%, respectively.

The contributions of this paper are threefold:

- We propose **A**ccumulative **W**orld **M**odeling (AWM), a continual world-modeling framework that decouples shared-model adaptation from task-specific preservation. Its evolving, trainable shared world model allows shared parameters to be reused across tasks, while task-specific experts mitigate representational drift.
- We introduce Composable Refinement Mechanisms (CRM) to adapt predictive dynamics via modular adapters, coupled with Return-Constrained Compression (RCC) to prune adapters under performance guarantees, strictly bounding parameter growth.
- We propose Confidence-Guided Expert Routing (CGER), an inference mechanism that dynamically identifies and activates the corresponding expert via cumulative reconstruction errors, eliminating the need for test-time environment labels.
- We validate AWM on the Atari and CoinRun benchmarks. On Atari, it improves normalized final average performance by 212.3% relative to the strongest baseline; on CoinRun, it also shows stronger normalized retention metrics than the strongest baseline.

2 Related Work

World models. Learning predictive transition models in latent space enables agents to optimize behavioral policies through imagined rollouts (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2021; 2025; Kaiser et al., 2020; Gelada et al., 2019; Grimm et al., 2020). Subsequent efforts have enhanced latent dynamics through expressive sequence backbones, such as Transformers (Micheli et al., 2023; Robine et al., 2023; Zhang et al., 2023) and state-space models (Deng et al., 2023; Samsami et al., 2024; Sun et al., 2024), or prioritized decision-relevant representations over pixel reconstruction (Schrittwieser et al., 2020; Hansen et al., 2022; 2024; Zhang et al., 2021; Zhou et al., 2025). Despite these advances, existing world models predominantly operate under stationary, single-task assumptions. When deployed in non-stationary sequences, joint parameter updates across heterogeneous environments inevitably trigger representational drift and catastrophic forgetting. Rather than proposing an isolated base architecture, AWM introduces modular adapter banks, enabling scalable sequential retention while retaining the benefits of shared representation learning.

Continual world modeling. Experience replay constitutes the prevailing paradigm for mitigating catastrophic forgetting in continual visual reinforcement learning (Rolnick et al., 2019; Chen et al., 2022; Kessler et al., 2023; Yang et al., 2024). Recent extensions adapt replay to latent dynamics, using distribution-matching buffers or short-term retrieval to stabilize temporal transitions (Alyahya et al., 2026; Pan et al., 2026). Some approaches also combine synthetic rehearsal with exploration for task sequences sharing state spaces and transition dynamics (Fu et al., 2025). However, rehearsal alone does not guarantee dynamics preservation: joint optimization across conflicting gradients often induces representational drift, causing severe policy degradation even under idealized, unbounded memory (Lopez-Paz & Ranzato, 2017; Nijjer, 2026). Online planning with continually updated world models offers a complementary approach to adaptation across task sequences (Liu et al., 2025). Alternative strategies deploy task-conditioned hypernetworks or parameter expansion (Huang et al., 2021; Rusu et al., 2016), yet they assume shared state spaces or incur unconstrained parameter growth. In contrast to replay, AWM addresses interference at the architectural level.

Continual Learning with Network Expansion. Continual network expansion methods isolate task knowledge by allocating task-specific parameters (Rusu et al., 2016; Ruvolo & Eaton, 2013; Mendez et al., 2022; Schwarz et al., 2018; Yan et al., 2021; Wang et al., 2022a). However, growing full network columns incurs prohibitive parameter scaling over extended sequences (Rusu et al., 2016). While parameter-efficient fine-tuning (PEFT) approaches—such as low-rank adapters and prompt pools (Hu et al., 2022; Wang et al., 2022b; Liang & Li, 2024)—alleviate capacity explosion in static domains, extending modular adaptation to continual world modeling remains non-trivial due to the tightly coupled temporal recursions and stochastic latent transitions (Zhou et al., 2026). Conversely, fixed-capacity consolidation schemes ultimately suffer from representation saturation as environments diversify (Schwarz et al., 2018). AWM bridges the gap between parameter growth and rigid capacity limits. Rather than freezing subnetworks, AWM adapts latent dynamics via lightweight, composable interfaces and subsequently compresses each adaptation into a compact atom.

Task-Agnostic Inference and Expert Routing. Deploying continual agents without explicit task identifiers requires inferring environment identity directly from online interactions (Xu et al., 2020; Iten et al., 2026). A prominent paradigm relies on task-specific autoencoders or gating networks, such as Expert Gate (Aljundi et al., 2017), which evaluate reconstruction likelihoods to route inputs to the appropriate model. Complementary continual-RL approaches explore self-activating neural ensembles to specialize behavior across tasks (Powers et al., 2022). However, dedicating independent gating models to each task introduces considerable parameter overhead and isolates routing from the underlying transition dynamics (Aljundi et al., 2017). In contrast, AWM exploits the intrinsic predictive architecture of world models to achieve task-agnostic routing with zero auxiliary gating parameters. Through Confidence-Guided Expert Routing (CGER), the agent leverages cumulative observation-reconstruction errors directly from the shared decoder as an unsupervised measure of model confidence. By coupling these temporal likelihoods with task-specific recurrent experts, CGER dynamically activates the correct policy without requiring external task labels or dedicated routing networks.

3 ACCUMULATIVE WORLD MODELING

3.1 OVERVIEW

We consider a world model-based agent that learns tasks $\mathcal{M}_1, \ldots, \mathcal{M}_K$ sequentially while saving experience in a bounded replay buffer $\mathcal{B}$. Each episode belongs to one task. The goal is to let the model acquire new tasks while retaining earlier capabilities. During training, task-boundary labels are available for experience collection. At test time, however, the agent operates in a strictly task-agnostic setting, requiring it to act purely from observations.

As the foundation for our work, AWM builds on DreamerV3 and its recurrent state-space model (RSSM) (Hafner et al., 2025). Given an interaction history $\mathcal{H}_t = (o_0, a_0, \ldots, a_{t-1}, o_t)$, the world model constructs a compact latent state $s_t = (h_t, z_t)$, combining a deterministic recurrent memory h_t with a categorical stochastic latent z_t. Specifically, the recurrent memory evolves according to

$$h_t = f_\theta(h_{t-1}, z_{t-1}, a_{t-1}), \tag{1}$$

where o_t denotes the image observation, a_{t-1} is the preceding action, and e_θ is the visual encoder. The stochastic state is sampled from the posterior as $z_t \sim q_\theta(\cdot \mid h_t, e_\theta(o_t))$, while the prior provides $\widehat{z}_t \sim p_\theta(\cdot \mid h_t)$ for imagined rollouts in actor–critic learning.

As shown in Figure 2, AWM organizes continual adaptation by decoupling an evolving shared world model from task-specific experts. During training on a new environment, Composable Refinement Mechanisms (CRM) introduce task-specific dynamics adapters across the recurrent, posterior, and prior interfaces to adapt task-specific dynamics while keeping earlier experts' private parameters frozen. Upon task completion, Return-Constrained Compression (RCC) prunes the current expert's dynamics adapters. Confidence-Guided Expert Routing (CGER) compares expert reconstruction errors at scheduled checks throughout the episode to select the policy for the current environment.

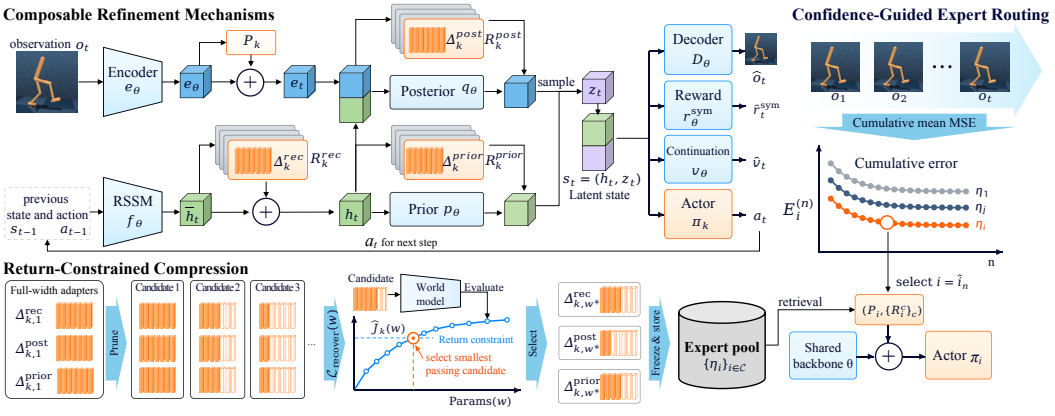

Figure 2: Overview of the AWM framework. During learning, CRM combines newly allocated adapters with frozen adapter atoms from earlier tasks. RCC prunes the current expert's adapters under a completed-task return constraint. During deployment, CGER selects the expert with the lowest cumulative mean reconstruction error without a supplied task identity.

3.2 COMPOSABLE REFINEMENT MECHANISMS (CRM)

CRM adds task-specific dynamics adapters to the recurrent state, posterior logits, and prior logits of a shared trainable world model. Each adapter combines with frozen atoms from earlier tasks to adapt these three computations to the current task.

The shared world model θ comprises the image encoder, recurrent transition, posterior and prior networks, and prediction heads. For task k, let $\bar{h}_t = f_\theta(h_{t-1}, z_{t-1}, a_{t-1})$ be the shared recurrent output and $e_t = e_\theta(o_t) + P_k(e_\theta(o_t))$ the encoding adjusted by its observation projector P_k. The adapted state and logits follow:

$$\begin{aligned} h_t &= \bar{h}_t + R_k^{\mathrm{rec}}(\bar{h}_t), \\ \ell_{k,t}^q &= \ell_\theta^q(e_t, h_t) + R_k^{\mathrm{post}}([e_t, h_t]), \\ \ell_{k,t}^p &= \ell_\theta^p(h_t) + R_k^{\mathrm{prior}}(h_t). \end{aligned} \tag{2}$$

The adjusted logits define categorical posterior and prior distributions through DreamerV3's normalization and uniform mixture. The recurrent adapter refines the state passed to both paths, the posterior adapter calibrates inference from the current observation, and the prior adapter calibrates latent transitions during imagination.

Each adapter Δ_k^c is a two-layer MLP whose hidden channels are partitioned into M groups. Each group's output contribution is an *atom*; summing these atoms gives the adapter output. For $c \in \{\mathrm{rec}, \mathrm{post}, \mathrm{prior}\}$, the composed adapter output R_k^c combines the new adapter with weighted outputs of frozen adapter atoms A_{jm}^c from earlier tasks:

$$R_k^c(u) = \Delta_k^c(u) + \sum_{j<k} \sum_{m=1}^{M} \alpha_{kjm}^c A_{jm}^c(u). \tag{3}$$

Here u is the corresponding interface input, and the signed scalar α_{kjm}^c is learned for task k and shared across its observations. These coefficients weight historical atom outputs within an expert; CGER instead selects which expert controls the agent. The adapter outputs and reuse coefficients start at zero to promote stable training.

Each task-specific expert η_k contains an observation projector P_k, three dynamics adapters and their reuse coefficients, and a private actor–critic pair (π_k, V_k):

$$\eta_k = \left(P_k, \{\Delta_k^c, \alpha_k^c\}_c, \pi_k, V_k\right), \qquad c \in \{\mathrm{rec}, \mathrm{post}, \mathrm{prior}\}. \tag{4}$$

Thus, an expert contains adapters, and each adapter consists of atoms. The actor–critic learns from trajectories imagined with the shared world model and that expert's adapters. After boundary updates and compression, η_k is frozen while the shared world model θ remains trainable.

Let $\mathcal{B}_k$ denote the current task replay buffer and $\mathcal{B}_j$ the buffer from an earlier task $j < k$. We optimize the model through two complementary objectives.

$$\begin{aligned} \mathcal{L}_{\mathrm{new}} &= \mathcal{L}_{\mathrm{WM}}(\mathcal{B}_k) + \lambda_\Delta \sum_c \mathbb{E}_{\mathcal{B}_k} \|\Delta_k^c(u)\|_2^2, \\ \mathcal{L}_{\mathrm{old}} &= \mathcal{L}_{\mathrm{WM}}(\mathcal{B}_j) + \mathcal{L}_{\mathrm{cons}}(\mathcal{B}_j), \end{aligned} \tag{5}$$

where $\mathcal{L}_{\mathrm{WM}}$ is the standard DreamerV3 objective. The regularizer penalizes new adapter outputs to encourage reuse of historical atoms. The teacher-distillation loss $\mathcal{L}_{\mathrm{cons}}$ matches old-task posterior distributions, recurrent states, actor outputs, and prediction-head outputs to a frozen boundary teacher (Li & Hoiem, 2018) as the shared world model is updated.

3.3 Return-Constrained Compression (RCC)

Retaining uncompressed dynamics adapters across sequential environments would cause continued parameter growth. We propose RCC to compress newly acquired adapters at task boundaries after full-width task adaptation, limiting retained adapter width per task before expert parameters are frozen for reuse. Total model capacity nevertheless accumulates as new experts are added.

RCC generates each candidate by structurally pruning the uncompressed adapter $\Delta_{k,1}^c$ to a discrete width fraction $w \in \mathcal{W}$, with $w = 1$ as the reference. For each atom, hidden channels q are ranked by the score $\|W_{\mathrm{in}}[:, q]\|_2 \cdot \|W_{\mathrm{out}}[q, :]\|_2$, where W_{in} and W_{out} are the incoming and outgoing weight matrices; an equal number of top-ranked channels is retained per atom.

Adapter Recovery. Each pruned candidate receives the same number of recovery updates on completed-task replay $\mathcal{B}_k$ using the losses below. We keep the task's actor–critic (π_k, V_k) and observation projector P_k fixed. Let $\Delta^c_{k,w}$ denote the candidate adapter at width fraction w, and $\Delta^c_{k,1}$ the frozen full-width teacher adapter. We optimize the following objective,

$$\begin{aligned}\mathcal{L}_{\text{recover}}(w) = \mathcal{L}^{(w)}_{\text{WM}}(\mathcal{B}_k) + \mathcal{L}^{1\to w}_{\text{cons}}(\mathcal{B}_k) \\ + \lambda_M \sum_c \mathbb{E}_{\mathcal{B}_k} \operatorname{MSE}\big(\Delta^c_{k,w}(u_{k,w}), \operatorname{sg}[\Delta^c_{k,1}(u_{k,1})]\big)\,,\end{aligned} \tag{6}$$

where $\mathcal{L}^{(w)}_{\text{WM}}$ fits real-task targets, $\mathcal{L}_{\text{cons}}$ matches the full-width teacher's state, actor, and prediction-head outputs, and the final term matches adapter outputs. During recovery, keep the shared weights and policy networks frozen, and only update the candidate adapters $\Delta^c_{k,w}$ across interfaces.

Return-Constrained Selection. Rather than relying on surrogate prediction loss, RCC selects retained adapter capacity directly by empirical task return under oracle routing. For candidate w, let $\widehat{J}_k(w)$ denote its raw return on a dedicated selection cohort for the just-completed task under its known (oracle) route. RCC selects the most compact candidate satisfying the return constraint:

$$w^* = \arg\min_{w\in\mathcal{W}} \text{Params}(w) \quad \text{s.t.} \quad \widehat{J}_k(w) \geq \widehat{J}_k(1) - \varepsilon_J \max(|\widehat{J}_k(1)|, 1), \tag{7}$$

where $\varepsilon_J = 0.05$ sets the permitted return shortfall relative to the full-width candidate. Because $w = 1$ is included in $\mathcal{W}$, the full-width candidate satisfies the constraint and provides a fallback. The selected adapters are installed as $\Delta^c_k \leftarrow \Delta^c_{k,w^*}$ in expert η_k and then frozen. Thus, RCC retains the smallest evaluated adapter configuration meeting the completed-task return criterion, with width selected separately for each task.

3.4 Confidence-Guided Expert Routing (CGER)

CGER uses reconstruction error as a proxy for how well each expert matches current observations. It compares errors across eligible experts using the shared decoder, without a separate task classifier.

Let $\mathcal{C}$ denote the pool of experts eligible for routing. During deployment, all candidates update their recurrent states using the actions executed by the agent. Specifically, each candidate $i \in \mathcal{C}$ maintains a deterministic recurrent state h^i_t and posterior belief ρ^i_t, from which the shared decoder reconstructs observations as $\widehat{o}^i_t = D_\theta([\text{vec}(\rho^i_t), h^i_t])$. The decoder uses continuous posterior probabilities for reconstruction scoring, reducing sampling variability.

For inference efficiency, CGER evaluates reconstruction errors on a sparse schedule. It samples at 10-step intervals in practice to limit inference overhead. Let τ_m denote the interaction step of scheduled evaluation m. After n scheduled evaluations, the cumulative error and active expert assignment are determined by

$$E^{(n)}_i = \frac{1}{n}\sum_{m=1}^{n} \operatorname{MSE}(\widehat{o}^i_{\tau_m}, o_{\tau_m}), \qquad \widehat{i}_n = \arg\min_{i\in\mathcal{C}} E^{(n)}_i, \tag{8}$$

where cumulative averaging over scheduled checkpoints attenuates transient sensory noise, and the selected assignment $\widehat{i}_n$ is held between intervals to prevent mid-trajectory policy oscillation.

4 Experiments

4.1 Experimental setup

Benchmarks. We evaluate knowledge acquisition and retention across two complementary continual reinforcement-learning settings. Atari evaluates adaptation to abrupt, heterogeneous dynamics as the agent learns sequentially across six distinct games, including Ms. Pac-Man, Boxing, Crazy Climber, Frostbite, Seaquest, and Enduro. All Atari tasks use 64×64 RGB observations, the full action space, and an action repeat of four, following conventions (Hafner et al., 2025; Alyahya et al., 2026). CoinRun complements this setting by testing retention and atom reuse under progressive visual and viewpoint perturbations over a shared dynamical foundation. Starting from the original

Method	ACC↑	$F\downarrow$	min-ACC↑	WC-ACC↑
DREAMERV3	0.1099 [0.1084, 0.2017]	1.9894 [1.7186, 2.2350]	-0.0549 [-0.1336, -0.0485]	-0.0106 [-0.0253, 0.0129]
ARROW	0.6187 [0.6058, 0.6259]	0.3326 [0.2774, 0.3359]	0.6226 [0.5865, 0.6250]	0.5429 [0.5077, 0.5481]
AWM	**1.9321** [1.5687, 2.9629]	**0.0078** [-0.3043, 0.0907]	**1.5927** [1.1599, 3.1281]	**1.3775** [1.0319, 2.6380]

Table 1: **Results on Atari.** Each cell reports the seed-level median and $[Q_{25}, Q_{75}]$ across five seeds. Bold indicates the best result.

Method	Raw avg.↑	ACC↑	$F\downarrow$	min-ACC↑	WC-ACC↑
DREAMERV3	5.5452 [4.6752,5.7048]	0.7352 [0.5277,0.7745]	0.4034 [0.3336,0.5863]	-0.3287 [-0.4869,-0.0843]	-0.1346 [-0.2343,0.0591]
ARROW	**6.9727** [5.9375,7.0247]	1.0222 [0.7699,1.0346]	-0.0076 [-0.0185,0.2311]	0.8812 [0.6432,0.8902]	0.8252 [0.6164,0.8495]
AWM	6.9401 [6.8652,6.9889]	**1.1804** [1.1695,1.2210]	**-0.0101** [-0.0410,0.0304]	**1.0017** [0.9515,1.0640]	**1.0268** [1.0187,1.0450]

Table 2: **Results on CoinRun.** Each cell reports the seed-level median and $[Q_{25}, Q_{75}]$ across five seeds. Bold indicates the best result.

CoinRun environment, the curriculum successively removes the background (+NB), restricts themes (+RT), introduces procedurally generated assets (+GA), applies monochrome styling (+MA), and removes agent-centered camera tracking (+CA).

Baseline methods. We benchmark AWM against DREAMERV3/FIFO and ARROW-50. DREAMERV3/FIFO updates a monolithic shared world model using standard first-in, first-out experience replay (Hafner et al., 2025). ARROW-50 extends this foundation by splitting replay capacity equally between FIFO storage and a long-term diversity memory designed to preserve historical transitions (Alyahya et al., 2026). All methods share the same task presentation sequence, per-task training-interaction budget, and total replay capacity of 1,024 trajectories.

Metrics. We assess lifelong-learning performance using four established criteria, namely average performance (ACC), average post-acquisition minimum performance (min-ACC), worst-case performance (WC-ACC), and backward forgetting (F) (Alyahya et al., 2026; De Lange et al., 2023). WC-ACC integrates active-task performance with historical minima to summarize the stability–plasticity trade-off. Following ARROW (Alyahya et al., 2026), all scores use the same task-specific random-policy and end-budget single-task returns as fixed normalization anchors. Unnormalized per-task returns and full metric definitions are provided in the supplementary material.

4.2 MAIN RESULTS

We evaluate lifelong retention and adaptation across two complementary benchmarks, as summarized in Tables 1 and 2. The sequential Atari suite assesses robustness to discrete, highly heterogeneous dynamics shifts, while procedurally generated CoinRun evaluates retention under continuous, cumulative visual distortions. Across both domains, AWM consistently establishes a superior retention frontier compared with replay-augmented and shared-model baselines.

As illustrated by the lifelong learning curves in Figure 3, AWM reliably shields historical competencies from interference as new environments are introduced. In contrast, standard DREAMERV3 undergoes severe representational collapse upon task switches, frequently deteriorating toward the random-policy floor. Across five random seeds, AWM attains a median ACC of 1.9321 compared with 0.6187 for ARROW, while reducing catastrophic forgetting to a near-zero level ($F = 0.0078$; Table 1). The margins in min-ACC and WC-ACC confirm that AWM maintains a robust performance floor across the task sequence, rather than stabilizing only at the final evaluation checkpoint.

At the individual-game level, AWM outperforms ARROW in final raw return across all six games, with pronounced gains on complex dynamics such as Frostbite, where median return improves from 260.6 to 1,871.9. The methods also have non-overlapping interquartile ranges across all four aggregate metrics. Section 4.3 examines the contributions of individual components.

In procedurally generated CoinRun environments, the primary strength of AWM lies in its resilience to cumulative visual corruptions, as shown in Figure 4. AWM achieves a median WC-ACC of 1.0268 compared with 0.8252 for ARROW, alongside negligible median forgetting ($F = -0.0101$; Table 2).



Figure 3: **Normalized lifelong performance across sequential Atari environments.** Thick segments indicate the actively trained task, vertical dashed lines denote task boundaries, and shading represents interquartile ranges across five seeds.

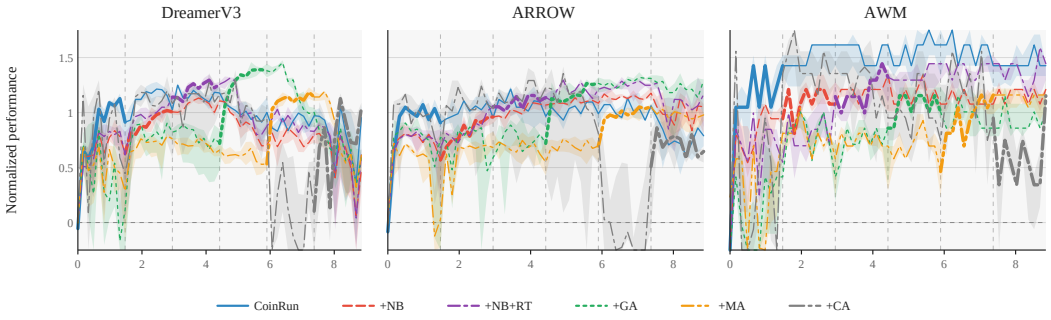

Figure 4: **Normalized lifelong performance across sequential CoinRun variants.** Thick segments indicate the actively trained variant, vertical dashed lines denote task boundaries, and shading represents interquartile ranges across five seeds.

Final raw-average return medians remain comparable (6.9401 for AWM and 6.9727 for ARROW), with AWM matching or exceeding ARROW across five of the six environment variants.

The final camera-shift variant highlights a classic stability-plasticity tradeoff. Specifically, unconstrained DREAMERV3 achieves a higher raw score than AWM (5.9180 versus 5.1172) on this isolated final variant by catastrophically overwriting prior representations, whereas AWM deliberately bounds plasticity on the active task to preserve the integrity of historical competence.

4.3 ABLATION STUDIES

Table 3 shows that Full AWM achieves the strongest median retention scores among the component configurations. CRM dynamics adapters support task-specific adaptation, and historical atom reuse makes previously learned dynamics available to new experts. Both contribute to improved min-ACC and WC-ACC, highlighting their importance for maintaining performance throughout the sequence. Teacher distillation provides the largest retention benefit among the tested components, emphasizing the value of preserving historical states, predictions, and policy behavior as the shared model evolves. Shared-gradient projection further improves retention by coordinating current-task learning with historical-task objectives. These results support the joint role of reusable dynamics and protected shared-model updates in continual learning.

The RCC controls in Table 3 demonstrate a favorable retention–capacity trade-off: AWM retains fewer world-model parameters while achieving higher median retention scores than both uncompressed variants. The advantage persists when the recovery-update budget is matched, supporting the contribution of return-guided adapter compression. By selecting adapter width according to completed-task return, RCC links retained capacity to the task behavior that each expert must preserve for subsequent reuse. The oracle-readout comparison in the same table examines deployment using

Variant	WM parameters (M)	ACC↑	min-ACC↑	WC-ACC↑
Shared world model	19.50	1.1520 [0.8479, 1.3562]	1.1089 [0.1328, 1.3851]	1.0865 [0.8539, 1.2190]
w/o CRM dynamics adapters	21.42 [21.17, 22.50]	1.5373 [1.2360, 1.8387]	1.2415 [0.9982, 1.4848]	1.1701 [0.9407, 1.3994]
w/o historical atom reuse	34.05 [32.62, 34.52]	1.4790 [1.1283, 2.0674]	1.0850 [0.9321, 2.0411]	1.0039 [0.8310, 1.7931]
w/o teacher distillation	33.57 [33.33, 34.76]	1.0011 [0.9222, 1.6951]	0.9369 [0.8095, 1.5733]	0.8664 [0.7579, 1.3691]
w/o shared-gradient projection	36.42 [35.00, 36.66]	1.4036 [1.1662, 1.9792]	1.3892 [1.1278, 1.8815]	1.2472 [1.0158, 1.6574]
w/o RCC	42.60	1.1179 [0.9732, 1.1462]	1.0636 [0.9927, 1.1866]	0.9815 [0.8892, 1.0478]
w/o RCC, matched updates	42.60	1.5051 [1.2101, 1.8001]	1.2377 [0.9951, 1.4803]	1.1032 [0.8870, 1.3194]
AWM, oracle readout	37.85 [35.00, 38.80]	1.9550 [1.5718, 2.3382]	1.6043 [1.2898, 1.9187]	1.4422 [1.1595, 1.7249]
Full AWM	37.85 [35.00, 38.80]	1.9321 [1.5687, 2.9629]	1.5927 [1.1599, 3.1281]	1.3775 [1.0319, 2.6380]

Table 3: **Ablation studies on Atari.** Results are reported as the median and $[Q_{25}, Q_{75}]$ across five random seeds. Shared world model excludes task-specific adapters and observation projectors. The teacher-distillation ablation removes $\mathcal{L}_{\text{cons}}$ during acquisition and RCC recovery, and adapter-output matching during recovery; shared-gradient projection and return checks remain. The matched-updates control removes RCC while matching its recovery-update budget.

identical trained checkpoints. CGER remains close to oracle routing across the retention metrics, supporting cumulative reconstruction error as an effective signal for selecting among accumulated experts during task-agnostic execution.

Figure 5 shows how retained capacity accumulates across the task sequence. In both Atari and CoinRun, the shared world model maintains a constant parameter footprint, and added capacity is localized to task-specific dynamics adapters and observation projectors. Posterior adapters account for the largest share of the growth, with smaller contributions from recurrent and prior adapters and lightweight observation projectors. Atari accumulates more adapter capacity than CoinRun, while both benchmarks retain the same shared-backbone size. Read together with the RCC controls in Table 3, these trends show how backbone sharing and boundary compression support an expanding library of reusable experts with controlled per-task capacity.

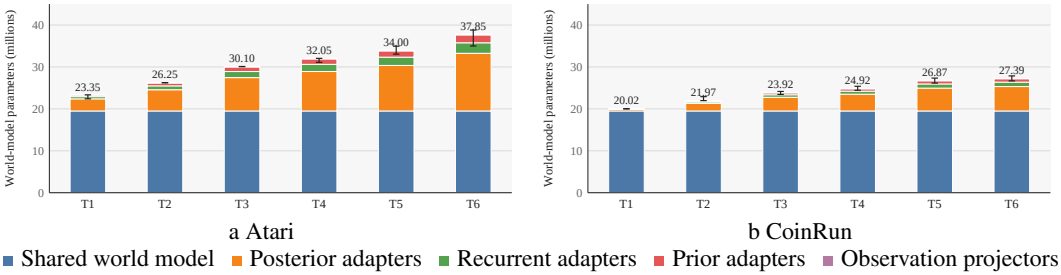

Figure 5: **Retained world model capacity.** Cumulative parameter counts after each task sequence with error bars indicating interquartile ranges across five seeds.

5 Conclusion

In this work, we introduced AWM, a continual model-based reinforcement learning framework that resolves the plasticity–stability dilemma by decoupling shared world-model evolution from task-specific dynamics preservation. Through the coordinated integration of composable dynamics adaptation, return-constrained parameter compression, and confidence-guided routing, AWM reliably shields historical competencies while adapting to novel transitions without requiring task identity labels at deployment. Empirical evaluations on sequential Atari and CoinRun benchmarks demonstrate that AWM virtually eliminates catastrophic forgetting and maintains robust performance frontiers across heterogeneous environments. Promising directions for future work include extending this modular dynamics formulation to continuous non-stationary environments without explicit task boundaries and scaling accumulative world models to real-world embodied control domains.

Reproducibility Statement

To support reproducibility, the supplementary material provides detailed pseudocode, implementation and evaluation details, experimental settings, and results from repeated runs. We will publicly release the complete source code and the configurations needed to reproduce the reported experiments.

AI Use Statement

Generative AI tools assisted with language editing, manuscript organization, and implementation support. The authors are responsible for checking all mathematical statements, experimental evidence, references, and final claims.

References

Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3366–3375, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Aljundi_Expert_Gate_Lifelong_CVPR_2017_paper.html.

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

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.00420.

Geng Chen, Wendong Zhang, Han Lu, Siyu Gao, Yunbo Wang, Mingsheng Long, and Xiaokang Yang. Continual predictive learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10728–10737, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Chen_Continual_Predictive_Learning_From_Videos_CVPR_2022_paper.pdf.

Matthias De Lange, Gido M. van de Ven, and Tinne Tuytelaars. Continual evaluation for lifelong learning: Identifying the stability gap. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zy350cRstc6.

Fei Deng, Junyeong Park, and Sungjin Ahn. Facing off world model backbones: RNNs, transformers, and S4. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6c65eb9b56719c1aa45ff73874de317-Abstract-Conference.html.

Haotian Fu, Yixiang Sun, Michael Littman, and George Konidaris. Knowledge retention in continual model-based reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17832–17851, 2025. URL https://proceedings.mlr.press/v267/fu25f.html.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2170–2179, 2019. URL https://proceedings.mlr.press/v97/gelada19a.html.

Ignat Georgiev, Varun Giridhar, Nick Hansen, and Animesh Garg. PWM: Policy learning with multi-task world models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/240ea1741b205ea295721d55184ac43b-Abstract-Conference.html.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2011.03506.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.01603.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02193.

Danijar Hafner, Kuang-Huei Lee, Ian Fischer, and Pieter Abbeel. Deep hierarchical planning from pixels. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a766f56d2da42cae20b5652970ec04ef-Abstract-Conference.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.16828.

Nicklas A. Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406, 2022. URL https://proceedings.mlr.press/v162/hansen22a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Yizhou Huang, Kevin Xie, Homanga Bharadhwaj, and Florian Shkurti. Continual model-based reinforcement learning with hypernetworks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 799–805, 2021. doi: 10.1109/ICRA48506.2021.9560793. URL https://ieeexplore.ieee.org/document/9560793.

Klemens Iten, Bruce Lee, Chenhao Li, Lenart Treven, Andreas Krause, and Bhavya Sukhija. Model-based reinforcement learning for control under time-varying dynamics. *arXiv preprint arXiv:2604.02260*, 2026. URL https://arxiv.org/abs/2604.02260.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for Atari. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xCPJHtDB.

Samuel Kessler, Mateusz Ostaszewski, Michał Paweł Bortkiewicz, Mateusz Żarski, Maciej Wolczyk, Jack Parker-Holder, Stephen J. Roberts, and Piotr Miłoś. The effectiveness of world models for continual reinforcement learning. In *Proceedings of the Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 184–204, 2023. URL https://proceedings.mlr.press/v232/kessler23a.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2404.00228.

Zichen Liu, Guoji Fu, Chao Du, Wee Sun Lee, and Min Lin. Continual reinforcement learning by planning with online world models. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38397–38423, 2025. URL https://proceedings.mlr.press/v267/liu25p.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08840.

Jorge A. Mendez, Harm van Seijen, and Eric Eaton. Modular lifelong reinforcement learning via neural composition. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2207.00429.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

Gurp Nijjer. The world model remembers, the actor forgets: Dream rehearsal for continual model-based RL. *arXiv preprint arXiv:2607.19749*, 2026. URL https://arxiv.org/abs/2607.19749.

Minting Pan, Wendong Zhang, Geng Chen, Xiangming Zhu, Siyu Gao, Yunbo Wang, and Xiaokang Yang. Continual visual reinforcement learning with a life-long world model. In *Machine Learning and Knowledge Discovery in Databases. Research Track (ECML PKDD 2025)*, volume 16018 of *Lecture Notes in Computer Science*, pp. 146–162, 2026. doi: 10.1007/978-3-032-06106-5_9.

Sam Powers, Eliot Xing, and Abhinav Gupta. Self-activating neural ensembles for continual reinforcement learning. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 683–704, 2022. URL https://proceedings.mlr.press/v199/powers22a.html.

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TdBaDGCpjly.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. URL https://arxiv.org/abs/1606.04671.

Paul Ruvolo and Eric Eaton. ELLA: An efficient lifelong learning algorithm. In *Proceedings of the International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 507–515, 2013. URL https://proceedings.mlr.press/v28/ruvolo13.html.

Mohammad Reza Samsami, Artem Zholus, Janarthanan Rajendran, and Sarath Chandar. Mastering memory tasks with world models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vDArHJ68h.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537, 2018. URL https://proceedings.mlr.press/v80/schwarz18a.html.

Ruixiang Sun, Hongyu Zang, Xin Li, and Riashat Islam. Learning latent dynamic robust representations for world models. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47234–47260, 2024. URL https://proceedings.mlr.press/v235/sun24n.html.

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. FOSTER: Feature boosting and compression for class-incremental learning. In *European Conference on Computer Vision*, 2022a. URL https://arxiv.org/abs/2204.04662.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b. URL https://arxiv.org/abs/2112.08654.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. DayDreamer: World models for physical robot learning. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2226–2240, 2023. URL https://proceedings.mlr.press/v205/wu23c.html.

Mengdi Xu, Wenhao Ding, Jiacheng Zhu, Zuxin Liu, Baiming Chen, and Ding Zhao. Task-agnostic online reinforcement learning with an infinite mixture of gaussian processes. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/47951a40efc0d2f7da8ff1ecbfde80f4-Abstract.html.

Yifan Xu, Nicklas Hansen, Zirui Wang, Yung-Chieh Chan, Hao Su, and Zhuowen Tu. On the feasibility of cross-task transfer with model-based reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KB1sc5pNKFv.

Shipeng Yan, Jiangwei Xie, and Xuming He. DER: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yan_DER_Dynamically_Expandable_Representation_for_Class_Incremental_Learning_CVPR_2021_paper.html.

Luke Yang, Levin Kuhlmann, and Gideon Kowadlo. Augmenting replay in world models for continual reinforcement learning. *arXiv preprint arXiv:2401.16650*, 2024. URL https://arxiv.org/abs/2401.16650.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995, 2017. URL https://proceedings.mlr.press/v70/zenke17a.html.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.10742.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. STORM: Efficient stochastic transformer based world models for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5647763d4245b23e6a1cb0a8947b38c9-Abstract.html.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

Haoyu Zhou, Joe Watson, Anson Lei, and Ingmar Posner. Benchmarking world models for continual learning on compositional tasks. *arXiv preprint arXiv:2609.22055*, 2026. URL https://arxiv.org/abs/2609.22055.