Leveraging Episodic Memory to Improve World Models for Reinforcement Learning

Julian Coda-Forno[1] **Changmin Yu**[2]
Qinghai Guo[3] **Zafeirios Fountas**[3] **Neil Burgess**[2]
[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany;
[2]Institute of Cognitive Neuroscience, UCL, London, United Kingdom;
[3]Huawei Noah's Ark Lab, London, United Kingdom
{julian.coda-forno.20, changmin.yu.19, n.burgess@}@ucl.ac.uk
{guoqinghai, zafeirios.fountas@huawei.com}

Abstract

Poor sample efficiency plagues the practical applicability of deep reinforcement learning (RL) algorithms, especially compared to biological intelligence. In order to close the gap, previous work have proposed to augment the RL framework with an analogue of biological episodic memory, leading to the emerging field of "episodic control". Episodic memory refers to the ability to recollect individual events independent of the slower process of learning accumulated statistics, and evidence suggests that humans can use episodic memory for planning. Existing attempts to integrate episodic memory components into RL agents have mostly focused on the model-free domain, leaving scope for investigating their roles under the model-based settings. Here we propose the **E**pisodic **M**emory Module for **W**orld **M**odels (EMWM) to aid learning of world-model transitions, instead of value functions for standard Episodic-RL. The EMWM stores latent state transitions that have high prediction-error under the model as memories, and uses linearly interpolated memories when the model shows high epistemic uncertainty. Memories are dynamically updated with a timescale reflecting their continual surprise and uncertainty. Implemented in combination with existing world-model agents, the EMWM produces a boost in performance over baseline agents on complex Atari games such as Montezuma's Revenge. Our results indicate that the EMWM can temporarily fill in gaps while a world model is being learned, giving significant advantages in complex environments where such learning is slow.

1 Introduction

The neuroscience-inspired concept of episodic memory (EM) [1–4] has been proposed as a potential solution to improving the sample inefficiency of Deep RL agents [4]. This concept refers to the recollection of personal experiences and is thought to be of critical importance for control at very early stages of learning [1]. Mathematical models of EM leverage non-parametric statistics - through single-event recalls - instead of the slower-to-learn accumulated statistics of semantic memory. Neuroscience studies suggest that EM in mammals is beneficial for control, and it is also used in the process of modelling the environment's dynamics, often referred to as 'world models' in RL [5, 6].

This has inspired new approaches in the field of Episodic-RL that integrate EM into model-free RL [4], where policy updates are based on the non-parametric approximation of value function, given the EM component. Despite the success of this method, there have been few prior works [7–9] that address the combination of episodic control and model-based RL, which is often interpreted

Memory in Artificial and Real Intelligence (MemARI) Workshop at 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

as the more sample-efficient alternative to model-free methods. This has been labeled as episodic model-based control (EMBC) [10]. In this work, we propose the Episodic Memory Module for World Model Transitions (EMWM), an architecture that integrates EM with a state-of-the art world model, the Recurrent State-Space Model (RSSM) from Dreamer-V2 [11], to assist the latent state transitions at low experience and optimize sample efficiency.

2 Background

Standard Episodic-RL architectures [2, 3, 12, 4] estimate the state-action value Q using a buffer of pre-recorded memories. While the agent interacts online with the environment, it stores all visited states as keys to an action-specific dictionary and the associated state-action value Q as the dictionary value. $Q(s_t, a_t) = \sum_{i=0}^{H} r_t \gamma^i$ is stored as the sum of discounted rewards over a fixed horizon with length H. Instead of utilising a single parametric estimate of the action values, the episodic-RL models utilise a weighted average of similar state-action values retrieved from the cached episodic memory buffer, $\hat{Q}(s, a) = \sum_i w_i Q_i(s_i, a_i)$ [12]. The policy evaluation and policy improvement are based on these values under the standard model-free RL context. More recently, new Episodic-RL architectures have been proposed that perform episodic recalls beyond value approximation [13–15] and therefore we will refer to Episodic-RL as a general term referring to methods leveraging non-parametric mappings of possibly old events within an RL algorithm.

World models in model-based RL entail the usage of latent state-space models as the dynamics model for projecting the direct modelling of transition dynamics in the observable space down to some low-dimensional latent space. The associated RL agent operates on the latent space, and is trained through generative predictions of future latent states and rewards. Dreamer-V2 is a prominent work in this field [11], leveraging a generative recurrent world model (RSSM [16]) when learning an actor and value networks (actor-critic) by backpropagating their analytical gradient through the model dynamics. The training of the world model in Dreamer-V2 is shown graphically in Figure 1A.

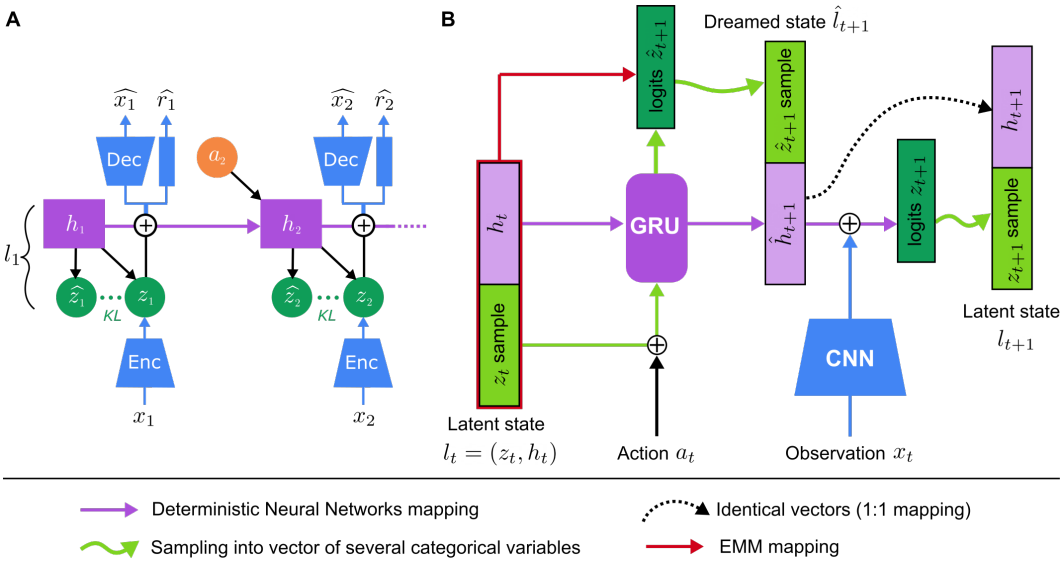

Figure 1: **A:** Training of the world model in DreamerV2. The model resembles the features of a sequential VAE, and is trained by minimising the variational free energy. $\hat{z}_t$ and z_t are the predicted prior and posterior latent codes at time t, respectively. h_t represents the deterministic hidden state of the GRU [17] in the RSSM. Graphical illustration adopted from [11], maintaining consistency in colours. **B:** Graphical illustration of RSSM with incorporation of EMWM mapping.

3 Methods

3.1 Episodic Memory Module for World Model Transitions (EMWM)

EMWM here is defined as a memory buffer that records individual state transitions that were marked as "surprising", and recalls these transitions in order to substitute semantic memory in the event of high epistemic uncertainty. This buffer is implemented in the form of a dictionary, whose entries use as a "key" the current latent state $l_t = (z_t, h_t)$ and as a value the predictive prior distribution over the next state $p(z_{t+1}|z_t, h_t) = \hat{z}_{t+1}$ as shown in Figure 1B.

Figure 2 summarizes the process of memory formation (storage) and recall (prediction) using the EMWM's dictionaries. EMWM employs action-dependent dictionaries, in order to estimate the complete prior $\hat{z}_{t+1} \approx p(z_{t+1}|z_t, a_t = a), \forall a \in \mathcal{A}$. Therefore, one limitation of the current version of EMWM is that it can only be implemented in environments with discrete action spaces. These dictionaries are populated as the agent interacts with the environment, while dictionary lookups are used to estimate the next latent state transition prior when predicting future states, either for planning or dreaming-based policy training. Dictionary lookups are performed by calculating the weighted average (kernel $k(l, l_i) = \frac{1}{\|l-l_i\|_2^2+\delta}$) of the k=5 nearest neighbours, with respect to a given state.

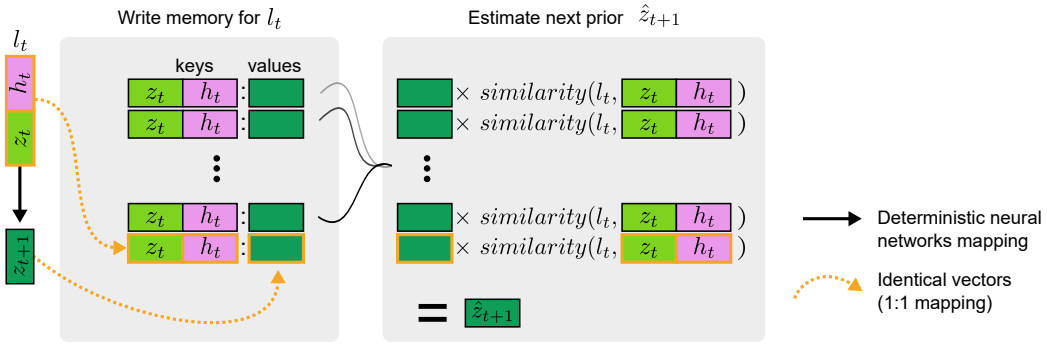

Figure 2: EMWM architecture for memory formation and recall, given an action a_t.

One might be sceptical that latent state representations learnt from accumulated statistics (analogous to semantic memory) would be stored in EM. However, episodic information is often defined with respect to semantic knowledge, which can be an important determinant of performance [18] and hippocampal activity [19] in episodic memory experiments. Here, the one-shot learning of episodic information (i.e. storing a specific transition between latent states) occurs on a much faster timescale than the learning of semantic (i.e. latent state) representations themselves. Nonetheless, after enough further learning has occurred, the stored and current representations may no longer align, effectively losing access to very old memories. Such a problem might underlie a phenomenon called *infantile amnesia*, which is the inability to retrieve EM before the age of two to four years old [20]. This motivates the importance of how is our EM module updated and used over time which is presented in the following section.

3.2 EMWM integration with Dreamer-V2

It has been suggested that leveraging EM is most useful during surprising events [1]. This insight can also be applied to our framework, if the non-parametric statistics of EMWM are used when the agent's predicted state is surprising and, therefore, the prediction of the neural networks will most likely be poor. Indeed, stochastic gradient descent optimisation requires the use of small learning rates [3]. Due to the global approximation nature of neural networks, high learning rates cause catastrophic interference. In other words, this means that new experience can only be incorporated into a neural network at a slow pace. States that are still "surprising" will not be accommodated by the global function approximation yet. Therefore, we only use EMWM for prediction when the neural networks are uncertain and we only store and keep the most surprising memories over time.

We capture uncertainty using the ensemble-based epistemic uncertainty estimate [21, 22]. The $x_{pred} = 5\%$ most uncertain states (l_t, a_t) are predicted by EMWM instead of the world model. The

amount of EM we can store is limited by computational (or neuronal) resources. Here, we store the $x_{storage} = 1$ % most surprising states and their respective surprise. Surprise is captured using the KL divergence between the prior and posterior (which is given during observations in world model training) which we denote KL_{obs}. We substitute the least relevant memories where relevance is inspired by Forgetting curves [23] in the brain and their recall probabilities $R = e^{-t/S}$ where t is the time step, S is a stability constant, which we capture by $S = KL_{obs} \times x_{scaling}$. Whenever a memory is used for prediction, we consider memory consolidation [24], and we reset it to $R = 1$. This provides the ability to only keep memories that were very surprising as well as consolidated over time. Finally, to ensure that neural networks take over a given state transition after enough experience, we introduce a final update process to the EMWM. As research suggests, the brain replaces EM with semantic memory mappings when enough experience has been encountered [1, 25]. Therefore, for each $x_{upd} = 100$ time step, we evaluate the epistemic uncertainty for all existing entries in EMWM and delete the $x_{del} = 0.1$ % most "certain" memory transitions. This ensures that even memories which were highly surprising at storage, but now accommodated by the global function approximation from neural networks, are still deleted. Algorithm 1 describes the generic scheme of how EMWM can be used in model-based RL for imagination.

Algorithm 1 Semi-parametric training of EMWM

Initialize policy π_ϕ, predictive world-model p_θ , dataset D and the EMWM, $\Psi = (R, a_t, z_t, z_{t+1})_{n=1}^{N}$.
for t in range(T) **do**
  Add experience to D by interacting with the environment using π_ϕ
  Draw B data sequences $(a_t, o_t, r_t, o_{t+1}) \sim D$
  $R^n \times = e^{\frac{-1}{KL_{obs}^n \times x_{scaling}}} \forall n \in \{0, 1...N\}$
  while Training world-model p_θ on $B^{(i)} \forall i \in B$ **do**
    $KL_{obs}{}^{(i)} = KL(p_\theta(\hat{z_{t+1}}|z_t, a_t)^{(i)} || p_\theta(z_{t+1}|z_t, a_t, o_{t+1})^{(i)})$
    if $KL_{obs}{}^{(i)} > KL_{obs}$ top $x_{storage}^{th}$-percentile **then**
      $R^{(i)} = e^{\frac{-1}{KL_{obs}^{(i)} \times x_{scaling}}}$
      Delete $(a_t, z_t, z_{t+1}, KL_{obs})^n$ in Ψ where $n = argmin_n R_{n=1}^N$
      Store $(a_t, z_t, z_{t+1}, KL_{obs})^{(i)}$ in Ψ
    end if
  end while
  Imagine C trajectories $(z_\tau, a_\tau)_{\tau=t}^{t+H}$
  while Training policy π_ϕ on C **do**
    if $\Phi_i > \Phi_{thresh} \forall i \in C$ (where $thresh$ is the x_{pred}^{th}-percentile highest $\Phi_j \forall j \in C$) **then**
      Replace the prediction $p_\theta(z_{\tau+1}|z_t, a_t)$ with $\sum_n z_{t+1}^n \times similarity(z_\tau, z_t^n|a_t)$, for $n = 1, \ldots, N$
    end if
  end while
  if t % x_{upd} == 0 **then**
    Compute Φ_n, for $n = 1, \ldots, N$
    Set top $x_{del}\%$ of the memories (with respect to Φ_n) in Ψ to $R = 0$
  end if
end for
where:

Φ = Disagreement = $\sum_{m1=1}^{M} KL\Big(p_{\theta_{m1}}(\hat{z_{t+1}}|z_t, a_t) \,||\, M^{-1} \sum_{m2=1}^{M} p_{\theta_{m2}}(\hat{z_{t+1}}|z_t, a_t)\Big)$
N = Size of EMWM dictionnary
M = Number of Ensemble Units
T = Total number of training steps
H = Horizon

4 Experiments

We investigated whether our proposed EMWM-DreamerV2 model improves the sample efficiency of learning over the baseline vanilla DreamerV2 agent on different Atari games [26]. The aim of the empirical studies is to show that our EMWM module optimizes the DreamerV2 transition predictions and, thus, lead to better policy training with "dreamed" roll-out trajectories. We evaluate on 6 selected Atari tasks, the implementation details can be found in Appendix B.

In Figure 3, we observe a substantial performance increase over DreamerV2, especially over the initial frames for three of the evaluated environments: River Raid, HERO and Montezuma-revenge, which interestingly are all exploration-demanding environments with extremely sparse rewards. The preliminary results cohere with our hypothesis that replacing prediction targets in model-training with linear interpolation of past experiences alleviates the heteroscedasticity and poor accuracy introduced by the parametric predictions, and is mostly reflected in the early stage of training.

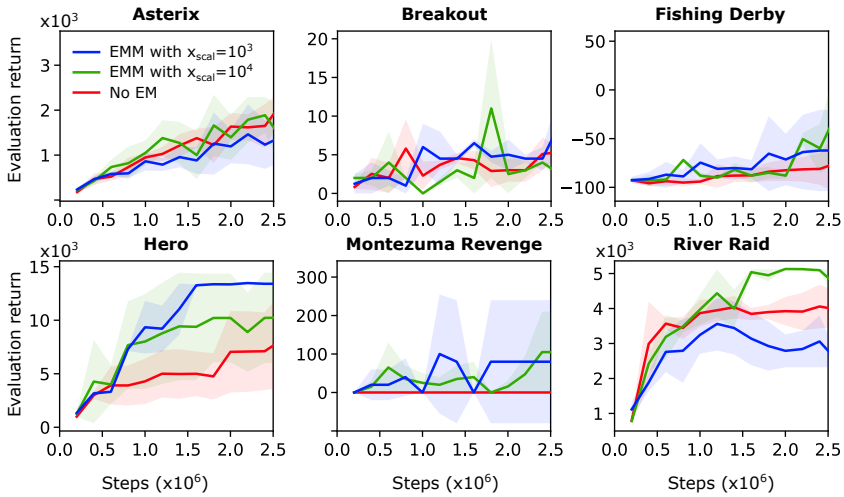

Figure 3: Evaluation return on selected Atari games. All evaluations are averaged over 5 random seeds.

5 Discussion and Future work

We have introduced a new way to leverage the properties of episodic memory in RL, namely for World-Model imaginations. This model serves as a proof of concept and many choices were only heuristics inspired by properties of human episodic memory. Nonetheless, the model has shown some success with sample efficiency improvements on complex environments. This highlights the importance for future research to gain more intuition on how to balance between semantic and episodic memory. This balance seems crucial for optimum learning from fast non-trivial predictions from non-parametric statistical mappings and slower to learn but asymptotically more performant accumulated statistics. This model's promise relies on how its choices of retrieval, storage and deletion of EM balance this memory system with the learnt accumulated statistics.

However, many aspects could be improved in future work. First, more analysis is required to gain better understanding of all the hyperparameter and architectural choices. This can be done with ablation studies, evaluations on other model-based RL agent (not only dreamer-v2) and different environments. This would provide a clearer picture of the potential of this proof of concept for which our analysis was limited due to lack of computational resources. One of the main focus but also maybe one of the most challenging one should be to understand how to gain intuition of the effect of forgetting curves, and how to modulate the proportion of predictions taken from EM with experience (the balance between semantic and episodic memory). Finally, this EMWM architecture is robust in terms of what it can combine with and therefore could be used across various different RL architecture for optimized sample-efficiency. One strong limitation is the retrieval time for predictions, and a hierarchical structure [27] could greatly improve this issue by only attending to several memories instead of looking at the entire EM.

References

[1] Máté Lengyel and Peter Dayan. Hippocampal contributions to control: The third way. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[2] Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. 06 2016.

[3] Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control, 2017.

[4] Matthew Botvinick, Sam Ritter, Jane X. Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23(5):408–422, 2019.

[5] Timothy E.J. Behrens, Timothy H. Muller, James C.R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

[6] Brad E. Pfeiffer and David J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79, May 2013.

[7] Greg Wayne, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack Rae, Piotr Mirowski, Joel Z. Leibo, Adam Santoro, Mevlana Gemici, Malcolm Reynolds, Tim Harley, Josh Abramson, Shakir Mohamed, Danilo Rezende, David Saxton, Adam Cain, Chloe Hillier, David Silver, Koray Kavukcuoglu, Matt Botvinick, Demis Hassabis, and Timothy Lillicrap. Unsupervised Predictive Memory in a Goal-Directed Agent. *arXiv:1803.10760 [cs, stat]*, March 2018. arXiv: 1803.10760.

[8] Hung Le, Thommen Karimpanal George, Majid Abdolshah, Truyen Tran, and Svetha Venkatesh. Model-Based Episodic Memory Induces Dynamic Hybrid Controls. *arXiv:2111.02104 [cs]*, November 2021. arXiv: 2111.02104.

[9] Sam Ritter, Ryan Faulkner, Laurent Sartran, Adam Santoro, Matt Botvinick, and David Raposo. Rapid Task-Solving in Novel Environments, April 2021. arXiv:2006.03662 [cs, stat].

[10] Daphna Shohamy Oliver Vikbladh and Nathaniel Daw. Episodic contributions to model-based reinforcement learning, 2017.

[11] Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *CoRR*, abs/2010.02193, 2020.

[12] Samuel J. Gershman and Nathaniel D. Daw. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68(1):101–128, 2017. PMID: 27618944.

[13] Qihong Lu, Uri Hasson, and Kenneth A Norman. A neural network model of when to retrieve and encode episodic memories. *eLife*, 11:e74445, February 2022. Publisher: eLife Sciences Publications, Ltd.

[14] Samuel Ritter, Jane X. Wang, Zeb Kurth-Nelson, Siddhant M. Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew Botvinick. Been There, Done That: Meta-Learning with Episodic Recall. *arXiv:1805.09692 [cs, stat]*, July 2018. arXiv: 1805.09692.

[15] Abigail N. Hoskin, Aaron M. Bornstein, Kenneth A. Norman, and Jonathan D. Cohen. Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. Technical report, bioRxiv, May 2018. Section: New Results Type: article.

[16] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[18] Chris M Bird, Rachel A Davies, Jamie Ward, and Neil Burgess. Effects of pre-experimental knowledge on recognition memory. *Learning & Memory*, 18(1):11–14, 2011.

[19] Iris Trinkler, John A King, Christian F Doeller, Michael D Rugg, and Neil Burgess. Neural bases of autobiographical support for episodic recollection of faces. *Hippocampus*, 19(8):718–730, 2009.

[20] Gregory L. Robinson-Riegler Bridget Robinson-Riegler. *Cognitive Psychology: Applying The Science of the Mind, 3rd Edition*. 2012.

[21] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[22] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. *CoRR*, abs/2005.05960, 2020.

[23] J. Murre Jaap M. and Joeri Dros. Replication and analysis of ebbinghaus' forgetting curve.

[24] Yadin Dudai. The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, 55(1):51–86, 2004. PMID: 14744210.

[25] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12):1704–1711, December 2005.

[26] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *J. Artif. Int. Res.*, 47(1):253–279, may 2013.

[27] Andrew Kyle Lampinen, Stephanie C. Y. Chan, Andrea Banino, and Felix Hill. Towards mental time travel: a hierarchical memory for reinforcement learning agents. *CoRR*, abs/2105.14039, 2021.

A Additional Analysis

In this section, we present additional results for EMWM-Dreamer. First, we investigated what metric could explain the improvements in sample efficiency. One interesting result, especially for HERO as shown in Figure 4, is the difference in Actor Loss. EMWM seem to exhibit larger actor loss for these environments, possibly meaning its latent predictions produce more accurate predictive latent state and reward signals leading to non-trivial policy gradients. Indeed, for HERO (which also has the most significant sample efficiency with respect to Dreamer-V2), the actor loss is larger by orders of magnitude.

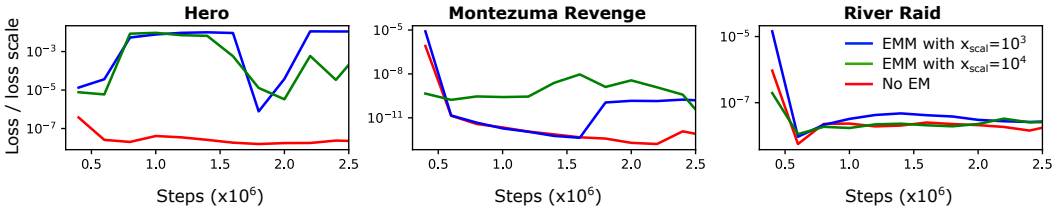

Figure 4: Actor Loss for environments where EMWM is significantly more sample efficient

We then investigated the difference in dream rendering between both models as shown in Figure 5. This experiment aimed to understand key differences from environments with extremely sparse rewards. We performed the same sequence of actions for 15 steps for both standard Dreamer-V2 and with EMWM after 2.5 million frames of learning (we used $x_{scaling} = 1000$). However, we only showed observations to the agents for the 5 first steps and then only performed the remaining 10 actions within their dreaming transitions. We tracked online environment interactions during the full 15 steps to have a ground truth baseline. We show the last 5 predictions using the world-model decoder and the first 5 online observations for all models. Interestingly, we notice a substantial difference between the two models. The main observation is that the agent location and actions is a lot more faithful to the ground truth with EMWM. The agent effectively stays stuck on the ladder in Dreamer-V2 predictions, whereas it can generate meaningful behaviours such as jumping with EMWM, even though both perform the same exact actions in the initial 5 states.

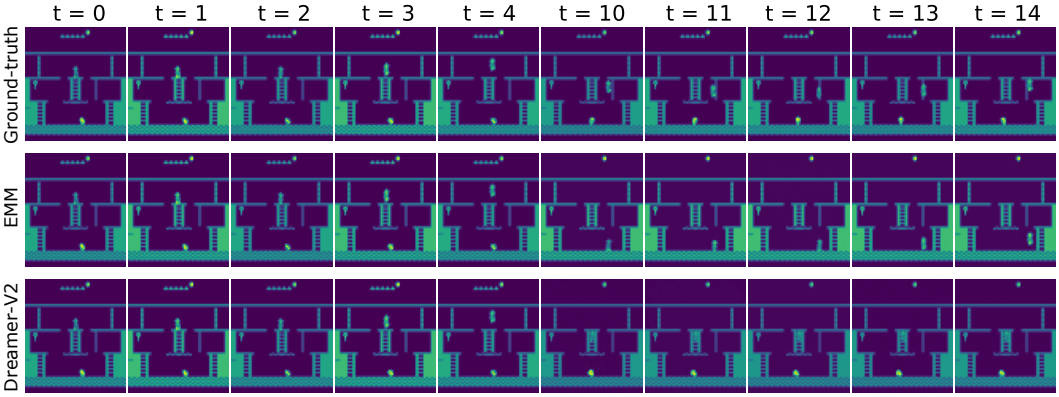

Figure 5: Dream rollout comparison in Montezuma-revenge

Figure 6 show both the surprise and uncertainty (disagreement ϕ of deep-ensembles) over time. More specifically, we plotted the top 1% surprise and 5% uncertainty replacements over 5 random seeds on H.E.R.O (presented in Section 5 with $x_{scaling} = 10^3$). Interestingly, we see that both the surprise and uncertainty have a similar behavior. They both highly decrease at the beginning (until approximately 600k steps) and then stabilizes to a more constant (non-zero) behavior. This highlights our expectation that during the early stage of training, "surprise" states are frequently encountered and the world model prediction exhibits high uncertainty. Additionally, the fact that the average surprise and uncertainty never reaches 0 suggests that the EMWM could aid learning even beyond the early stage of training by alleviating the negative impacts induced by undertraining is certain

parts of the environment. Therefore, these experiments indicate that the prediction replacements from EMWM, which we suggested should be decreased over time for future work, should not be decreased to 0%.

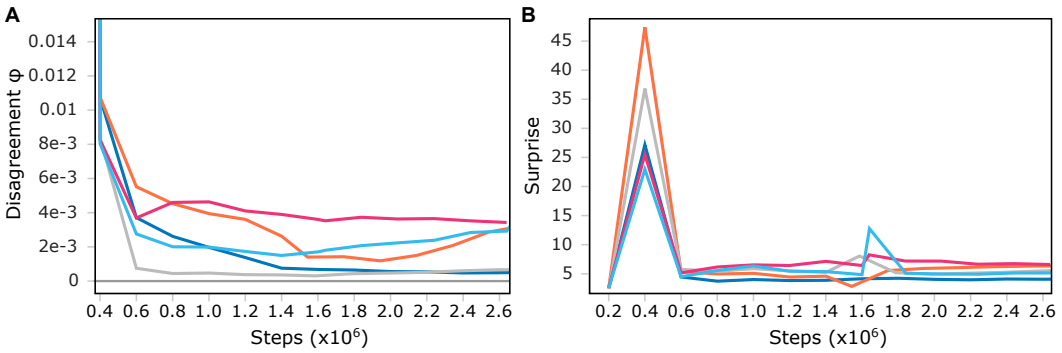

Figure 6: Top 5% disagreements (a) and 1% surprises (b) over time for H.E.R.O with each colour representing a different seed.

We also conducted a preliminary ablation study with different replacement proportion on H.E.R.O. We used as baseline runs from Section 5, EMWM-Dreamer which showed highest performance ($x_{pred} = 5\%$ and $x_{scaling} = 1000$) shown in Figure 7. We completed the plot with three different runs $x_{pred} \in \{1\%, 10\%, 100\%\}$ with fixed $x_{scaling} = 1000$. We also included a run with a simpler EM forgetting process (First in First Out as opposed to using our more involved forgetting process). All runs were performed for 5 different seeds. We can see that 1% and 10% behave similarly to the one with simpler EM forgetting process and suggest that for H.E.R.O, the choice of $x_{pred} = 5\%$ seems coherent with performance. It also suggests that our more involved forgetting process based on heuristics seems relevant at least on the Hero environment. Interestingly, $x_{pred} = 100\%$ which means the prior predictions are fully using EM, perform terribly, highlighting again the need to balance well between accumulated statistics and EM mappings.

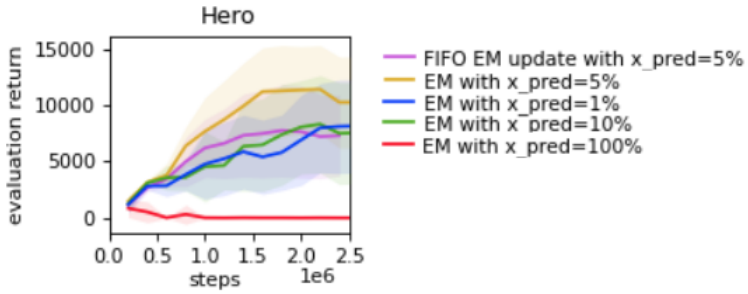

Figure 7: Ablation study on H.E.R.O

Finally, Table 1 show the comparison of performance on a longer time scale (10^7 training steps) with existing Episodic-RL models on the 4 selected Atari environments (MFEC [2], NEC [3], MBEC++ [8]). Note that the performance statistics for these baseline algorithms are shown directly as reported in their respective original works, and the evaluation of EMWM-Dreamer and Dreamer are based on one seed due to computational limitations on our side. As explained in Section 5, it is worth noting that the additional baselines fall under the model-free RL category and therefore expected to acquire better performance asymptotically. For instance, in H.E.R.O, EMWM-Dreamer converges very fast to this magnitude (1M frames) and does not improve given additional training. This could be explained by the fact that EMWM-Dreamer uses an actor-critic controller, and Policy gradient-based RL algorithms are well-known to be prone to local minimas as opposed to value-based RL algorithms. Indeed, we see that after $10M$ frames, both Dreamer-v2 and EMWM-Dreamer have very similar performance across the 4 presented games even though we showed in Section 4 that improved sample efficiency is achieved at lower experience. This highlights again the intuition that EMWM usage should be diminished over time for optimized results.

	MFEC [2]	NEC [3]	MBEC++ [8]	DREAMER [11]	EMWM-DREAMER
FISHING DERBY	-90.3	-72.2	17.6	**63**	59
H.E.R.O.	14767.7	**16265.3**	12148.5	13495	13775
MONTEZUMA REVENGE	76.4	42.1	0	400	**500**
RIVER RAID	4195.0	5498.1	**10656.4**	6880	7470

Table 1: Performance after 1×10^7 training steps of EMWM-Dreamer and the baseline agents on selected Atari games. The best performance for each game shown in bold [3, 8].

B Implementation Details

The most difficult EMWM hyperparameter to choose was the $x_{scaling}$. This has a significant impact to the behavior of the forgetting recall curve as it scales the exponential decay, while different Atari environments have different timescales. We tried both $x_{scaling} \in \{1000, 10000\}$ and leave the question of finding intuitively robust and adaptable forgetting curves for a given environment. The rest of the hyperparameters were set and explained in the Method section. Figure 3 shows the evaluation return on 6 environments, all ran for 5 different seeds until 2.5 million frames.
Regarding Dreamer-v2 we chose to keep the default hyperparameters except for two components, the frequency of the world-model training from every 16 to 1 steps and the RSSM training batch and length size from 50 to 30. The former choice was made because we are analyzing sample efficiency from learning fast world models and makes it therefore a more competitive comparison. This also implies that comparison of sample-efficiency to previously stated results on the Dreamer-v2 paper are not comparable. The latter choice was made due to GPU memory restrictions when using our EM. We aimed to keep the usage of our model within a single 12 GB GPU. We therefore also had a limitation on the dictionary sizes. We only kept 18 (Atari action space) dictionaries of size 500 which also made the selectivity of storage even more relevant.