POPGYM: BENCHMARKING PARTIALLY OBSERVABLE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Real world applications of Reinforcement Learning (RL) are often partially observable, thus requiring memory. Despite this, partial observability is still largely ignored by contemporary RL benchmarks and libraries. We introduce Partially Observable Process Gym (POPGym), a two-part library containing (1) a diverse collection of 14 partially observable environments, each with multiple difficulties and (2) implementations of 13 memory model baselines – the most in a single RL library. Existing partially observable benchmarks tend to fixate on 3D visual navigation, which is computationally expensive and only one type of many possible POMDPs. In contrast, POPGym environments are diverse, produce smaller observations, use less memory, and often converge within two hours of training on a consumer-grade GPU. We implement our high-level memory API and memory baselines on top of the popular RLlib framework, providing plug-and-play compatibility with various training algorithms, exploration strategies, and distributed training paradigms. Using POPGym, we execute the largest comparison across RL memory models to date. POPGym is available at `https://anonymous.4open.science/r/popgym-51D8/README.md`.

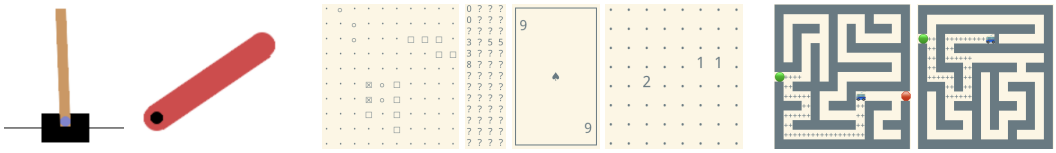
1 INTRODUCTION

Datasets like MNIST (Lecun et al., 1998) have driven advances in Machine Learning (ML) as much as new architectural designs (Levine et al., 2020). Comprehensive datasets are paramount in assessing the progress of learning algorithms and highlighting shortcomings of current methodologies. This is evident in the context of RL, where the absence of fast and comprehensive benchmarks resulted in a reproducability crisis (Henderson et al., 2018). Large collections of diverse environments, like the Arcade Learning Environment, OpenAI Gym, ProcGen, and DMLab provide a reliable measure of progress in deep RL. These fundamental benchmarks play a role in RL equivalent to that of MNIST in supervised learning (SL).

The vast majority of today's RL benchmarks are designed around Markov Decision Processes (MDPs). In MDPs, agents observe a *Markov state*, which contains all necessary information to solve the task at hand. When the observations are Markov states, the Markov property is satisfied, and traditional RL approaches guarantee convergence to an optimal policy (Sutton & Barto, 2018, Chapter 3). But in many RL applications, observations are ambiguous, incomplete, or noisy – any of which makes the MDP *partially observable* (POMDP) (Kaelbling et al., 1998), breaking the Markov property and all convergence guarantees. Furthermore, Ghosh et al. (2021) find that policies trained under the ideal MDP framework cannot generalize to real-world conditions when deployed, with epistemic uncertainty transforming real-world MDPs into POMDPs. By introducing *memory* (referred to as sequence to sequence models in SL), we can summarize the observations[1] therefore restoring policy convergence guarantees for POMDPs (Sutton & Barto, 2018, Chapter 17.3).

Despite the importance of memory in RL, most of today's comprehensive benchmarks are fully observable or near-fully observable. Existing partially observable benchmarks are often navigation-based – representing only spatial POMDPs, and ignoring applications like policymaking, disease diagnosis, teaching, and ecology (Cassandra, 1998). The state of memory-based models in RL

[1]Strictly speaking, the agent actions are also required to guarantee convergence. We consider the previous action as part of the following observation.



(a) Stateless Cartpole and Stateless Pendulum (b) Battleship, Concentration, Higher Lower and Mine Sweeper (c) Labyrinth Escape and Explore

Figure 1: Renders from select POPGym environments.

libraries is even more dire. We are not aware of any RL libraries that implement more than three or four distinct memory baselines. In nearly all cases, these memory models are limited to frame stacking and LSTM.

To date, there are no popular RL libraries that provide a diverse selection of memory models. Of the few existing POMDP benchmarks, even fewer are comprehensive and diverse. As a consequence, there are no large-scale studies comparing memory models in RL. We propose to fill these three gaps by introducing our framework, POPGym.

1.1 Contributions

POPGym is a collection of 14 partially observable gym environments (Fig. 1) and 13 memory baselines. All environments come with at least three difficulty settings and randomly generate levels to prevent overfitting. The POPGym environments use low-dimensional observations, making them fast and memory efficient. Many of our baseline models converge in under two hours of training on a single consumer-grade GPU (Appendix 1, 2). The POPGym memory baselines utilize a simple API built on top of the popular RLlib library (Liang et al., 2018), seamlessly integrating memory models with an assortment of RL algorithms, sampling, exploration strategies, logging frameworks, and distributed training methodologies. Utilizing POPGym and its memory baselines, we execute a large-scale evaluation, analyzing the capabilities of memory models on a wide range of tasks. To summarize, we contribute

1. A comprehensive collection of diverse POMDP tasks
2. The largest collection of memory baseline implementations in an RL library
3. A large-scale, principled comparison across memory models

2 Related Work

There are many existing RL benchmarks, which we categorize as fully (or near-fully) observable and partially observable. In near-fully observable environments, large portions of the the Markov state are visible in an observation, though some information may be missing. We limit our literature review to *comprehensive* benchmarks (those that contain a wide set of tasks), as environment diversity is essential for the accurate evaluation of RL agents (Cobbe et al., 2020).

2.1 Fully and Near-Fully Observable Benchmarks

The Arcade Learning Environment (ALE) (Bellemare et al., 2013) wraps Atari 2600 ROMs in a Python interface. Most of the Atari games, such as Space Invaders or Missile Command are fully observable (Cobbe et al., 2020). Some games like asteroids require velocity observations, but models can recover full observability by stacking four consecutive observations (Mnih et al., 2015), an approach that does not scale for longer timespans. Even seemingly partially-observable multi-room games like Montezuma's Revenge are made near-fully observable by displaying the player's score and inventory (Burda et al., 2022).

OpenAI Gym (Brockman et al., 2016) came after ALE, implementing classic fully observable RL benchmarks like CartPole and MountainCar. Their Gym API found use in many other environments, including our proposed benchmark.

Cobbe et al. (2020) find that randomly generated environments are critical to training general agents, showing policies will overfit to specific levels otherwise. They propose ProcGen: 16 procedurally generated environments with pixel-space observations. Most environments are fully or near-fully observable, although a few environments provide a partially observable mode, effectively turning them into 2D area coverage (navigation) tasks. ProcGen motivates POPGym's use of random level generation.

2.2 Partially Observable Benchmarks

When enumerating partially observable benchmarks, we find many are based on 3D first-person navigation. DeepMind Lab (Beattie et al., 2016) (DMLab) is a 3D first-person view navigation simulator based on the Quake 3 physics engine. It implements various tasks such as collecting fruits, maze exploration, and laser tag. VizDoom (Kempka et al., 2016) is another 3D navigation simulator based on the PC game Doom. It gives the agent weapons and adds computer-control characters that can shoot at the player. Miniworld (Chevalier-Boisvert, 2018) is a third 3D first-person view navigation simulator that is easier to install than DMLab or VizDoom. MiniGrid (Chevalier-Boisvert et al., 2018) and GridVerse (Baisero & Katt, 2021) are 2D navigation simulators with a first-person view. Unlike the previously mentioned 3D simulators, agents converge on gridworld environments much faster due to the smaller observation space. This makes it a popular benchmark for memory models.

There are few POMDP libraries that provide tasks beyond navigation. Behaviour suite (BSuite) evaluates agents on a variety of axes, one of which is memory (Osband et al., 2020), but they only provide two POMDPs. Similar to our benchmark, (Zheng & Tellex, 2020) provide classic POMDPs with low-dimensional observation spaces. But their tasks are solvable without neural networks, and are not difficult enough for modern deep RL. Ni et al. (2022) provide 21 environments, 8 of which are regular POMDPs. The regular POMDPs are all classic control tasks with held out position or velocity measurements, which we show might not be sufficiently diverse to properly benchmark memory.

We briefly mention the Starcraft (Samvelyan et al., 2019) and VMAS (Bettini et al., 2022) benchmarks because multi-agent environments are intrinsically partially observable, but we focus specifically on single-agent POMDPs.

2.3 Shortcomings of Current Benchmarks

Popular fully observable benchmarks use pixel-based observation spaces, adding a layer of complexity which takes an order of magnitude longer to train when compared against state-based observation counterparts (Seita, 2020). In fully observable environments, visually pleasing results are worth a few extra hours training. This dogma persists into partial observability, where environments often take 10x longer to converge than their fully observable counterparts. Popular benchmarks using 3D graphics take hundreds of billions of timesteps (Parisotto et al., 2020) and multiple weeks (Morad et al., 2021) on a GPU to train a single policy to convergence. Until sample efficiency in partially observable RL improves, we must forgo pixel-based observations or continue to struggle with reproducability.

Many partially observable tasks with pixel-based observation spaces are based on some form of navigation (Ramani, 2019). Although navigation can be a partially observable task, wall following behavior in perfect mazes guarantees complete area coverage without the need for memory. When mazes are imperfect (i.e. contain cycles), deterministic wall following can get stuck in infinite loops. However, RL policies often have some amount of stochasticity that can break out of these loops. Kadian et al. (2020) and Morad et al. (2021) show that memory-free navigation agents learn wall following behaviors that are surprisingly effective in imperfect real-world mazes. We confirm these findings with our experiments, showing that memory-free agents are competitive with memory-endowed agents in certain navigation benchmarks.

All other (imperfect) mazes can be fully explored by storing no more than two past locations (observations) in memory (Blum & Kozen, 1978). Navigation-based tasks like area coverage, moving to a coordinate, or searching for items can be reduced to the maze exploration task. We do not claim

that navigation tasks are easy, but rather that it is important to have a variety of tasks to ensure we evaluate all facets of memory, such as *memory capacity*, that navigation tasks might miss.

2.4 Existing Memory Baselines

The state of memory models in RL is even more bleak than the benchmarks. Most libraries provide frame stacking and a single type of RNN. OpenAI Baselines (Dhariwal et al., 2017), Stable-Baselines3 (Raffin et al., 2021), and CleanRL (Huang et al., 2021) provide implementations of PPO with frame stacking and an LSTM. Ray RLlib (Liang et al., 2018) currently implements frame stacking, LSTM, and a transformer for some algorithms. Ni et al. (2022) implement LSTM, GRUs, and two model-based memory models. Yang & Nguyen (2021) provides recurrent versions of the DDPG, TD3, and SAC RL algorithms, which utilize GRUs and LSTM. Zheng & Tellex (2020) implement multiple classical POMDP solvers, but these do not use neural networks, preventing their application to more complex tasks. There is currently no go-to library for users who want to compare or apply non-standard memory models.

2.5 A Brief Summary on Memory

When designing a library of memory models, it is important to select competitive models. Ni et al. (2022) show that sequence to sequence models from SL are competitive or better than RL-specific memory methods while being more straightforward to implement, so we focus specifically on sequence to sequence models (called memory throughout the paper). Although a strict categorization of memory is elusive, most methods are some mix of RNNs, attention, and convolution.

RNNs (Elman, 1990) take an input and hidden state, feed them through a network, and produce a corresponding output and hidden state. RNNs depend on the previous state and must be executed sequentially, resulting in slow training but fast inference when compared with other methods.

Attention-based methods (Vaswani et al., 2017) have supplanted RNNs in many applications of SL, but traditional transformers have quadratically-scaling memory requirements, preventing them from running on long episodes in RL. Recent linear attention formulations (Schlag et al., 2021; Katharopoulos et al., 2020) claim to produce transformer-level performance in linear time and space, potentially enabling widespread use of attention in RL.

Like attention, convolutional methods are computationally efficient (Bai et al., 2018), lending themselves well to RL. They are less common than recurrent or attention-based methods in SL, and there is little literature on their use in RL.

3 POPGym Environments

We tag our proposed environments as *diagnostic*, *control*, *noisy*, *game*, and *navigation*. Each tag is designed to represent a different class of POMDP, and each environment has at least three distinct difficulty settings, creating the most diverse POMDP benchmark thus far.

Diagnostic environments serve as memory introspection and debugging tools. They allow the user to probe model capability with respect to the duration of memories, forgetting, and compression and recall.

Control environments are control RL environments made partially observable by removing part of the observation. Solving these tasks only requires short-term memory.

Noisy environments require the memory model to estimate the true underlying state by computing an expectation over many observations. These are especially useful for real-world robotics tasks.

Game environments provide a more natural and thorough evaluation of memory through card and board games. They stress test memory capacity, duration, and higher-level reasoning.

Navigation environments are common in other benchmarks, and we include a few to ensure our benchmark is comprehensive. More than anything, our navigation environments examine how memory fares over very long sequences.

3.1 Environment Descriptions

1. **Repeat First (Diagnostic):** At the first timestep, the agent receives one of four values and a remember indicator. Then it randomly receives one of the four values at each successive timestep without the remember indicator. The agent receives a reward for outputting (remembering) the first value.
2. **Repeat Previous (Diagnostic):** Like Repeat Fist, observations contain one of four values. The agent is rewarded for outputting the observation from some constant k timesteps ago, i.e. observation o_{t-k} at time t.
3. **Autoencode (Diagnostic):** During the WATCH phase, a deck of cards is shuffled and played in sequence to the agent with the watch indicator set. The watch indicator is unset at the last card in the sequence, where the agent must then output the sequence of cards in order. This tests whether the agent can encode a series of observations into a latent state, then decode the latent state one observation at a time.
4. **Stateless Cartpole (Control):** The cartpole environment from Barto et al. (1983), but with the angular and linear positions removed from the observation. The agent must integrate to compute positions from velocity.
5. **Stateless Pendulum (Control):** The swing-up pendulum (Doya, 1995), with the angular position information removed.
6. **Noisy Stateless Cartpole (Control, Noisy):** Stateless Cartpole (Env. 4) with Gaussian noise.
7. **Noisy Stateless Pendulum (Control, Noisy):** Stateless Pendulum (Env. 5) with Gaussian noise.
8. **Multiarmed Bandit (Noisy, Diagnostic):** The multiarmed bandit problem (Slivkins & others, 2019) posed as an episodic task. Every episode, bandits are randomly initialized. Over the episode, the player must trade off exploration and exploitation, remembering which bandits pay best. Each bandit has some probability of paying out a positive reward, otherwise paying out a negative reward. Note that unlike the traditional multiarmed bandit task where the bandits are fixed once initialized, these bandits reset every episode, forcing the agent to learn a policy that can adapt between episodes.
9. **Higher Lower (Game, Noisy):** Based on the card game higher-lower, the agent must guess if the next card is of higher or lower rank than the previous card. The next card is then flipped face-up and becomes the previous card. Using memory, the agent can utilize card counting strategies to predict the expected value of the next card, improving the return.
10. **Concentration (Game):** A deck of cards is shuffled and spread out face down. The player flips two cards at a time face up, receiving a reward if the flipped cards match. The agent must remember the value and position of previously flipped cards to improve the rate of successful matches.
11. **Battleship (Game):** A partially observable version of Battleship, where the agent has no access to the board and must derive its own internal representation. Observations contain either HIT or MISS and the position of the last salvo fired. The player receives a positive reward for striking a ship, zero reward for hitting water, and negative reward for firing on a specific tile more than once.
12. **Mine Sweeper (Game):** The computer game Mine Sweeper, but like our Battleship implementation, the agent does not have access to the board. Each observation contains the position and number of adjacent mines to the last square "clicked" by the agent. Clicking on a square mine will end the game and produce a negative reward. The agent must remember where it has already searched and must integrate information from nearby tiles to narrow down the location of mines. Once the agent has selected all non-mined squares, the game ends.
13. **Labyrinth Explore (Navigation):** The player is placed in a discrete, 2D procedurally-generated maze, receiving a reward for each previously unreached tile it reaches. The player can only observe adjacent tiles and the previous action. The agent also receives a small negative reward at each timestep, encouraging the agent to reach all squares quickly and end the episode.
14. **Labyrinth Escape (Navigation):** The player must escape the procedurally-generated maze, using the same observation space as Labyrinth Explore. This is a sparse reward setting, where the player receives a positive reward only after solving the maze.

Table 1: Frames per second (FPS) of our environments, computed on the Google Colab free tier and a Macbook Air (2020) laptop.

Environment	Colab FPS	Laptop FPS	Environment	Colab FPS	Laptop FPS
Repeat First	23,895	155,201	Multiarmed Bandit	48,751	469,325
Repeat Previous	50,349	136,392	Battleship	117,158	235,402
Autoencode	121,756	251,997	Concentration	47,515	157,217
Stateless Cartpole	73,622	218,446	Higher Lower	24,312	76,903
Stateless Pendulum	8,168	26,358	Minesweeper	8,434	32,003
Noisy Stateless Cartpole	6,269	66,891	Labyrinth Escape	1,399	41,122
Noisy Stateless Pendulum	6,808	20,090	Labyrinth Explore	1,374	30,611

4 POPGym Baselines

Our memory model API relies on an abstract memory model class, only requiring users to implement `memory_forward` and `initial_state` methods. Our memory API builds on top of RLlib, exposing various algorithms, exploration methods, logging, distributed training, and more.

We collect well-known memory models from SL domains and wrap them in this API, enabling their use on RL tasks. We rewrite models where the existing implementation is slow, unreadable, not amenable to our API, or not written in Pytorch. Some of these sequence models have yet to be applied in the context of reinforcement learning.

1. **MLP:** An MLP that cannot remember anything. This serves to form a lower bound for memory performance, as well and ensuring memory models are actively using memory, rather than just leveraging their higher parameter counts.
2. **Positional MLP (PosMLP):** An MLP that can access the current episodic timestep. The current timestep is fed into the positional encoding from Vaswani et al. (2017), which is summed with the incoming features. PosMLP enables agents to learn time-dependent policies (those which evolve over the course of an episode) without explicitly using memory.
3. **Elman Networks:** The original RNN, from Elman (1990). Elman networks sum the recurrent state and input, passing the resulting vector through a linear layer and activation function to produce the next hidden state. Elman networks are not used much in SL nowadays due to vanishing and exploding gradients.
4. **Long Short-Term Memory (LSTM):** Hochreiter & Schmidhuber (1997) designed LSTM to address the vanishing and exploding gradient problems present in earlier RNNs like the Elman Network. LSTM utilizes a constant error carousel to handle longer dependencies and gating to ensure recurrent state stability during training. It has two recurrent states termed the hidden and cell states.
5. **Gated Recurrent Unit (GRU):** The GRU is a simplification of LSTM, using only a single recurrent state. The GRU appears to have similar performance to LSTM in many applications while using fewer parameters (Chung et al., 2014).
6. **Independently Recurrent Networks (IndRNN):** Stacking LSTM and GRU cells tends to provide few benefits when compared with traditional deep neural networks. IndRNNs update the recurrent state using elementwise connections rather than a dense layer, enabling much deeper RNNs and handling longer dependencies than the LSTM and GRU (Li et al., 2018). In our experiments, we utilize a 2-layer IndRNN.
7. **Differentiable Neural Computers (DNC):** Graves et al. (2016) introduce a new type of recurrent model using external memory. The DNC utilizes an RNN as a memory controller, reading and writing to external storage in a differentiable manner.
8. **Fast Autoregressive Transformers (FART):** Unlike the traditional attention matrix whose size scales with the number of inputs, FART computes a fixed-size attention matrix at each timestep, taking the cumulative elementwise sum over successive timesteps (Katharopoulos et al., 2020). FART maintains two recurrent states, one for the running attention matrix and one for a normalization term, which helps mitigate large values and exploding gradients as the attention increases grows over time. The original paper omits a positional encoding, but we find it necessary for our benchmark.

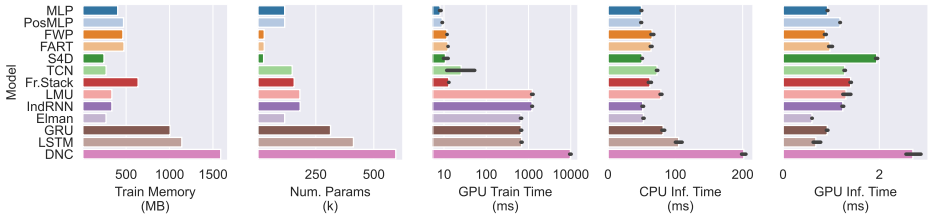


Figure 2: Performance characteristics for POPGym memory baselines on random inputs. We use a recurrent state size of 256, a batch size of 64, and a episode length of 1024. We compute CPU statistics on a 3GHz Xeon Gold and GPU statistics on a 2080Ti, reporting the mean and 95% confidence interval over 10 trials. Train times correspond to a full batch while inference times are per-element (i.e. the latency to compute a single action). Note that GPU Train Time has logarithmic scale.

9. **Fast Weight Programmers (FWP):** The theory behind FART and FWP is different, but the implementation is relatively similar. FWP also maintains a running cumulative sum. Unlike FART, FWP normalizes the key and query vectors to sum to one, requiring only a single recurrent state and keeping the attention matrix of reasonable scale (Schlag et al., 2021). Unlike the original paper, we add a positional encoding to FWP.

10. **Frame Stacking (Fr.Stack):** Mnih et al. (2015) implemented frame stacking to solve Atari games. Frame stacking is the concatenation of k observations along the feature dimension. Frame stacking is not strictly convolutional, but is implemented similarly to other convolutional methods. Frame stacking is known to work very well in RL, but the number of parameters scales with the receptive field, preventing it from learning long-term dependencies.

11. **Temporal Convolutional Networks (TCN):** TCNs slide 1D convolutional filters over the temporal dimension. On long sequences, they are faster and use less memory than RNNs. TCNs avoid the vanishing gradient problem present in RNNs because the gradient feeds through each sequence element individually, rather than propagating through the entire sequence (Bai et al., 2018).

12. **Legendre Memory Units (LMU):** LMUs are a mixture of convolution and RNNs. They apply Legendre polynomials across a sliding temporal window, feeding the results into an RNN hidden state (Voelker et al., 2019). LMUs can handle temporal dependencies spanning up to 100K timesteps.

13. **Diagonal State Space Models (S4D):** S4D treats memory as a controls problem. It learns a linear time invariant (LTI) state space model for the recurrent state. S4D applies a Vandermonde matrix to the sequence of inputs, which we can represent using either convolution or a recurrence. Computing the result convolutionally makes it very fast. In SL, S4D was able to solve the challenging 16,000 timestep Path-X task, demonstrating significant capacity for long-term dependencies (Gu et al., 2022).

5 Experiments

Our memory framework hooks into RLlib, providing integration with IMPALA, DQN, and countless other algorithms. Due to computational constraints, we only execute our study on Proximal Policy Optimization (PPO) (Schulman et al., 2017) and exclude the computationally-expensive DNC from experiments. We tend to use conservative hyperparameters to aid in reproducability – this entails large batch sizes, low learning rates, and many minibatch passes over every epoch. We run three trials of each model over three difficulties for each environment, resulting in over 1500 trials. We provide a summary over models in Fig. 3 and a more detailed summary in Tab. 2. Fig. 2 reports model throughput and Tab. 1 provides environment throughput. The full experimental parameters and more detailed results are available in Appendix A and Appendix B respectively.

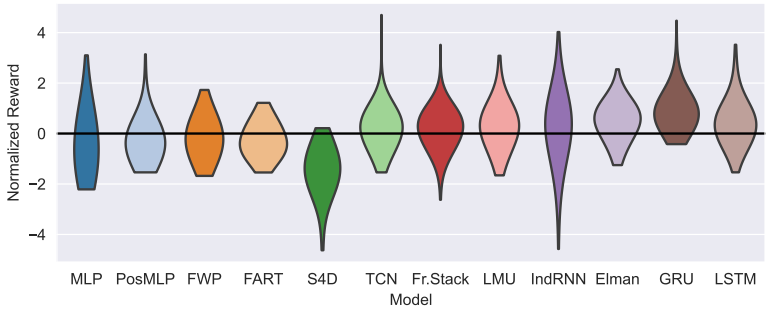


Figure 3: A violin plot of reward, aggregated over all environments. We normalize the rewards by subtracting by the mean and dividing by the standard deviation on a per-environment basis. The black line denotes the "mean" model across all environments. We use a Gaussian kernel with a bandwidth of 1.

6 Discussion

In general, RNNs seem to outperform convolutional and attention-based approaches with the GRU as the most robust and well-rounded model (Fig. 3), with the caveat that they are slower and use more parameters than attentional and pure convolutional methods (Fig. 2). The more recent methods (e.g. IndRNN) tend to excel in some tasks but fail in others. This highlights the need for diverse RL benchmarks and questions whether SL benchmarks provide enough indication for how memory will perform in RL. From our limited study, there is little correlation between memory performance on SL tasks (Gu et al., 2022; Voelker et al., 2019; Li et al., 2018) and RL tasks. SL's focus on long-term temporal dependencies and RL's focus on sample efficiency might be partially responsible for this gap. The longer the temporal window, the larger the search space for the RL optimization routines. Future work should investigate the cause of this disconnect, and whether any modifications could mitigate this SL to RL gap.

Elman networks outperformed LSTM and GRUs on shorter-horizon tasks while being conceptually simpler and using fewer parameters (Fig. 2). We noticed no exploding gradient problems, suggesting they could be applied to computationally-expensive environments with shorter-term temporal dependencies (e.g. robotic control).

The MLP was state of the art in the majority of navigation environments, reaching a nearly perfect score in 5/6 navigation tasks and again highlighting the importance of benchmarking on disparate tasks. This is in line with our hypothesis in Section 2.3. PosMLP likely did worse than MLP on navigation because for wall following, the positional encoding becomes a nuisance variable, greatly expanding the observation space. With episode length at 1024 timesteps, this increases the number of distinct observations for PosMLP by a factor of 1024.

Memory baselines had no trouble with the control tasks and their noisy counterparts, but still struggled with Multiarmed Bandit, highlighting for a third time the need for diverse environments. Classic control tasks with partially obscured observations are very common in literature (Ni et al., 2022). PosMLP nearly solved both the noisy and noise-free variants of Stateless CartPole on the easy difficulty, suggesting that just hiding portions of the observation in control tasks might not be a sufficient memory benchmark moving forward. The hard version of Stateless Cartpole uses longer episodes (up to 600 timesteps), which seems significantly reduce PosMLP performance, while the addition of noise seemed to make little difference. We propose that future benchmarks make their controls tasks sufficiently long ($\sim$ 600 timesteps) to stress-test memory.

All methods had trouble with game environments, suggesting some limitations are not due to memory models, but the underlying RL algorithm (PPO in our case). For example, memory models performed on-par with the MLP for Concentration on easier difficulties (where the player must match card colors), suggesting the memory models were not remembering anything. On the hard difficulty (where the player must match card ranks), some memory models pulled ahead of MLP, implying that they are remembering previous cards. We see similar low-variance results on Battle-

Table 2: Maximum reward for each environment, denoted by environment number and split by difficulty. We report the mean over three trials. A value of "-" denotes the model crashed consistently due to exploding weights. Environments are abbreviated by their index in Sec. 3.1: E1: Repeat First, E2: Repeat Previous, E3: Autoencode, E4: Stateless Cartpole, E5: Stateless Pendulum, E6: Noisy Stateless Cartpole, E7: Noisy Stateless Pendulum, E8: Multiarmed Bandit, E9: Higher Lower, E10: Concentration; E11: Battleship, E12: Mine Sweeper, E13: Labyrinth Explore, E14: Labyrinth Escape

Model	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
Easy														
MLP	-0.43	-0.48	-0.47	0.27	0.38	0.27	0.37	0.42	0.50	0.43	-0.35	0.26	**0.97**	**0.96**
PosMLP	-0.45	-0.48	-0.46	0.76	0.48	0.74	0.46	0.29	0.50	0.44	**-0.10**	0.44	0.87	0.54
FWP	0.98	0.32	-0.32	0.93	0.52	0.94	0.51	0.52	0.52	0.43	-0.40	0.47	-0.15	-0.08
FART	**1.00**	-0.16	-0.46	0.98	0.70	0.98	0.59	0.45	0.52	0.43	-0.41	0.55	0.10	-0.03
S4D	-0.45	-0.49	-0.49	0.18	-	0.18	-	0.05	0.32	0.43	-0.44	-0.12	-0.31	-0.20
TCN	-0.35	**1.00**	-0.45	**1.00**	**0.91**	**1.00**	0.91	0.47	0.50	0.43	-0.17	0.66	0.93	0.44
Fr.Stack	-0.35	**1.00**	-0.43	**1.00**	**0.91**	**1.00**	0.91	0.44	0.50	0.43	-0.41	0.62	0.94	0.50
LMU	**1.00**	**1.00**	-0.36	**1.00**	0.82	**1.00**	0.91	0.21	0.50	0.44	-0.40	0.81	0.81	0.61
IndRNN	**1.00**	0.88	-0.40	**1.00**	0.44	**1.00**	0.80	0.59	**0.53**	0.41	-0.31	0.75	0.94	0.19
Elman	**1.00**	**1.00**	-0.37	**1.00**	**0.91**	**1.00**	**0.93**	**0.74**	0.52	**0.44**	-0.33	0.76	0.96	**0.96**
GRU	**1.00**	**1.00**	-0.34	**1.00**	**0.91**	**1.00**	0.91	0.62	0.52	0.43	-0.36	0.85	0.96	**0.96**
LSTM	**1.00**	**1.00**	-0.31	**1.00**	0.87	**1.00**	0.91	0.68	0.52	0.43	-0.38	**0.86**	0.65	0.91
Medium														
MLP	-0.43	-0.49	-0.48	0.13	0.38	0.26	0.37	0.37	0.50	0.47	-0.37	-0.13	**0.92**	**0.24**
PosMLP	-0.47	-0.49	-0.48	0.38	0.49	0.74	0.47	0.20	0.50	0.47	-0.38	-0.09	0.29	-0.33
FWP	0.90	-0.27	-0.45	0.57	0.55	0.97	0.52	0.39	0.50	0.47	-0.38	-0.16	-0.48	-0.64
FART	0.94	-0.48	-0.49	0.61	0.63	0.98	0.62	0.30	**0.51**	0.47	-0.40	-0.17	-0.14	-0.65
S4D	-0.34	-0.49	-0.49	0.08	-	0.16	-	0.12	0.30	-	-0.40	-0.40	-0.56	-0.70
TCN	-0.46	-0.48	-0.47	0.99	0.88	**1.00**	0.91	0.33	0.50	0.47	-0.39	-0.11	0.55	-0.46
Fr.Stack	-0.45	-0.48	-0.48	**1.00**	0.88	**1.00**	0.91	0.38	0.50	0.47	-0.39	-0.23	0.75	-0.43
LMU	0.03	**0.82**	-0.47	**1.00**	0.88	**1.00**	0.91	0.04	0.47	0.47	-0.40	-0.09	0.24	-0.55
IndRNN	**0.99**	-0.38	-0.47	**1.00**	0.47	**1.00**	0.84	**0.70**	**0.51**	0.39	**-0.34**	0.02	0.74	-0.71
Elman	-0.46	-0.45	-0.47	**1.00**	0.90	**1.00**	**0.93**	0.42	0.50	0.46	-0.39	0.07	0.89	-0.01
GRU	**1.00**	-0.10	-0.45	**1.00**	0.89	**1.00**	0.91	0.59	**0.51**	0.46	-0.37	**0.08**	0.89	0.22
LSTM	0.29	-0.25	**-0.41**	0.99	**0.90**	**1.00**	0.91	0.41	**0.51**	0.47	-0.39	-0.13	0.05	-0.58
Hard														
MLP	-0.39	-0.49	-0.49	0.09	0.39	0.27	0.38	0.30	**0.51**	-0.82	-0.38	-0.31	**0.86**	-0.79
PosMLP	-0.45	-0.49	-0.48	0.25	0.51	0.74	0.47	0.19	0.50	-0.82	-0.38	-0.30	-0.08	**-0.73**
FWP	0.62	-0.44	-0.47	0.46	0.59	0.96	0.51	0.19	0.50	-0.82	-0.38	-0.35	-0.61	-0.84
FART	-0.38	-0.48	-0.48	0.39	0.70	0.98	0.69	0.25	**0.51**	-0.82	-0.38	-0.42	-0.43	-0.83
S4D	-0.33	-0.49	-0.49	0.05	-	0.17	-	0.11	0.24	-0.82	-0.39	-0.45	-0.73	-0.88
TCN	-0.46	-0.49	-0.48	0.92	**0.83**	**1.00**	**0.91**	0.14	0.50	**-0.71**	-0.38	-0.39	0.01	-0.79
Fr.Stack	-0.42	-0.49	-0.48	0.98	**0.83**	**1.00**	**0.91**	0.29	0.50	-0.74	-0.38	-0.39	0.38	-0.80
LMU	-0.47	**0.25**	-0.48	**1.00**	0.82	**1.00**	**0.91**	0.04	0.47	-0.82	-0.38	-0.30	-0.01	-0.82
IndRNN	0.36	-0.45	-0.49	**1.00**	0.74	**1.00**	0.57	**0.56**	**0.51**	-0.78	**-0.36**	-0.23	0.43	-0.93
Elman	-0.47	-0.47	-0.48	**1.00**	0.82	**1.00**	**0.91**	0.04	0.50	-0.82	-0.38	-0.19	0.76	-0.80
GRU	**1.00**	-0.21	-0.46	**1.00**	**0.83**	**1.00**	**0.91**	0.54	0.50	-0.82	-0.38	**-0.17**	0.79	**-0.73**
LSTM	-0.33	-0.37	**-0.45**	**1.00**	0.82	**1.00**	**0.91**	0.39	0.50	-0.74	-0.38	-0.38	-0.16	-0.81

ship. It is unclear why memory models fail to solve Battleship and Concentration, while doing well on Mine Sweeper. Memory models do succeed in Higher Lower, but the benefit from card counting is relatively small, translating to only small increases in reward.

7 Conclusion

We presented the POPGym benchmark, a collection of POMDPs and memory baselines designed to standardize partially observable RL. We discovered that the GRU is the best general-purpose memory model, beating out linear transformers and more modern RNNs across most tasks. Furthermore, we noticed a significant disconnect between the performance of memory models in supervised learning and in reinforcement learning. The importance of diversity in benchmarks, particularly for POMDPs, cannot be overstated.

REFERENCES

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Andrea Baisero and Sammie Katt. gym-gridverse: Gridworld domains for fully and partially observable reinforcement learning, 2021. URL https://github.com/abaisero/gym-gridverse. Publication Title: GitHub repository.

Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983. Publisher: IEEE.

Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. DeepMind Lab. Technical Report arXiv:1612.03801, arXiv, December 2016. URL http://arxiv.org/abs/1612.03801. arXiv:1612.03801 [cs] type: article.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, 47:253–279, June 2013. ISSN 1076-9757. doi: 10.1613/jair.3912. URL https://www.jair.org/index.php/jair/article/view/10819.

Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. VMAS: A Vectorized Multi-Agent Simulator for Collective Robot Learning. *The 16th International Symposium on Distributed Autonomous Robotic Systems*, 2022. Publisher: Springer.

Manuel Blum and Dexter Kozen. On the power of the compass (or, why mazes are easier to search than graphs). In *19th Annual Symposium on Foundations of Computer Science (sfcs 1978)*, pp. 132–142, October 1978. doi: 10.1109/SFCS.1978.30. ISSN: 0272-5428.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym, 2016. _eprint: arXiv:1606.01540.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. February 2022. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Anthony R Cassandra. A survey of POMDP applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*, volume 1724, 1998.

Maxime Chevalier-Boisvert. MiniWorld: Minimalistic 3D Environment for RL & Robotics Research, 2018. URL https://github.com/maximecb/gym-miniworld. Publication Title: GitHub repository.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic Gridworld Environment for OpenAI Gym, 2018. URL https://github.com/maximecb/gym-minigrid. Publication Title: GitHub repository.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging Procedural Generation to Benchmark Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2048–2056. PMLR, November 2020. URL https://proceedings.mlr.press/v119/cobbe20a.html. ISSN: 2640-3498.

Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. OpenAI Baselines, 2017. URL https://github.com/openai/baselines. Publication Title: GitHub repository.

Kenji Doya. Temporal Difference Learning in Continuous Time and Space. In D. Touretzky, M. C. Mozer, and M. Hasselmo (eds.), *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper/1995/file/1e1d184167ca7676cf665225e236a3d2-Paper.pdf.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. Publisher: Wiley Online Library.

Dibya Ghosh, Jad Rahme, Aviral Kumar, Amy Zhang, Ryan P Adams, and Sergey Levine. Why Generalization in RL is Difficult: Epistemic POMDPs and Implicit Partial Observability. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25502–25515. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d5ff135377d39f1de7372c95c74dd962-Abstract.html.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, and others. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016. Publisher: Nature Publishing Group.

Albert Gu, Karan Goel, and Christopher Re. Efficiently Modeling Long Sequences with Structured State Spaces. March 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pp. 3207–3214, New Orleans, Louisiana, USA, February 2018. AAAI Press. ISBN 978-1-57735-800-8.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735. Place: Cambridge, MA, USA Publisher: MIT Press.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, and Jeff Braga. CleanRL: High-quality Single-file Implementations of Deep Reinforcement Learning Algorithms. 2021. _eprint: 2111.08819.

Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance? *IEEE Robotics and Automation Letters*, 5(4):6670–6677, October 2020. ISSN 2377-3766. doi: 10.1109/LRA.2020.3013848. Conference Name: IEEE Robotics and Automation Letters.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998. Publisher: Elsevier.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5156–5165. PMLR, November 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html. ISSN: 2640-3498.

Micha\l Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. ViZDoom: A Doom-based AI Research Platform for Visual Reinforcement Learning. In *IEEE Conference on Computational Intelligence and Games*, pp. 341–348, Santorini, Greece, September 2016. IEEE. URL http://arxiv.org/abs/1605.02097.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791. Conference Name: Proceedings of the IEEE.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems, November 2020. URL http://arxiv.org/abs/2005.01643. arXiv:2005.01643 [cs, stat].

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5457–5466, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00572. URL https://ieeexplore.ieee.org/document/8578670/.

Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. RLlib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning*, pp. 3053–3062. PMLR, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, and others. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015. Publisher: Nature Publishing Group.

Steven D. Morad, Roberto Mecca, Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. Embodied Visual Navigation With Automatic Curriculum Learning in Real Environments. *IEEE Robotics and Automation Letters*, 6(2):683–690, April 2021. ISSN 2377-3766. doi: 10.1109/LRA.2020.3048662. Conference Name: IEEE Robotics and Automation Letters.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent Model-Free RL Can Be a Strong Baseline for Many POMDPs. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16691–16723. PMLR, June 2022. URL https://proceedings.mlr.press/v162/ni22a.html. ISSN: 2640-3498.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvári, Satinder Singh, Benjamin Van Roy, Richard Sutton, David Silver, and Hado van Hasselt. Behaviour Suite for Reinforcement Learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygf-kSYwH.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphaël Lopez Kaufman, Aidan Clark, Seb Noury, Matthew Botvinick, Nicolas Heess, and Raia Hadsell. Stabilizing Transformers for Reinforcement Learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7487–7498. PMLR, July 2020. URL https://proceedings.mlr.press/v119/parisotto20a.html.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Dhruv Ramani. A Short Survey On Memory Based Reinforcement Learning. Technical Report arXiv:1904.06736, arXiv, April 2019. URL http://arxiv.org/abs/1904.06736. arXiv:1904.06736 [cs] type: article.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear Transformers Are Secretly Fast Weight Programmers. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9355–9366. PMLR, July 2021. URL https://proceedings.mlr.press/v139/schlag21a.html. ISSN: 2640-3498.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Daniel Seita. Can RL From Pixels be as Efficient as RL From State?, July 2020. URL http://bair.berkeley.edu/blog/2020/07/19/curl-rad/.

Aleksandrs Slivkins and others. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019. Publisher: Now Publishers, Inc.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Aaron Voelker, Ivana Kajić, and Chris Eliasmith. Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper/2019/hash/952285b9b7e7a1be5aa7849f32ffff05-Abstract.html.

Zhihan Yang and Hai Nguyen. Recurrent Off-policy Baselines for Memory-based Continuous Control. Technical Report arXiv:2110.12628, arXiv, October 2021. URL http://arxiv.org/abs/2110.12628. arXiv:2110.12628 [cs] type: article.

Kaiyu Zheng and Stefanie Tellex. pomdp_py: A Framework to Build and Solve POMDP Problems. In *ICAPS 2020 Workshop on Planning and Robotics (PlanRob)*, 2020. URL https://icaps20subpages.icaps-conference.org/wp-content/uploads/2020/10/14-PlanRob_2020_paper_3.pdf.

Table 3: PPO hyperparameters used in all of our experiments.

HParam	Value
Decay factor γ	0.99
Value fn. loss coef.	1.0
Entropy loss coef.	0.0
Learning rate	5e-5
Num. SGD iters	30
Batch size	65536
Minibatch size	8192
GAE λ	1.0
KL target	0.01
KL coefficient	0.2
PPO clipping	0.3
Value clipping	0.3
BPTT Truncation Length	1024
Maximum Episode Length	1024

A Experimental Parameters

Given the number of environments and models, it is not computationally feasible to optimize hyperparameters in a structured way. Through trial and error, we evaluated all models on the Repeat First and Repeat Previous environments and found suitable values that worked across all models. We then picked a more conservative estimate (larger batch size, lower learning rate) to promote monotonic improvement and prevent catastrophic forgetting across other tasks, at the expense of some sample efficiency.

There is no clear axis for a truly fair comparison between memory models – model throughput and parameter count vary drastically throughout the memory models. We decide to limit the amount of storage (i.e. recurrent state size) of each memory model to 256 dimensions, which is greater than the common values of 64 and 128 in literature (Ni et al., 2022). This results in lower parameter counts for models that produce the recurrent state using a tensor product (e.g. FWP, FART, S4D). We could make an exception for these models, allowing them to produce recurrent states of size $256^2 = 63356$ dimensions instead of $16^2 = 256$ dimensions to bring up the parameter count, but we believe this is unfair to recurrent models. In this case, recurrent models would need to forget and compress information over longer episodes, while outer product models could store every single observation in memory without any compression or forgetting. Storing everything is unlikely to scale to real-world applications where episodes could span hours, days, or run indefinitely.

A.1 General Model Hyperparameters

For all our memory experiments, we use the same outer model, just swapping out the inner memory model. The outer model first projects observations from the environment to 128 zero-mean variance-one dimensions. Here, the positional encoding is applied if the memory model requests it. The projection goes through a single linear layer and LeakyReLU activation of size 128, then feeds into the memory model. Output from the memory model is projected to 128 dimensions, then split into the actor and critic head. The actor and critic heads are two layer MLPs with LeakyReLU activation.

A.2 Model-Specific Hyperparameters

The Elman, GRU, LSTM, and LMU RNNs use a single cell. We use a 2-cell IndRNN as they claim to utilize deeper networks. The FART and FWP models use a single attention block. We use the attention-only formulations of FWP, rather than the combined attention and RNN variant. We utilize a temporal window of four for frame stacking and TCN. LMU utilizes a θ window size of 64 timesteps.

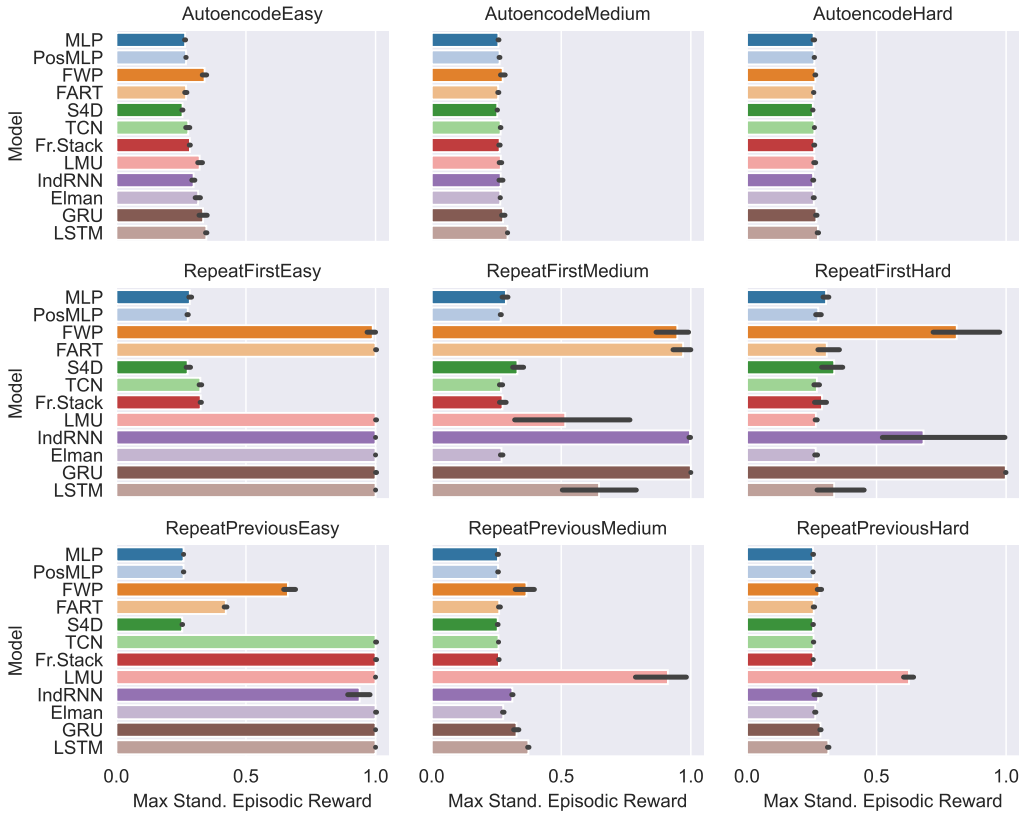


Figure 4: POPGym baselines on the diagnostic environments.

B Full Experimental Results

We report our results in three forms: bar charts denoting the maximum reward achieved over an episode, line plots showing cumulative maximum reward over training epochs, and tables of maximum reward achieved over each episode.

We run each trial 3 times, aggregating results using the mean over trials. The bar plots represent the mean and 95% confidence interval. For the bar charts, we standardize the reward between 0 and 1. In the line plots, the solid region refers to the mean and the shaded region to the 95% confidence interval. The tables report the mean and standard deviation.

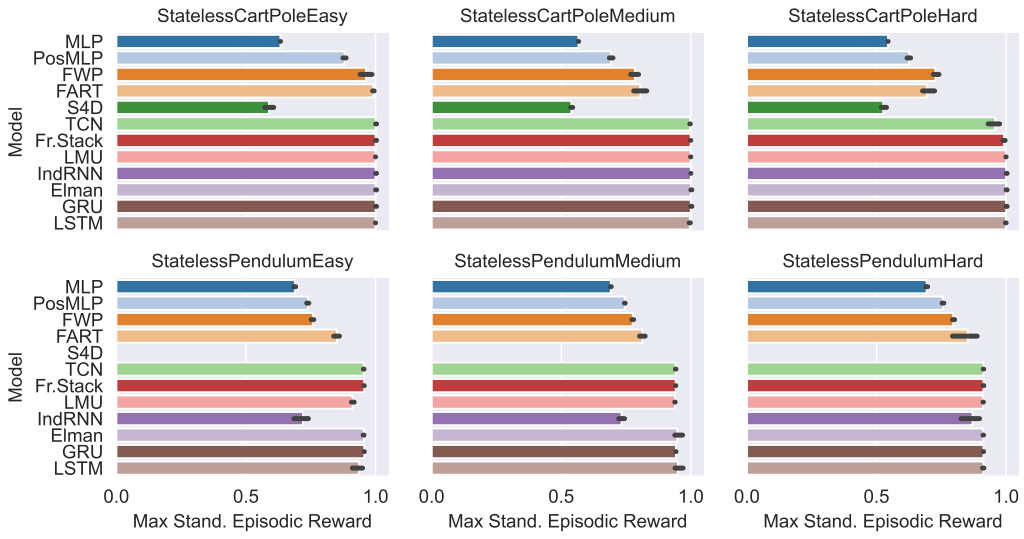


Figure 5: POPGym baselines on the control environments. We do not report S4D as model weights exploded, crashing all of the pendulum-based runs.

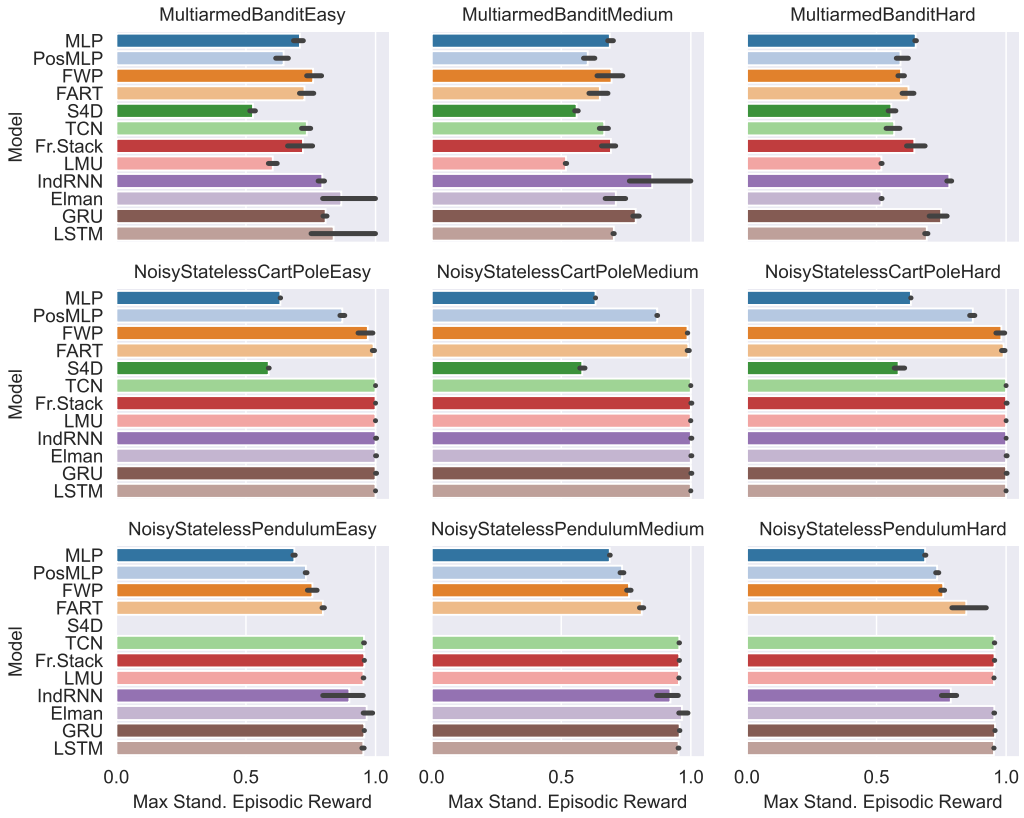


Figure 6: POPGym baselines on the noisy environments. We do not report S4D as model weights exploded, crashing all of the pendulum-based runs.

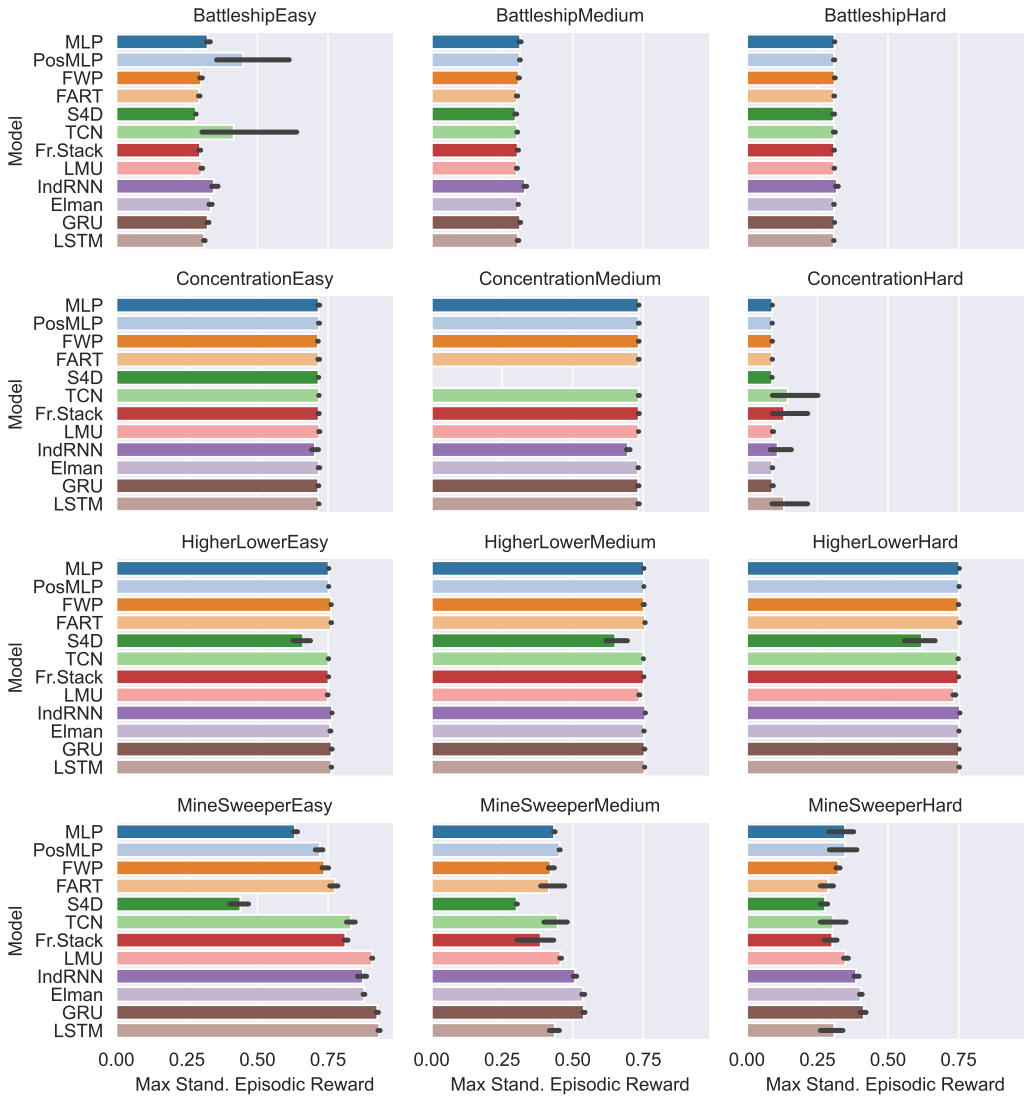


Figure 7: POPGym baselines on the game environments.

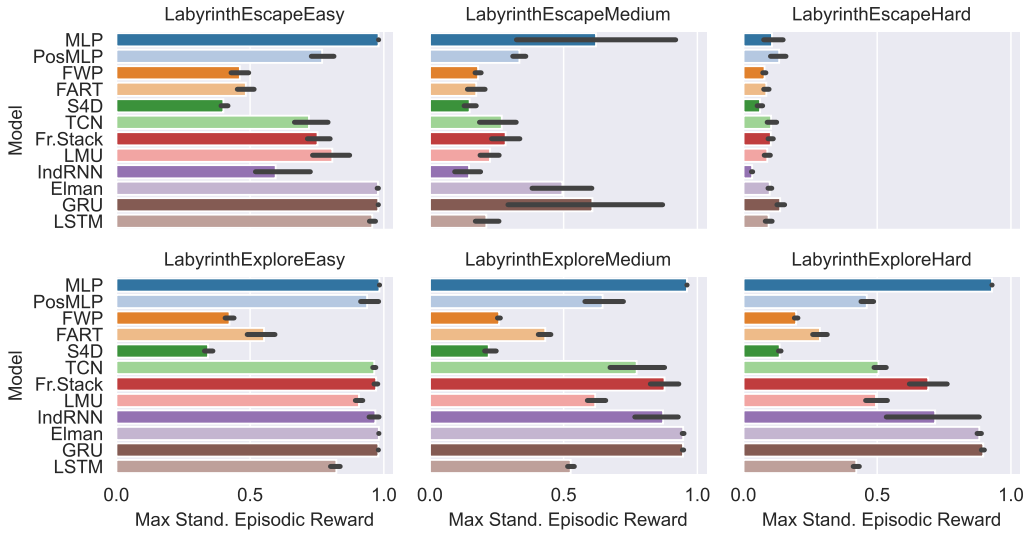


Figure 8: POPGym baselines on the navigation environments.

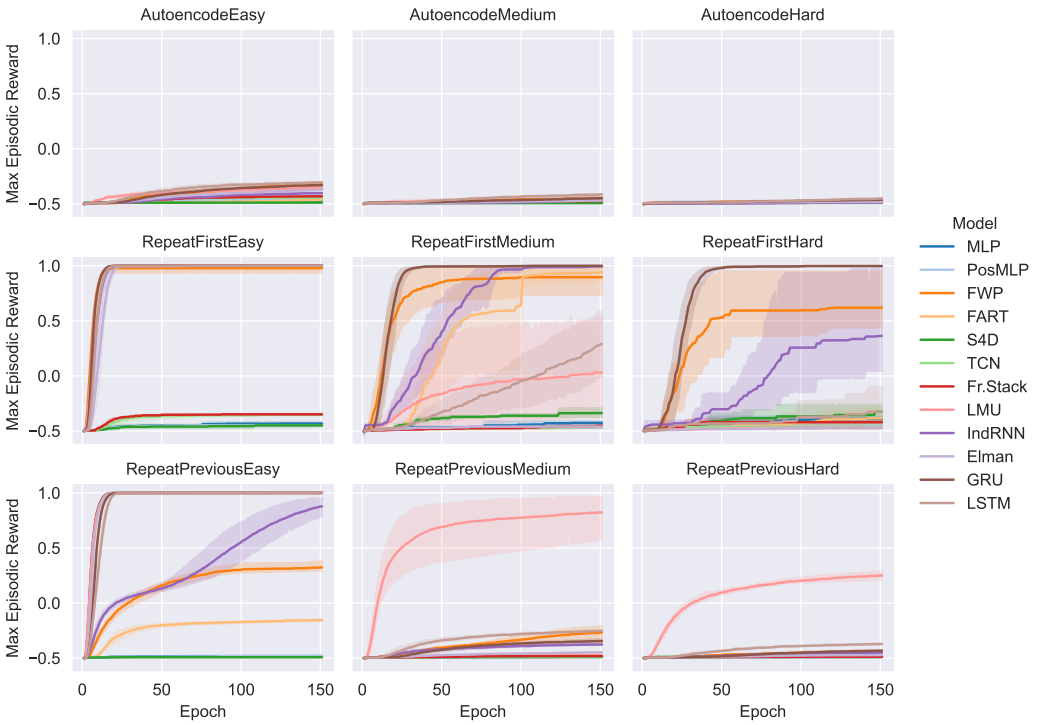


Figure 9: POPGym baselines on the diagnostic environments.

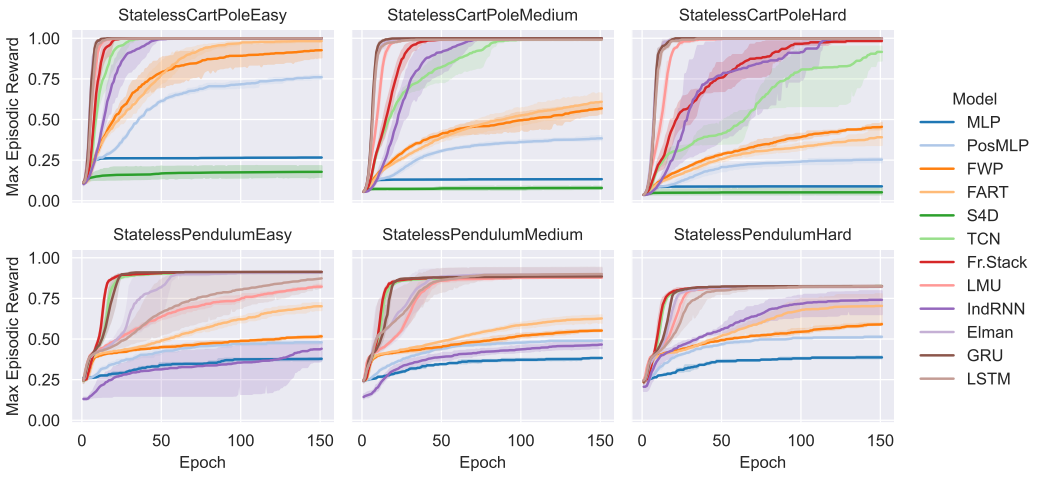


Figure 10: POPGym baselines on the control environments.

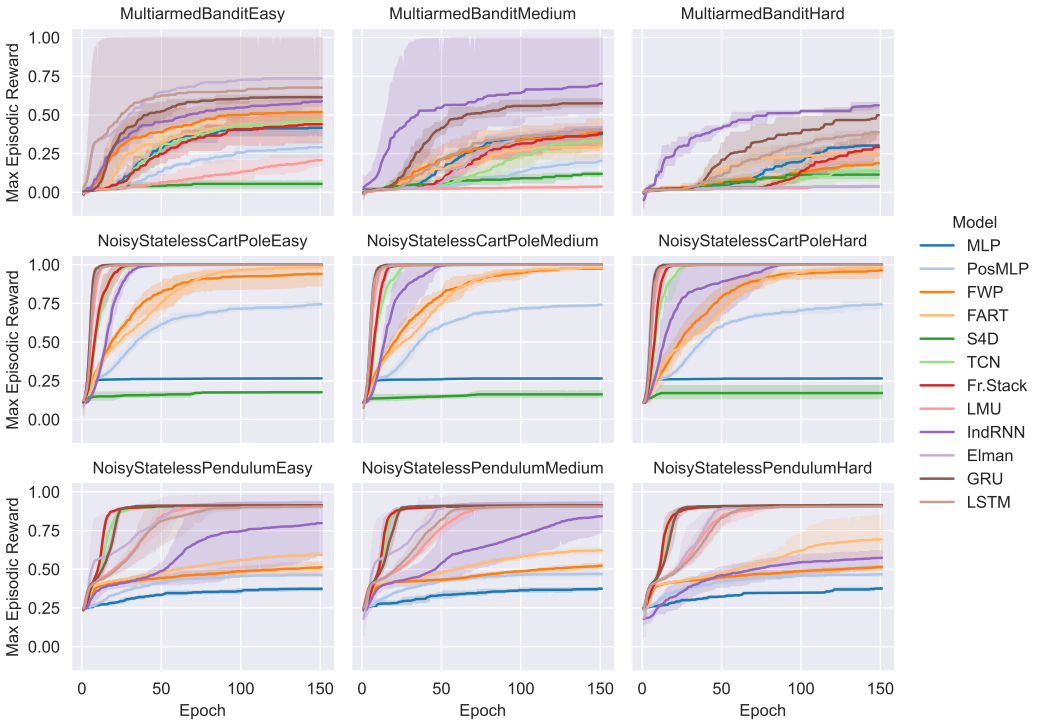


Figure 11: POPGym baselines on the noisy environments.

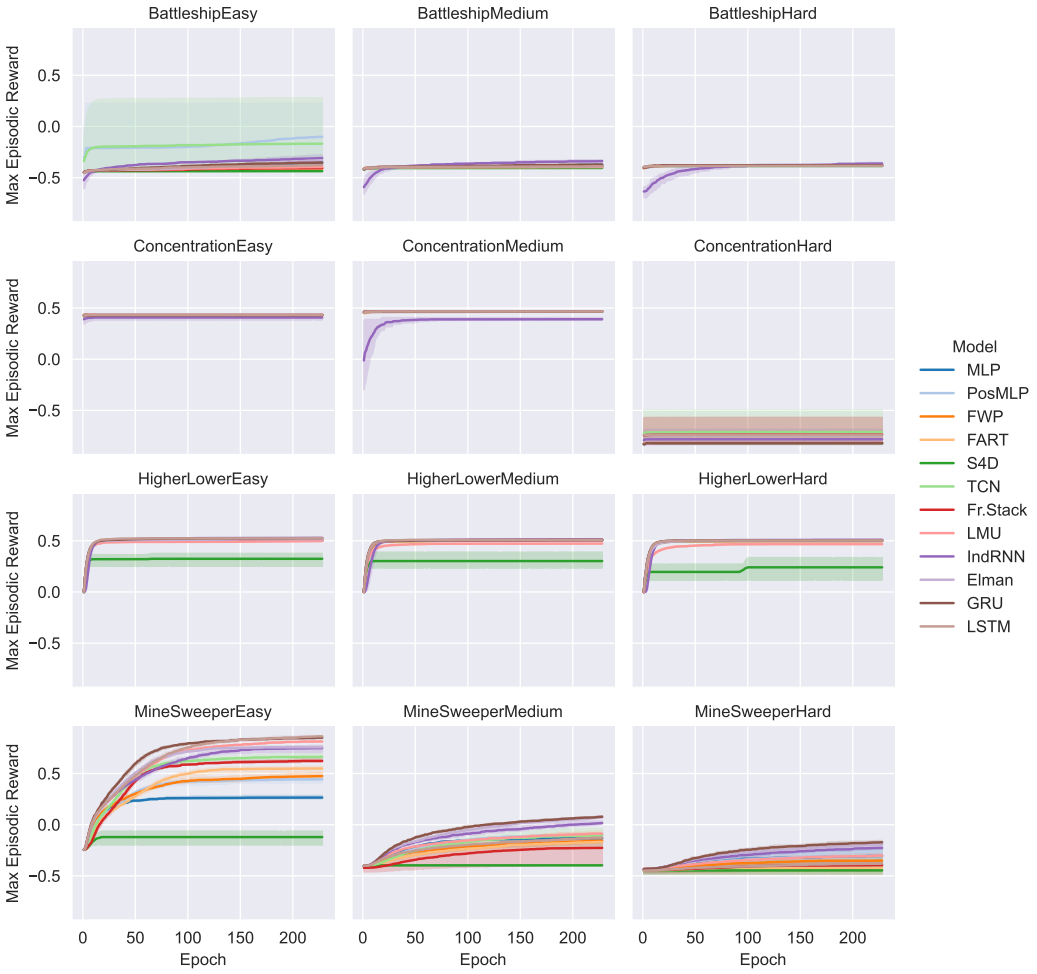


Figure 12: POPGym baselines on the game environments.

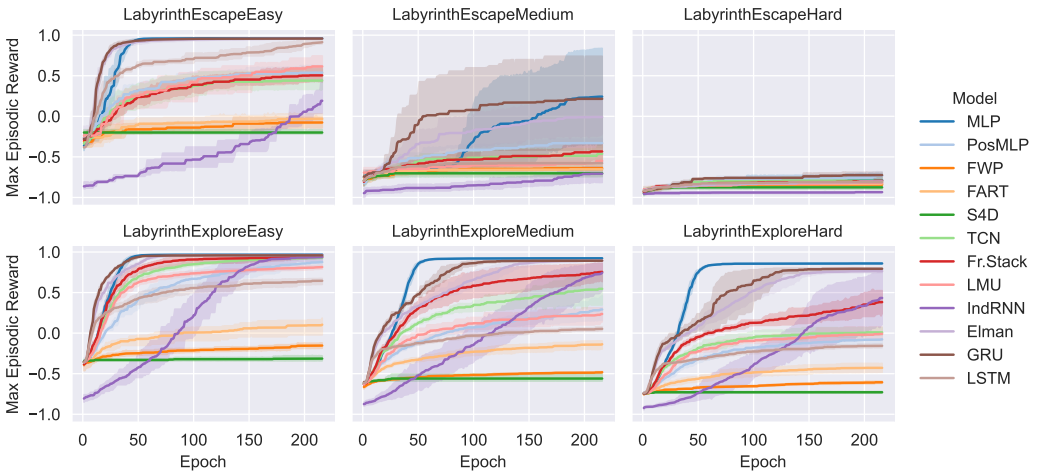


Figure 13: POPGym baselines on the navigation environments.

Env.	Model	μ	σ
AutoencodeEasy	Elman	-0.369	0.021
	FART	-0.462	0.009
	FWP	-0.319	0.017
	Fr.Stack	-0.432	0.005
	GRU	-0.331	0.043
	IndRNN	-0.404	0.013
	LMU	-0.357	0.019
	LSTM	**-0.306**	**0.006**
	MLP	-0.469	0.004
	PosMLP	-0.462	0.002
	S4D	-0.489	0.005
	TCN	-0.445	0.017
AutoencodeHard	Elman	-0.484	0.006
	FART	-0.484	0.002
	FWP	-0.472	0.004
	Fr.Stack	-0.481	0.004
	GRU	-0.464	0.008
	IndRNN	-0.489	0.004
	LMU	-0.478	0.008
	LSTM	**-0.452**	**0.008**
	MLP	-0.485	0.008
	PosMLP	-0.480	0.002
	S4D	-0.492	0.002
	TCN	-0.479	0.001
AutoencodeMedium	Elman	-0.470	0.001
	FART	-0.487	0.005
	FWP	-0.452	0.018
	Fr.Stack	-0.475	0.006
	GRU	-0.449	0.018
	IndRNN	-0.467	0.016
	LMU	-0.466	0.011
	LSTM	**-0.414**	**0.001**
	MLP	-0.485	0.004
	PosMLP	-0.475	0.004
	S4D	-0.494	0.002
	TCN	-0.467	0.003
RepeatFirstEasy	**Elman**	**1.000**	**0.000**
	FART	**1.000**	**0.000**
	FWP	0.978	0.038
	Fr.Stack	-0.349	0.005
	GRU	**1.000**	**0.000**
	IndRNN	**1.000**	**0.000**
	LMU	**1.000**	**0.000**
	LSTM	**1.000**	**0.000**
	MLP	-0.433	0.012
	PosMLP	-0.449	0.005
	S4D	-0.450	0.019
	TCN	-0.352	0.010
RepeatFirstHard	Elman	-0.469	0.013
	FART	-0.384	0.087
	FWP	0.619	0.288
	Fr.Stack	-0.420	0.050
	GRU	**0.997**	**0.003**
	IndRNN	0.363	0.544

Continued on next page

Env.	Model	μ	σ
	LMU	-0.468	0.009
	LSTM	-0.327	0.201
	MLP	-0.389	0.023
	PosMLP	-0.449	0.022
	S4D	-0.326	0.086
	TCN	-0.460	0.022
RepeatFirstMedium	Elman	-0.461	0.010
	FART	0.940	0.091
	FWP	0.896	0.181
	Fr.Stack	-0.453	0.033
	GRU	**1.000**	**0.000**
	IndRNN	0.994	0.009
	LMU	0.031	0.573
	LSTM	0.292	0.374
	MLP	-0.427	0.026
	PosMLP	-0.467	0.004
	S4D	-0.338	0.055
	TCN	-0.465	0.016
RepeatPreviousEasy	**Elman**	**1.000**	**0.000**
	FART	-0.155	0.011
	FWP	0.324	0.051
	Fr.Stack	**1.000**	**0.000**
	GRU	**1.000**	**0.000**
	IndRNN	0.877	0.087
	LMU	**1.000**	**0.000**
	LSTM	**1.000**	**0.000**
	MLP	-0.481	0.001
	PosMLP	-0.481	0.002
	S4D	-0.492	0.003
	TCN	**1.000**	**0.000**
RepeatPreviousHard	Elman	-0.473	0.005
	FART	-0.484	0.006
	FWP	-0.442	0.017
	Fr.Stack	-0.488	0.001
	GRU	-0.432	0.007
	IndRNN	-0.450	0.029
	LMU	**0.251**	**0.039**
	LSTM	-0.371	0.006
	MLP	-0.489	0.003
	PosMLP	-0.489	0.002
	S4D	-0.489	0.002
	TCN	-0.486	0.002
RepeatPreviousMedium	Elman	-0.448	0.006
	FART	-0.479	0.008
	FWP	-0.268	0.077
	Fr.Stack	-0.481	0.002
	GRU	-0.345	0.026
	IndRNN	-0.377	0.005
	LMU	**0.823**	**0.217**
	LSTM	-0.253	0.005
	MLP	-0.488	0.005
	PosMLP	-0.487	0.003
	S4D	-0.490	0.003
	TCN	-0.484	0.001

Env.	Model	μ	σ
StatelessCartPoleEasy	**Elman**	**1.000**	**0.000**
	FART	0.983	0.007
	FWP	0.926	0.053
	Fr.Stack	**1.000**	**0.000**
	GRU	**1.000**	**0.000**
	IndRNN	**1.000**	**0.000**
	LMU	**1.000**	**0.000**
	LSTM	**1.000**	**0.000**
	MLP	0.266	0.001
	PosMLP	0.760	0.015
	S4D	0.178	0.039
	TCN	**1.000**	**0.000**
StatelessCartPoleHard	**Elman**	**1.000**	**0.000**
	FART	0.391	0.054
	FWP	0.455	0.024
	Fr.Stack	0.984	0.009
	GRU	**1.000**	**0.000**
	IndRNN	**1.000**	**0.000**
	LMU	**0.999**	**0.001**
	LSTM	**0.998**	**0.003**
	MLP	0.088	0.001
	PosMLP	0.253	0.015
	S4D	0.052	0.022
	TCN	0.916	0.048
StatelessCartPoleMedium	**Elman**	**1.000**	**0.000**
	FART	0.609	0.070
	FWP	0.568	0.039
	Fr.Stack	**0.999**	**0.001**
	GRU	**1.000**	**0.000**
	IndRNN	**1.000**	**0.000**
	LMU	**0.998**	**0.001**
	LSTM	0.992	0.006
	MLP	0.133	0.002
	PosMLP	0.384	0.018
	S4D	0.078	0.009
	TCN	0.994	0.002
StatelessPendulumEasy	**Elman**	**0.908**	**0.004**
	FART	0.702	0.023
	FWP	0.516	0.012
	Fr.Stack	**0.912**	**0.001**
	GRU	**0.913**	**0.000**
	IndRNN	0.440	0.062
	LMU	0.822	0.011
	LSTM	0.873	0.046
	MLP	0.378	0.009
	PosMLP	0.479	0.009
	TCN	**0.909**	**0.003**
StatelessPendulumHard	**Elman**	**0.824**	**0.002**
	FART	0.704	0.103
	FWP	0.592	0.010
	Fr.Stack	**0.825**	**0.003**
	GRU	**0.826**	**0.001**
	IndRNN	0.741	0.078
	LMU	**0.823**	**0.002**

Continued on next page

Env.	Model	μ	σ
	LSTM	**0.825**	**0.006**
	MLP	0.387	0.008
	PosMLP	0.513	0.007
	TCN	**0.826**	**0.002**
StatelessPendulumMedium	**Elman**	**0.897**	**0.044**
	FART	0.627	0.028
	FWP	0.552	0.008
	Fr.Stack	0.883	0.001
	GRU	0.885	0.001
	IndRNN	0.466	0.029
	LMU	0.878	0.002
	LSTM	**0.900**	**0.045**
	MLP	0.383	0.003
	PosMLP	0.490	0.003
	TCN	0.883	0.003

Env.	Model	μ	σ
MultiarmedBanditEasy	**Elman**	**0.736**	**0.229**
	FART	0.452	0.062
	FWP	0.518	0.059
	Fr.Stack	0.440	0.104
	GRU	0.615	0.018
	IndRNN	0.589	0.028
	LMU	0.207	0.034
	LSTM	0.677	0.280
	MLP	0.416	0.043
	PosMLP	0.291	0.053
	S4D	0.054	0.022
	TCN	0.470	0.037
MultiarmedBanditHard	Elman	0.039	0.004
	FART	0.247	0.046
	FWP	0.189	0.025
	Fr.Stack	0.292	0.075
	GRU	0.498	0.080
	IndRNN	**0.562**	**0.019**
	LMU	0.038	0.005
	LSTM	0.388	0.014
	MLP	0.302	0.007
	PosMLP	0.185	0.053
	S4D	0.114	0.031
	TCN	0.136	0.056
MultiarmedBanditMedium	Elman	0.421	0.079
	FART	0.298	0.076
	FWP	0.388	0.102
	Fr.Stack	0.384	0.063
	GRU	0.575	0.025
	IndRNN	**0.702**	**0.260**
	LMU	0.037	0.006
	LSTM	0.406	0.006
	MLP	0.375	0.024
	PosMLP	0.204	0.048

Continued on next page

Env.	Model	μ	σ
NoisyStatelessCartPoleEasy	S4D	0.119	0.014
	TCN	0.332	0.035
	Elman	**1.000**	**0.000**
	FART	0.985	0.011
	FWP	0.940	0.075
	Fr.Stack	**1.000**	**0.000**
	GRU	**1.000**	**0.000**
	IndRNN	**1.000**	**0.000**
	LMU	**0.999**	**0.001**
	LSTM	**1.000**	**0.000**
	MLP	0.265	0.002
	PosMLP	0.744	0.023
NoisyStatelessCartPoleHard	S4D	0.176	0.003
	TCN	**1.000**	**0.000**
	Elman	**1.000**	**0.000**
	FART	0.981	0.014
	FWP	0.963	0.041
	Fr.Stack	**1.000**	**0.000**
	GRU	**1.000**	**0.000**
	IndRNN	**1.000**	**0.000**
	LMU	**1.000**	**0.000**
	LSTM	**1.000**	**0.000**
	MLP	0.265	0.001
	PosMLP	0.743	0.021
NoisyStatelessCartPoleMedium	S4D	0.171	0.042
	TCN	**1.000**	**0.000**
	Elman	**1.000**	**0.000**
	FART	0.981	0.008
	FWP	0.974	0.002
	Fr.Stack	**1.000**	**0.000**
	GRU	**1.000**	**0.000**
	IndRNN	**1.000**	**0.000**
	LMU	**0.999**	**0.001**
	LSTM	**1.000**	**0.000**
	MLP	0.265	0.001
	PosMLP	0.741	0.004
NoisyStatelessPendulumEasy	S4D	0.162	0.020
	TCN	**1.000**	**0.001**
	Elman	**0.931**	**0.046**
	FART	0.594	0.010
	FWP	0.513	0.038
	Fr.Stack	0.913	0.001
	GRU	0.913	0.001
	IndRNN	0.798	0.178
	LMU	0.907	0.003
	LSTM	0.905	0.010
	MLP	0.373	0.010
	PosMLP	0.463	0.008
NoisyStatelessPendulumHard	TCN	0.911	0.003
	Elman	0.909	0.003
	FART	0.692	0.138
	FWP	0.515	0.017
	Fr.Stack	**0.911**	**0.001**
	GRU	**0.915**	**0.001**

Continued on next page

Env.	Model	μ	σ
NoisyStatelessPendulumMedium	IndRNN	0.573	0.063
	LMU	0.906	0.001
	LSTM	0.906	0.001
	MLP	0.375	0.005
	PosMLP	0.468	0.012
	TCN	**0.910**	**0.002**
	Elman	**0.931**	**0.046**
	FART	0.621	0.017
	FWP	0.523	0.017
	Fr.Stack	0.911	0.001
	GRU	0.914	0.002
	IndRNN	0.843	0.093
	LMU	0.908	0.001
	LSTM	0.906	0.003
	MLP	0.375	0.003
	PosMLP	0.469	0.015
	TCN	0.911	0.002

Env.	Model	μ	σ
BattleshipEasy	Elman	-0.329	0.013
	FART	-0.414	0.006
	FWP	-0.400	0.009
	Fr.Stack	-0.408	0.007
	GRU	-0.353	0.008
	IndRNN	-0.309	0.026
	LMU	-0.396	0.007
	LSTM	-0.378	0.007
	MLP	-0.353	0.016
	PosMLP	**-0.101**	**0.283**
	S4D	-0.436	0.002
	TCN	-0.168	0.447
BattleshipHard	Elman	-0.383	0.003
	FART	-0.382	0.003
	FWP	-0.378	0.004
	Fr.Stack	-0.382	0.003
	GRU	-0.379	0.001
	IndRNN	**-0.362**	**0.010**
	LMU	-0.381	0.003
	LSTM	-0.385	0.002
	MLP	-0.380	0.002
	PosMLP	-0.382	0.005
	S4D	-0.385	0.006
	TCN	-0.380	0.007
BattleshipMedium	Elman	-0.387	0.002
	FART	-0.395	0.005
	FWP	-0.383	0.005
	Fr.Stack	-0.390	0.005
	GRU	-0.372	0.004
	IndRNN	**-0.339**	**0.011**
	LMU	-0.397	0.004
	LSTM	-0.388	0.006

Continued on next page

Env.	Model	μ	σ
	MLP	-0.372	0.008
	PosMLP	-0.375	0.004
	S4D	-0.405	0.007
	TCN	-0.393	0.003
ConcentrationEasy	**Elman**	**0.435**	**0.006**
	FART	0.433	0.006
	FWP	0.429	0.003
	Fr.Stack	**0.434**	**0.005**
	GRU	0.430	0.004
	IndRNN	0.407	0.024
	LMU	**0.439**	**0.004**
	LSTM	0.433	0.002
	MLP	**0.434**	**0.007**
	PosMLP	**0.436**	**0.004**
	S4D	0.433	0.001
	TCN	**0.435**	**0.002**
ConcentrationHard	Elman	-0.819	0.003
	FART	-0.820	0.001
	FWP	-0.821	0.002
	Fr.Stack	-0.736	0.186
	GRU	-0.819	0.006
	IndRNN	-0.782	0.107
	LMU	-0.817	0.005
	LSTM	-0.738	0.188
	MLP	-0.820	0.001
	PosMLP	-0.821	0.001
	S4D	-0.822	0.001
	TCN	**-0.711**	**0.215**
ConcentrationMedium	**Elman**	**0.463**	**0.002**
	FART	**0.467**	**0.003**
	FWP	**0.468**	**0.002**
	Fr.Stack	**0.468**	**0.004**
	GRU	**0.464**	**0.005**
	IndRNN	0.391	0.013
	LMU	**0.466**	**0.001**
	LSTM	**0.466**	**0.006**
	MLP	**0.467**	**0.002**
	PosMLP	**0.467**	**0.004**
	TCN	**0.467**	**0.005**
HigherLowerEasy	Elman	0.516	0.004
	FART	0.522	0.003
	FWP	0.522	0.001
	Fr.Stack	0.503	0.001
	GRU	**0.525**	**0.003**
	IndRNN	**0.528**	**0.001**
	LMU	0.498	0.004
	LSTM	**0.524**	**0.001**
	MLP	0.504	0.001
	PosMLP	0.505	0.001
	S4D	0.324	0.064
	TCN	0.503	0.001
HigherLowerHard	Elman	0.503	0.001
	FART	**0.506**	**0.002**
	FWP	0.499	0.002

Continued on next page

Env.	Model	μ	σ
HigherLowerMedium	Fr.Stack	0.499	0.001
	GRU	0.504	0.001
	IndRNN	**0.509**	**0.001**
	LMU	0.468	0.012
	LSTM	0.503	0.003
	MLP	**0.505**	**0.000**
	PosMLP	0.503	0.001
	S4D	0.241	0.112
	TCN	0.497	0.002
	Elman	0.504	0.002
	FART	**0.512**	**0.002**
	FWP	0.503	0.004
MineSweeperEasy	Fr.Stack	0.503	0.001
	GRU	0.509	0.003
	IndRNN	**0.514**	**0.001**
	LMU	0.472	0.004
	LSTM	0.509	0.001
	MLP	0.504	0.000
	PosMLP	0.504	0.000
	S4D	0.302	0.078
	TCN	0.501	0.001
	Elman	0.756	0.007
	FART	0.549	0.032
	FWP	0.474	0.026
MineSweeperHard	Fr.Stack	0.623	0.016
	GRU	0.850	0.008
	IndRNN	0.747	0.035
	LMU	0.813	0.003
	LSTM	**0.863**	**0.008**
	MLP	0.265	0.017
	PosMLP	0.443	0.030
	S4D	-0.121	0.069
	TCN	0.662	0.031
	Elman	-0.193	0.012
	FART	-0.423	0.050
MineSweeperMedium	FWP	-0.350	0.015
	Fr.Stack	-0.395	0.058
	GRU	**-0.172**	**0.023**
	IndRNN	-0.226	0.019
	LMU	-0.302	0.019
	LSTM	-0.381	0.087
	MLP	-0.305	0.116
	PosMLP	-0.302	0.123
	S4D	-0.446	0.029
	TCN	-0.387	0.091
	Elman	0.071	0.012
	FART	-0.169	0.097
	FWP	-0.157	0.025
	Fr.Stack	-0.227	0.169
	GRU	**0.078**	**0.008**
	IndRNN	0.016	0.020
	LMU	-0.087	0.007
	LSTM	-0.127	0.036
	MLP	-0.132	0.006

Continued on next page

Env.	Model	μ	σ
	PosMLP	-0.093	0.004
	S4D	-0.397	0.004
	TCN	-0.107	0.087

Env.	Model	μ	σ
LabyrinthEscapeEasy	**Elman**	**0.956**	**0.005**
	FART	-0.032	0.074
	FWP	-0.076	0.086
	Fr.Stack	0.504	0.114
	GRU	**0.958**	**0.001**
	IndRNN	0.191	0.225
	LMU	0.615	0.140
	LSTM	0.914	0.023
	MLP	**0.961**	**0.001**
	PosMLP	0.539	0.101
	S4D	-0.200	0.030
	TCN	0.440	0.148
LabyrinthEscapeHard	Elman	-0.803	0.019
	FART	-0.829	0.026
	FWP	-0.842	0.015
	Fr.Stack	-0.795	0.026
	GRU	**-0.725**	**0.037**
	IndRNN	-0.935	0.006
	LMU	-0.822	0.028
	LSTM	-0.812	0.035
	MLP	-0.786	0.092
	PosMLP	-0.733	0.075
	S4D	-0.876	0.027
	TCN	-0.793	0.044
LabyrinthEscapeMedium	Elman	-0.009	0.320
	FART	-0.653	0.085
	FWP	-0.639	0.029
	Fr.Stack	-0.433	0.134
	GRU	0.215	0.586
	IndRNN	-0.707	0.124
	LMU	-0.550	0.072
	LSTM	-0.578	0.115
	MLP	**0.242**	**0.765**
	PosMLP	-0.331	0.066
	S4D	-0.703	0.057
	TCN	-0.462	0.147
LabyrinthExploreEasy	**Elman**	**0.963**	**0.001**
	FART	0.104	0.111
	FWP	-0.153	0.042
	Fr.Stack	0.943	0.014
	GRU	0.959	0.001
	IndRNN	0.937	0.041
	LMU	0.814	0.033
	LSTM	0.645	0.050
	MLP	**0.967**	**0.001**
	PosMLP	0.875	0.075

Continued on next page

Env.	Model	μ	σ
LabyrinthExploreHard	S4D	-0.314	0.041
	TCN	0.928	0.015
	Elman	0.763	0.023
	FART	-0.428	0.070
	FWP	-0.606	0.018
	Fr.Stack	0.382	0.173
	GRU	0.793	0.013
	IndRNN	0.434	0.429
	LMU	-0.009	0.082
	LSTM	-0.156	0.029
	MLP	**0.859**	**0.001**
	PosMLP	-0.078	0.059
LabyrinthExploreMedium	S4D	-0.729	0.013
	TCN	0.010	0.051
	Elman	0.894	0.008
	FART	-0.137	0.067
	FWP	-0.482	0.015
	Fr.Stack	0.755	0.141
	GRU	0.893	0.008
	IndRNN	0.744	0.233
	LMU	0.236	0.090
	LSTM	0.053	0.034
	MLP	**0.922**	**0.001**
	PosMLP	0.291	0.192
	S4D	-0.560	0.062
	TCN	0.546	0.242