Drama: Mamba-Enabled Model-Based Reinforcement Learning Is Sample and Performance Efficient

Wenlong Wang, Ivana Dusparic, Yucheng Shi, Ke Zhang, Vinny Cahill

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Taighde Éireann
Research Ireland

ICLR
International Conference On
Learning Representations

Outline

Introduction

Method

Result

Conclusion

Introduction

Challenges in Deep RL

- **Successes:** Mastery in games like Go [Sil+16; Sil+17], Dota [Ber+19], and Atari [Mni+13], as well as simulated environments like MuJoCo [Sch+17].
- **Key Limitation:** Training requires millions of environment interactions, which is impractical for real-world deployment due to cost and safety constraints.
- **Goal:** Improve sample efficiency to bridge the gap between theoretical advancements and real-world applications.

Model-Based RL using World Models

- **Approach**: Learn environment dynamics using sequence models (e.g., Transformers, RNNs) to generate *synthetic training data*.
- **Advantage**: Reduces reliance on costly real-world interactions.
- **Challenges**:
  - Model-Based RL (MBRL) architectures often require large parameter counts (25M–200M), increasing computational overhead.
  - Early prediction errors in world models can propagate, leading to biased policies that are prone to local optima (difficult to correct).

Sequence Modeling for World Models

- **RNNs (LSTM/GRU)**: Linear complexity, but struggle to capture long-range dependencies and suffer from vanishing/exploding gradients [Haf+23].
- **Transformers**: Powerful performance, but $O(n^2)$ complexity makes them computationally costly for long sequence processing. They also inefficiently allocate representation capacity by storing all positional interactions [MAF23; Rob+23].
- **SSMs (e.g., Mamba/Mamba2)**: Linear complexity, excel in handling long-range dependencies, and enhance representation efficiency through selective information compression [GD24; DG24].

Key Contributions

- **Drama**: Achieves SOTA on Atari100k with a 7M-parameter world model.
- **Mamba2>Mamba1 in MBRL**: We evaluate the performance of state-of-the-art SSMs as world models on the Atari100k benchmark and demonstrate the superiority of Mamba-2 for modelling dynamics in Atari games.
- **Dynamic Frequency Sampling (DFS)**: Mitigates imperfect dynamics via adaptive sampling.

Method

Drama structure

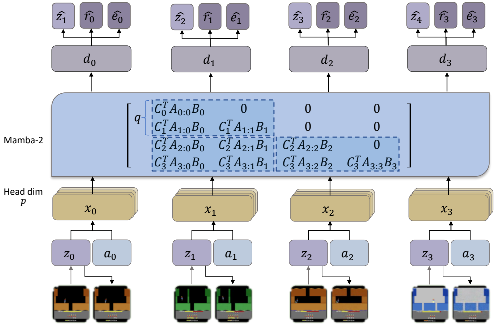


Figure: Raw frames are encoded into $\boldsymbol{z}_t$ and combined with action a_t as input to Mamba blocks. The input is split by head dimension p to compute the recurrent deterministic state d_t, which predicts $\hat{\boldsymbol{z}}_{t+1}$, reward $\hat{r}_t$, and termination $\hat{e}_t$.

Discrete Variational Auto-encoder

- Extends standard VAE architecture.
- Incorporates fully-connected layer to discretise latent embeddings.
- Raw observation: $\mathbf{O}_t \in [0, 255]^{(3,64,64)}$.
- Encoder compresses observation into discrete vector: $z_t \sim p(z_t|\mathbf{O}_t)$.
- Decoder reconstructs raw image: $\hat{\mathbf{O}}_t$.
- Gradients passed using straight-through estimator.

Sequences Model

- Simulates environment in latent variable space $\mathbf{z}_t$.
- Deterministic state variable: $\boldsymbol{d}_t$.
- Implemented with Mamba/Mamba-2.
- Dynamics model equation:

$$\boldsymbol{d}_t = f(\mathbf{z}_{t-l:t}, \mathbf{a}_{t-l:t}; \omega)$$

- Latent variable predictor:

$$\hat{\mathbf{z}}_{t+1} \sim p(\hat{\mathbf{z}}_{t+1} | \boldsymbol{d}_t; \omega)$$

Behaviour Policy Learning

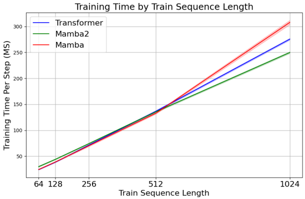
- Trained within 'imagination' process driven by dynamics model.
- Rollout begins from last transition in each sequence.
- Key difference: Mamba updates inference parameters independently of sequence length.
- State concatenates prior discrete variable $\hat{\boldsymbol{z}}_t$ with deterministic variable $\boldsymbol{d}_t$.
- Uses standard actor-critic architecture, but also compatible with other RL algorithm: e.g., PPO.

Dynamic Frequency-Based Sampling

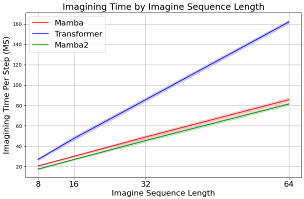
- Mitigate issues arising from an inaccurate world model in model-based RL.
- Introduces two vectors during training:
  - $\boldsymbol{v}$: Tracks world model usage.
  - $\boldsymbol{b}$: Tracks behaviour policy usage.
- Sampling probabilities:
  - For world model: $(p_1, p_2, \ldots, p_{|\mathcal{E}|}) = \mathtt{softmax}(-\boldsymbol{v})$.
  - For imagination: $(p_1, p_2, \ldots, p_{|\mathcal{E}|}) = \mathtt{softmax}(f(\boldsymbol{v}, \boldsymbol{b}))$.
- Ensures transitions are sampled based on learning progress.

Result

Time comparison



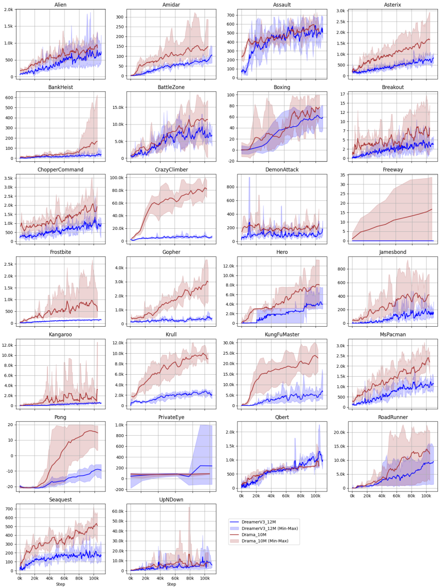
Training the world model



Autogenerative 'imagination'

Figure: Wall-clock time comparison of sequence models in MBRL. Experiments were conducted on a consumer-grade laptop with an NVIDIA RTX 2000 Ada Mobile GPU, ensuring practical relevance to resource-constrained settings.

Drama(10M) vs. DreamerV3 (12M)

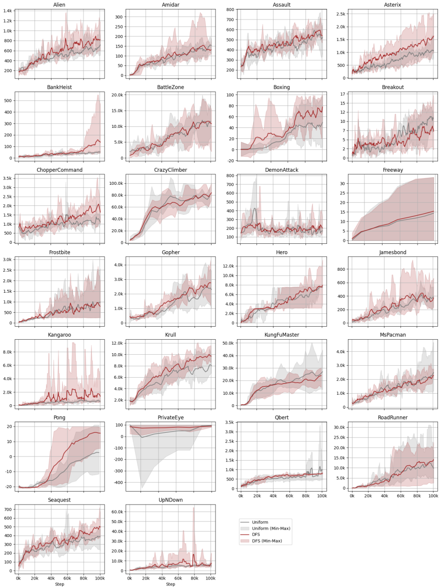


DramaXS(10M) vs. DreamerV3XS(12M)

Metric	Random	Human	DramaXS	DreamerV3XS
Mean (%)	0	100	105	37
Median (%)	0	100	27	7

Table: With limited parameters, Drama significantly outperform DreamerV3 in the Atari100k benchmark.

Uniform sampling vs. DFS

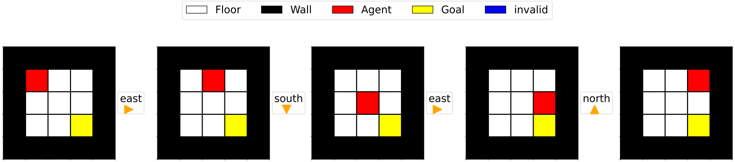


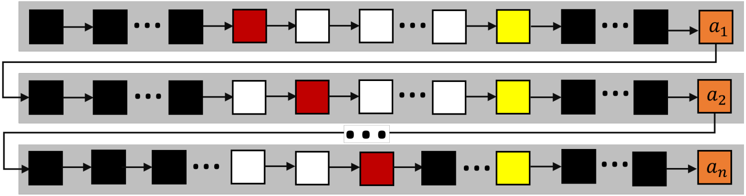
Uniform sampling vs. DFS

Game	Random	Human	DFS	Uniform
Mean (%)	0	100	105	80
Median (%)	0	100	27	28

Table: The Atari100K performance table demonstrates that the Drama XS model, when paired with DFS, achieves a higher normalized mean score compared to using the uniform sampling method. This highlights the effectiveness of DFS in enhancing performance on Atari100K benchmarks within Mamba-powered MBRL.

Dynamics models for long-sequence predictability tasks





Long-seq predictability task result

Method	l	Training Time (ms)	Memory Usage (%)	Error (%)
Mamba-2	208	25	13	15.6 ± 2.6
	1664	214	55	14.2 ± 0.3
Mamba-1	208	34	14	13.9 ± 0.4
	1664	299	52	14.0 ± 0.4
GRU	208	75	66	21.3 ± 0.3
	1664	628	68	34.7 ± 25.4
Transformer	208	45	17	75.0 ± 1.1
	1664	-	OOM	-

Table: Performance comparison of different methods on the grid world environment. Memory usage is reported as a percentage of an 8GB GPU. The error is represented as the mean $\pm$ standard deviation. The training time refers to the average duration of one training step. Note that the Transformer encounters an out-of-memory (OOM) error during training with long sequences.

Conclusion

Conclusion

- **Key Contributions of Drama:**
  - Addresses the challenges of RNN and transformer-based world models.
  - Achieves $O(n)$ memory and computational complexity, enabling longer training sequences.
  - Novel sampling method mitigates suboptimality during early training.
  - Lightweight world model with only 7M trainable parameters, trainable on standard hardware.

Acknowledgement

This publication has emanated from research conducted with the financial support of Taighde Éireann - Research Ireland under Frontiers for the Future grant number 21/FFP-A/8957 and grant number 18/CRT/6223. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.