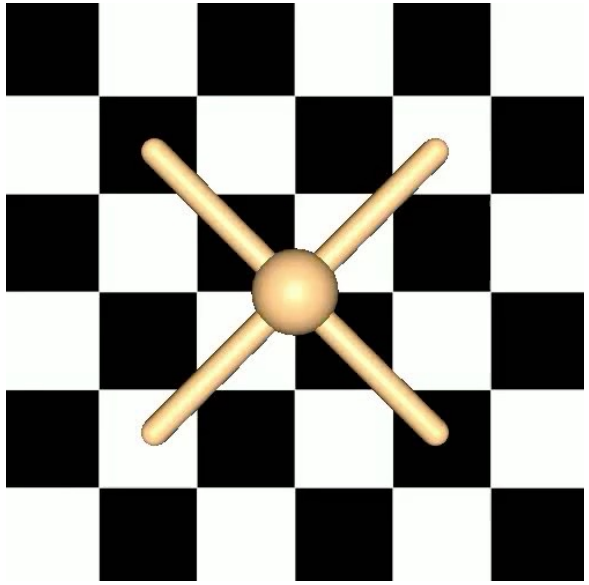
TECHNION
Israel Institute
of Technology

MAMBA: an Effective World Model Approach for Meta-Reinforcement Learning

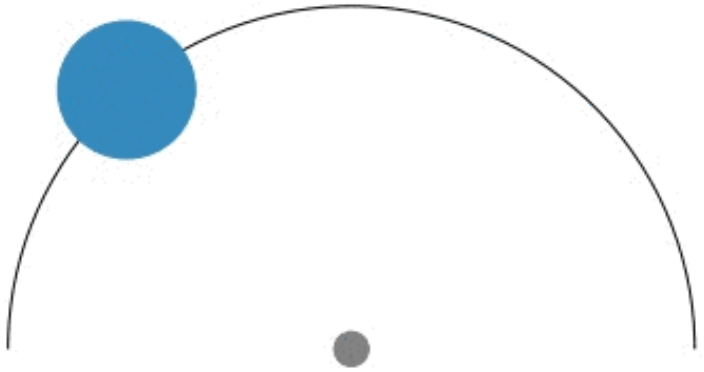
Zohar Rimon*, Tom Jurgenson*, Orr Krupnik, Gilad Adler, Aviv Tamar

Meta Reinforcement Learning

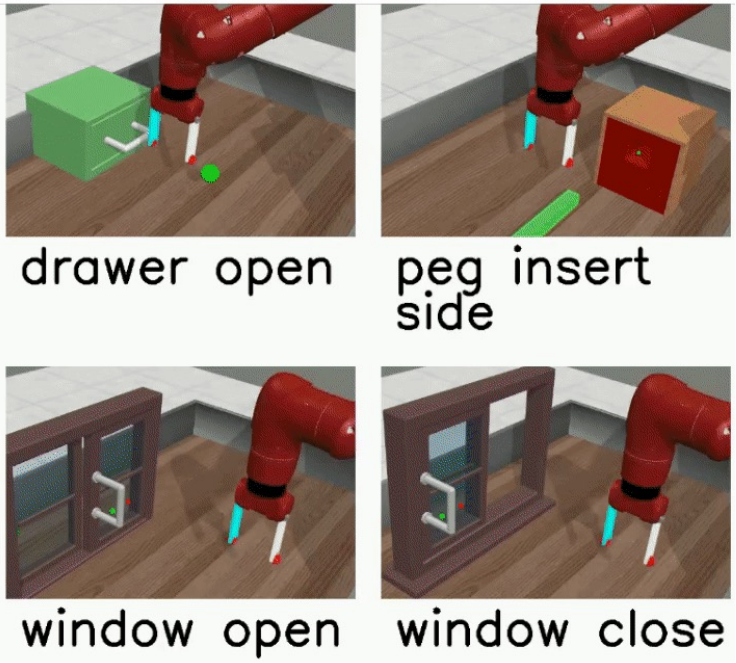
[1]



[2]



[3]



Fail to scale to complex distributions

[1] Dorfman et al. Offline Meta RL – Identifiability Challenges and Effective Data Collection Strategies, 2021
[2] Yu et al. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning, 2019
[3] Rakelly et al. Efficient off-policy meta-reinforcement learning via probabilistic context variables, 2019

High Dimension Task Distributions

Problem: Meta-RL is hard in distributions with many DoF

- Shown empirically [1]
- And theoretically [2]

Number of training tasks needed scale exponentially with the DoF

Is all hope lost?

[1] Mandi et al., On the Effectiveness of Fine-tuning Versus Meta-reinforcement Learning, 2022
[2] Rimon et al., Meta Reinforcement Learning with Finite Training Tasks - a Density Estimation Approach, 2022

Decomposable Task Distributions (DTDs)

Solution: Tasks can be decomposed → improved bounds

- Exponential dependency on the dimensionality of the sub-tasks
- Linear dependency on the number of sub tasks

[1]

[1] Herzog et al. Deep RL at Scale: Sorting Waste in Office Buildings with a Fleet of Mobile Manipulators (2023)

Effectiveness of Current Approaches

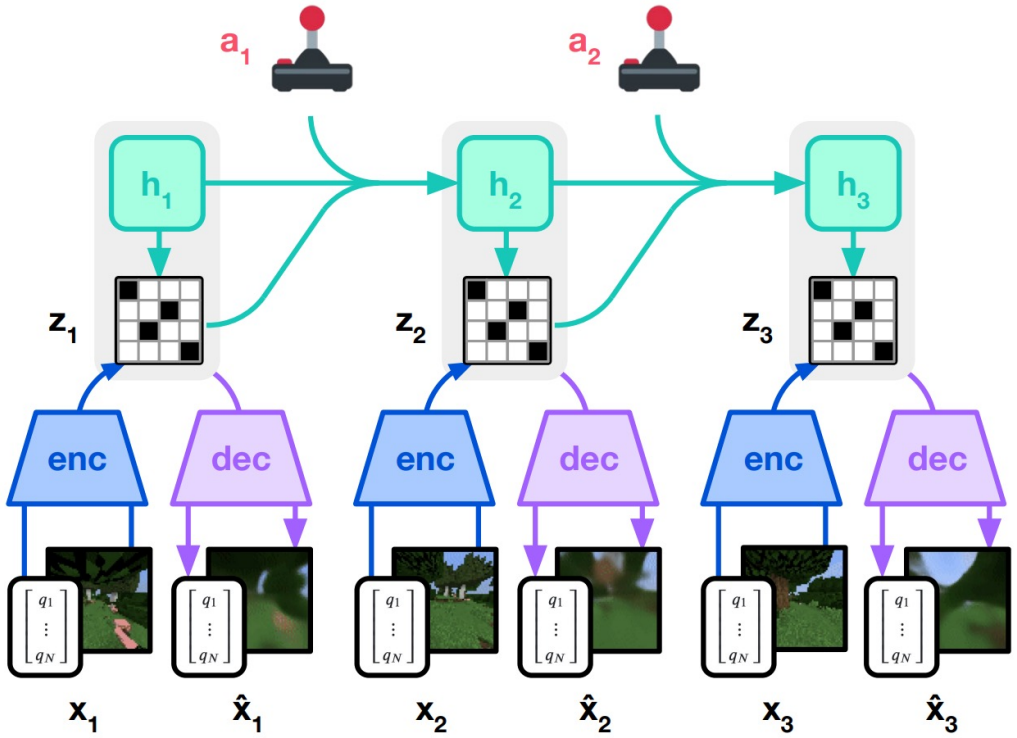
Problem: Current algorithms are

- Sample inefficient
- Very sensitive to hyperparameter tuning

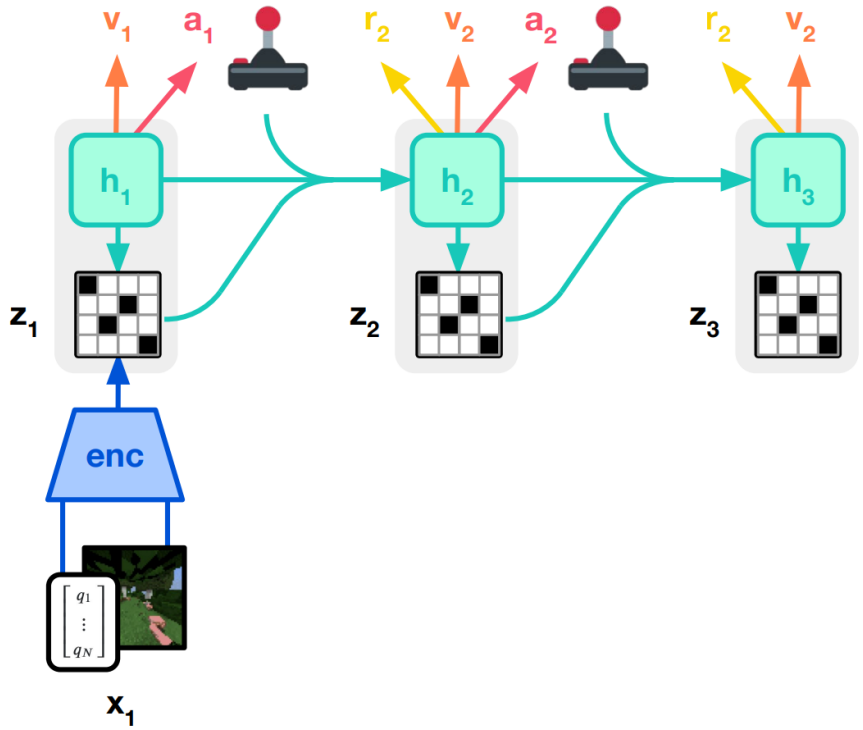
Solution:

Incorporate advancements in model-based RL

DreamerV3



(a) World Model Learning



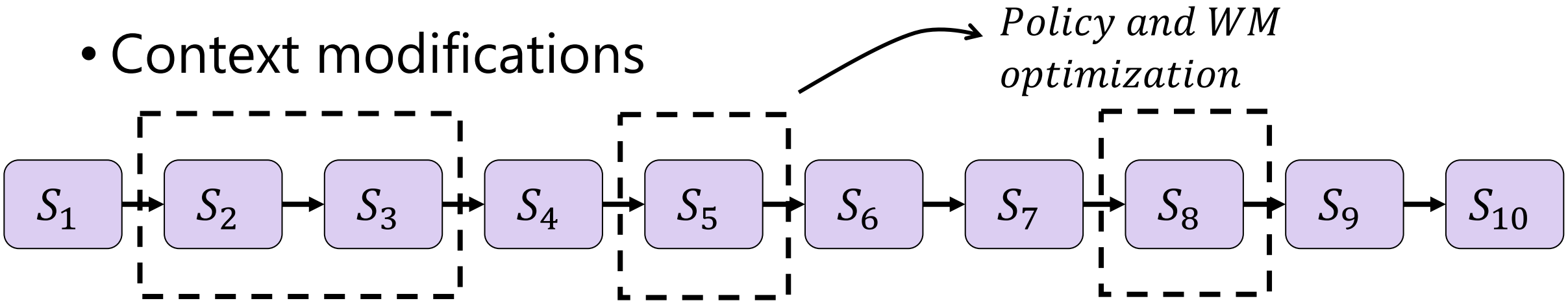
(b) Actor Critic Learning

- SOTA on many POMDPs
- Meta-RL is a special case of POMDPs

Danijar et al. Mastering diverse domains through world models (2023)

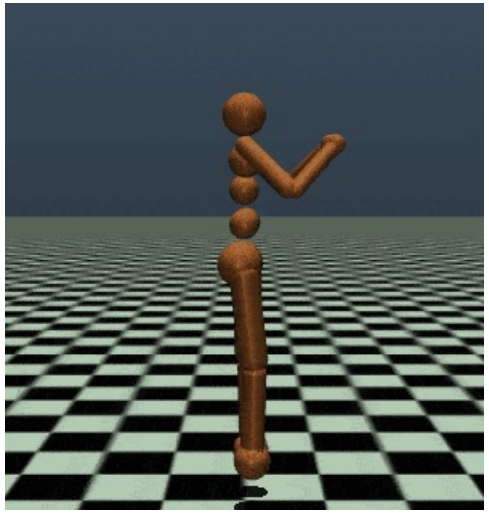
MAMBA – Met**A** RL **M**odel **B**ased **A**lgorithm

- Based on DreamerV3 architecture
- Modified to work for meta-rl:
  - Reward and time-step added to observations
  - Context modifications

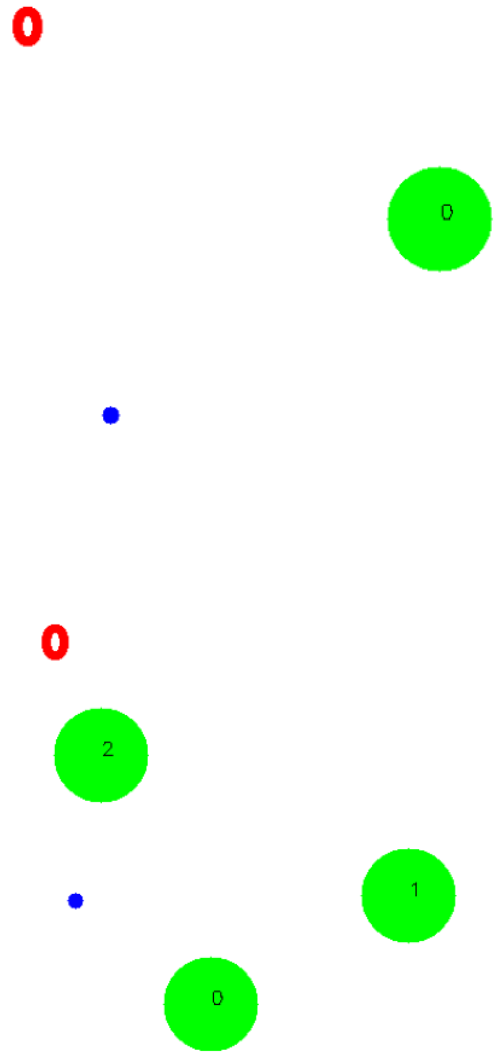


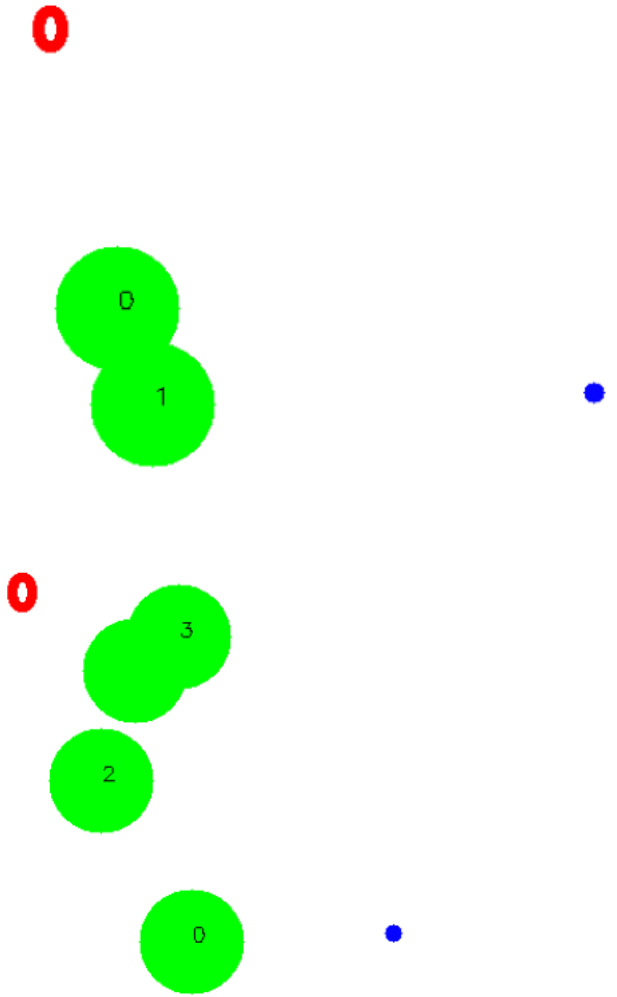
  - Added horizon scheduling

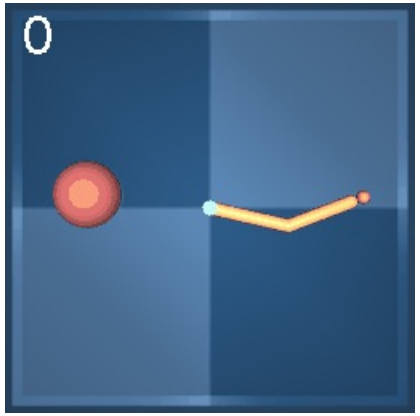
Experiments - Standard Benchmarks



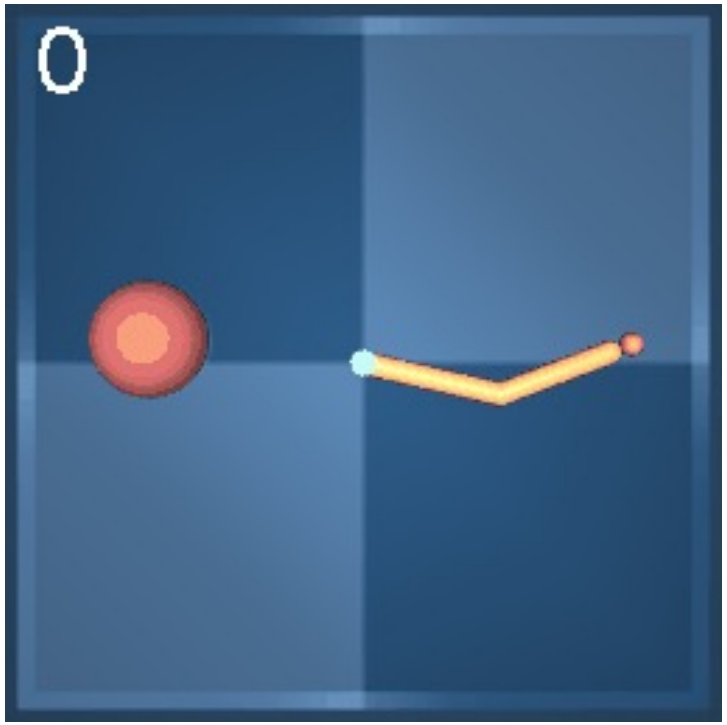
Experiments - Reacher Multi Goal







Experiments – Learning From Pixels



Experiments

	RL^2	VariBAD	HyperX	Dreamer-Vanilla	Dreamer-Tune	MAMBA (Ours)
HalfCircle:	239.4±3.1	218.0 ± 26.0	204.0 ± 19.9	**241.1 ± 9.9**	**239.7 ± 3.1**	**242.2 ± 7.9**
	∼10M	∼10M	∼15M	∼2M	∼2M	∼2M
Rooms-3:	108.0±31.7	**158.7 ± 2.6**	111.8 ± 20.2	137.5 ± 6.2	142.6 ± 4.1	**156.2 ± 1.7**
Rooms-4:	85.1±19.0	88.1 ± 39.2	14.1 ± 3.0	115.0 ± 6.7	119.8 ± 3.	**136.7 ± 1.4**
Rooms-5:	72.4±36.8	0.9 ± 2.5	-15.8 ± 0.4	96.5 ± 2.5	93.9 ± 1.3	**113.1 ± 5.7**
Rooms-6:	64.9±15.9	-11.0 ± 0.9	-15.7 ± 0.7	69.5 ± 6.3	71.0 ± 6.3	**94.5 ± 1.2**
Rooms-7:	42.9±22.9	-	-	-	50.2±4.4	**73.2±2.9**
Rooms-8:	29.0±17.	-	-	-	29.1±8.3	**55.6±3.0**
	∼100M	∼100M	∼100M	∼5M	∼6M	∼6M
Reacher-1:	-	473.7 ± 27.5	5[illegible] ± 36.0	552.0 ± 27.8	503.4 ± 69.0	**655.5 ± 12.3**
Reacher-2:	-	46.6 ± 3.	30.0 ± 48.5	**247.4 ± 80.5**	**217.6 ± 64.3**	**285.8 ± 89.6**
Reacher-3:	-	0.2 ± 0.	0.5 ± 0.2	183.6 ± 100.0	76.9 ± 80.5	**325.0 ± 47.0**
Reacher-4:	-	0.0 ± 0.0	-0.5 ± 0.1	0.4 ± 0.0	0.1 ± 0.2	**77.7 ± 61.1**
	-	∼150M	∼150M	∼10M	∼10M	∼10M
EscapeRoom:	79.9±4.4	70.7 ± 3.	60.9 ± 6.5	68.2 ± 2.4	**73.2 ± 7.8**	**73.9 ± 3.1**
	∼20M	∼20M	∼20M	∼4M	∼4M	∼4M
Point-Wind:	-	14.5±44.7	177.0 ± 118.5	226.1 ± 3.5	224.5 ± 4.4	**224.1 ± 5.2**
	-	∼20M	∼20M	∼2M	∼2M	∼2M
Humanoid-Dir:	-	1369.3 ± 75.3	-	2068.3 ± 156.7	2096.3 ± 79.8	**2405.9 ± 119.0**
	-	∼100M	-	∼30M	∼30M	∼30M

Thank You!

Come find us in the poster session

github.com/zoharri/mamba

zohar.rimon@campus.technion.ac.il