

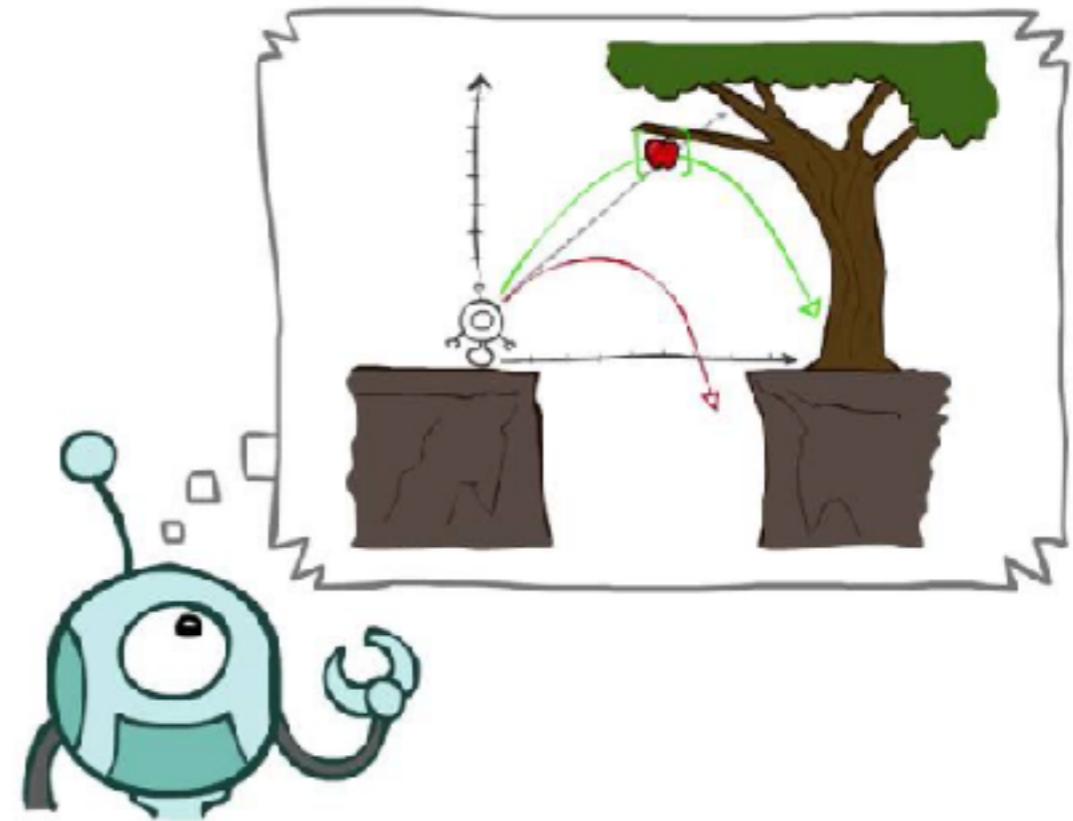


**Learning rollout policy for
internal planning in deep
reinforcement learning agent** || **Metody planowania w
głębokim uczeniu ze
wzmocnieniem**

Piotr Januszewski

Motivation

Humans make plans

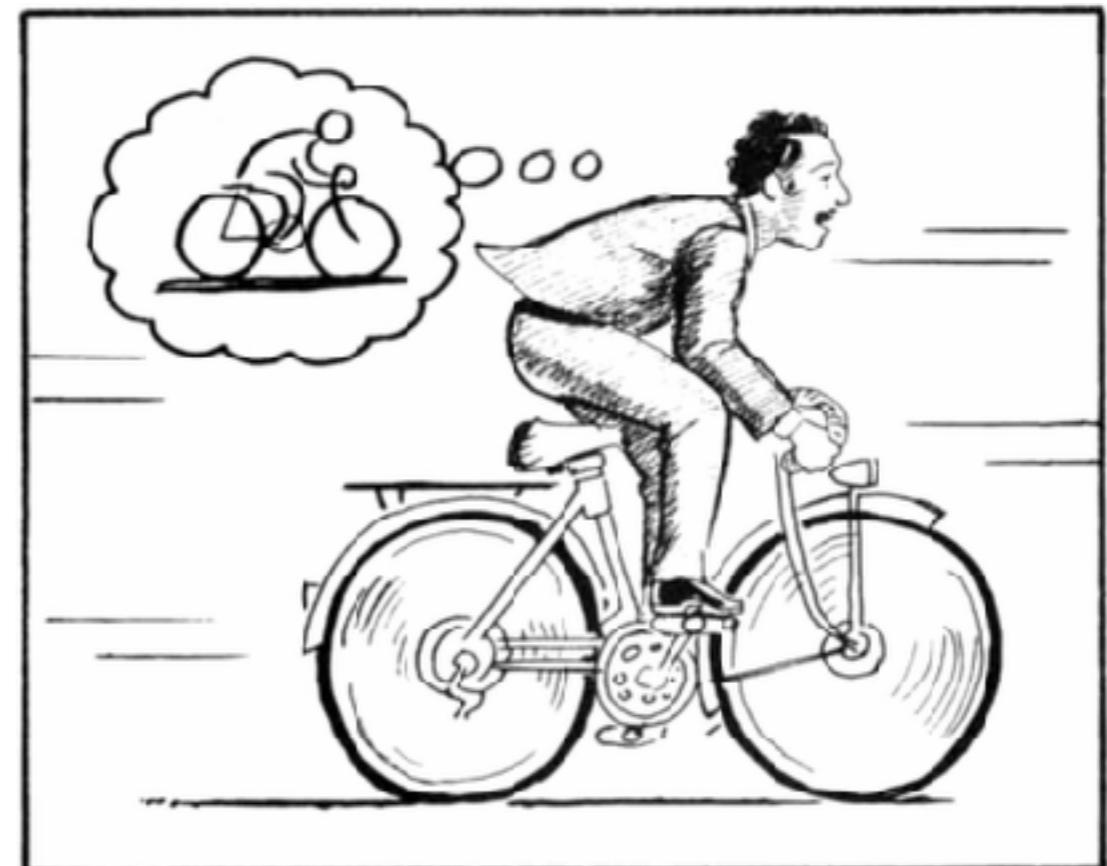


Source: <http://ai.berkeley.edu/>

Mental model

The image of the world around us, which we carry in our head, is just a model. Nobody in his head imagines all the world, government or country. He has only selected concepts, and relationships between them, and uses those to represent the real system

~ Jay Wright Forrester, 1971



A World Model, from Scott McCloud's
Understanding Comics.

MUSCLE MEMORY



Problem

Goal



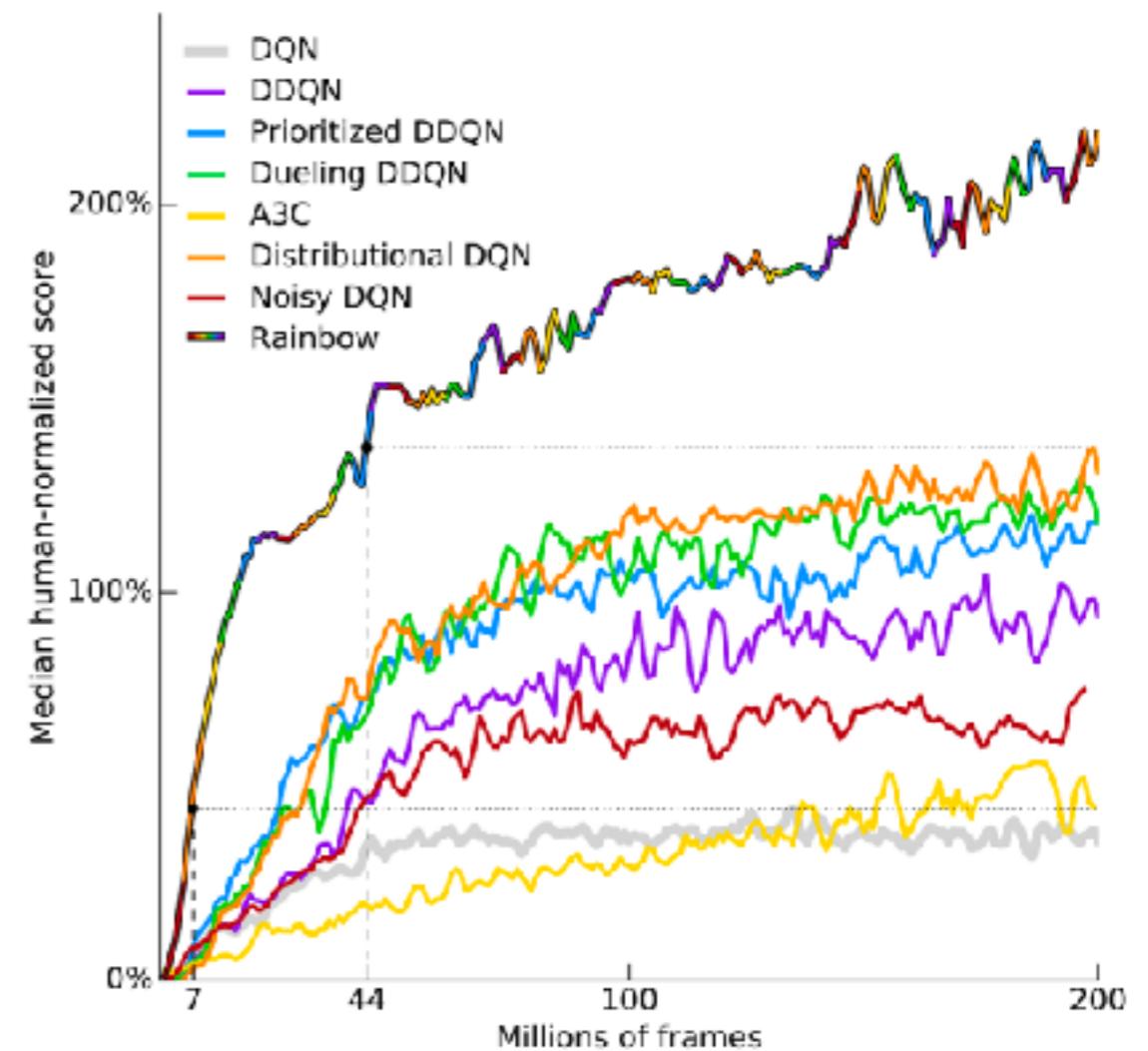
The goal is to improve data efficiency without loss in performance compared to model-free methods. This work focuses on three benchmarks: an arcade game with dense rewards Boxing, a challenging environment with sparse rewards Freeway and a complex puzzle game Sokoban.

*State of the art and
a little bit of theory*

State of the art

Model-free:

- Rainbow - a compilation of several independent improvements to the DQN algorithm made by the deep reinforcement learning community.
- PPO - the new family of policy gradient methods for reinforcement learning, which alternate between sampling data through interaction with the environment, and optimising a surrogate objective function using stochastic gradient ascent.

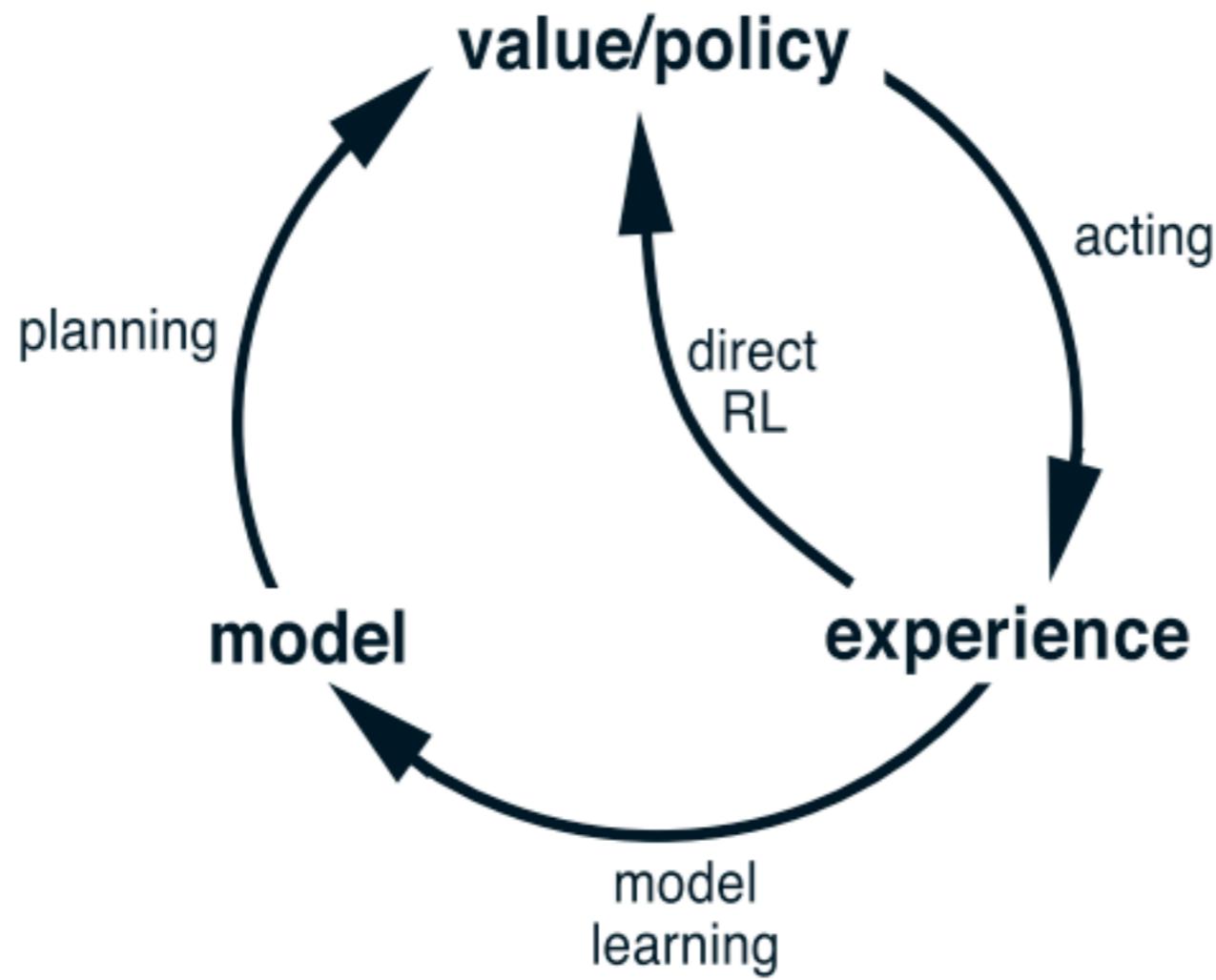


Median human-normalised performance across 57 Atari games from the Rainbow paper.

State of the art

Model-based:

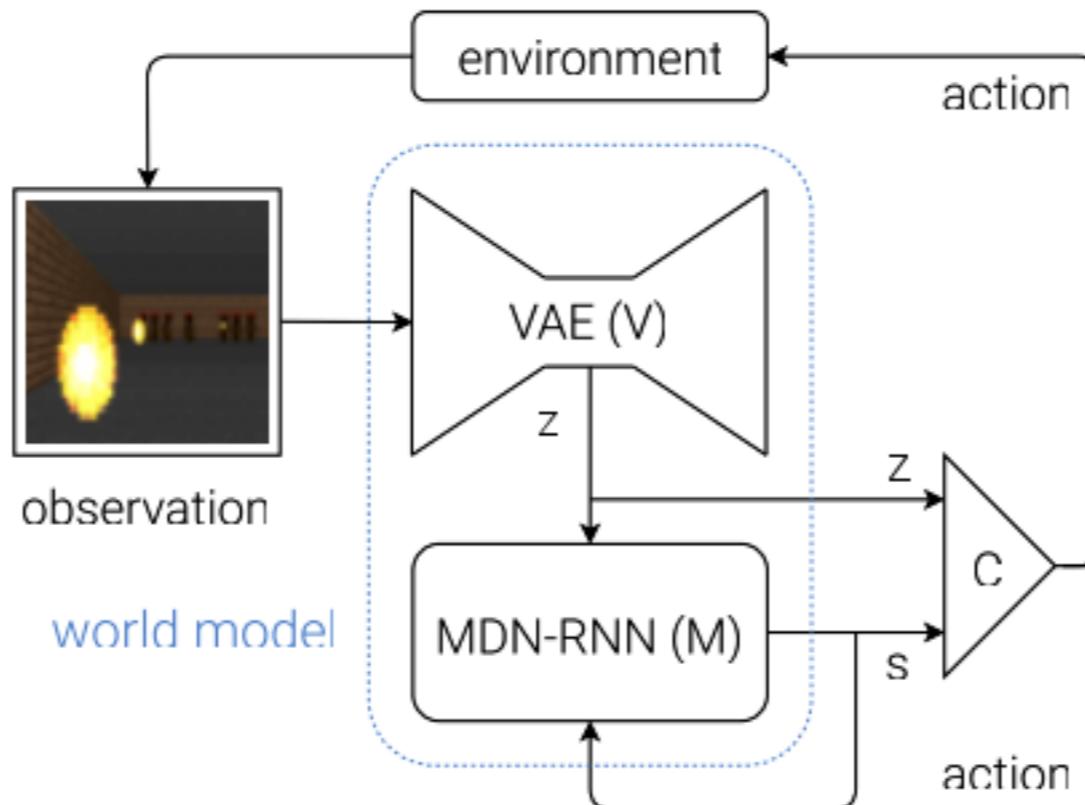
- SimPLe - the authors explore how video prediction models can enable RL agents to solve Atari games with orders of magnitude fewer interactions than model-free methods.
- World Models - a simple latent state-space model inspired by human cognition.
- PlaNet - a purely model-based agent that learns the environment dynamics from images and chooses actions through fast online planning in latent space.



Planning and learning

The experience can improve value functions and policies either directly or indirectly via the transition model.

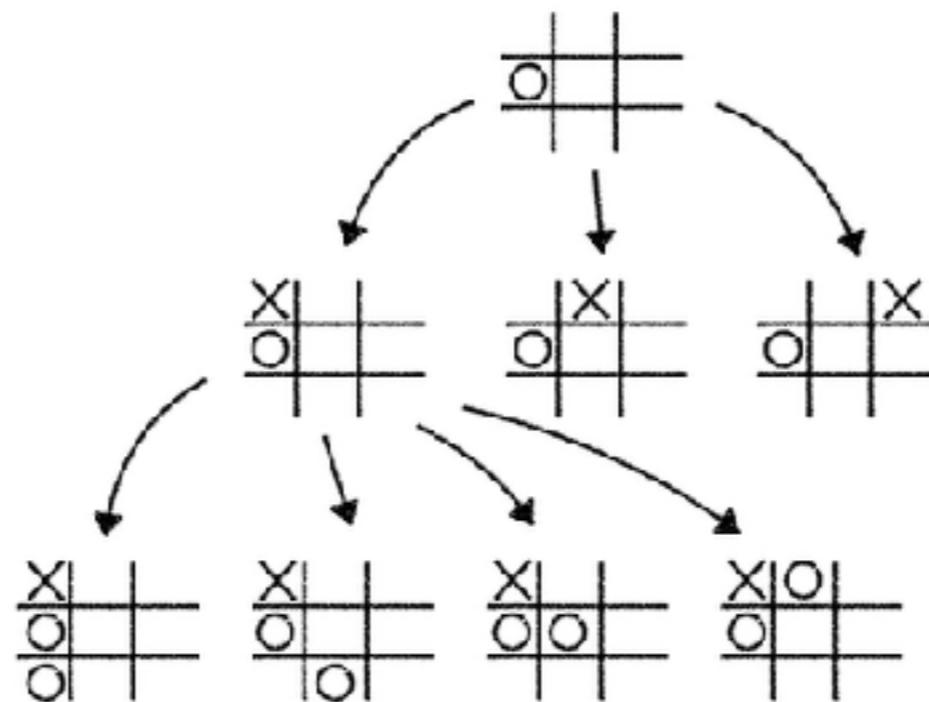
Latent state-space models



Flow diagram of World Models agent model.

Latent state-space models make predictions at a higher level of abstraction, both spatially and temporally, than at the level of raw pixel observations. Moreover, in order to increase model accuracy, they often exploit the benefits of explicitly modeling uncertainty in state transitions.

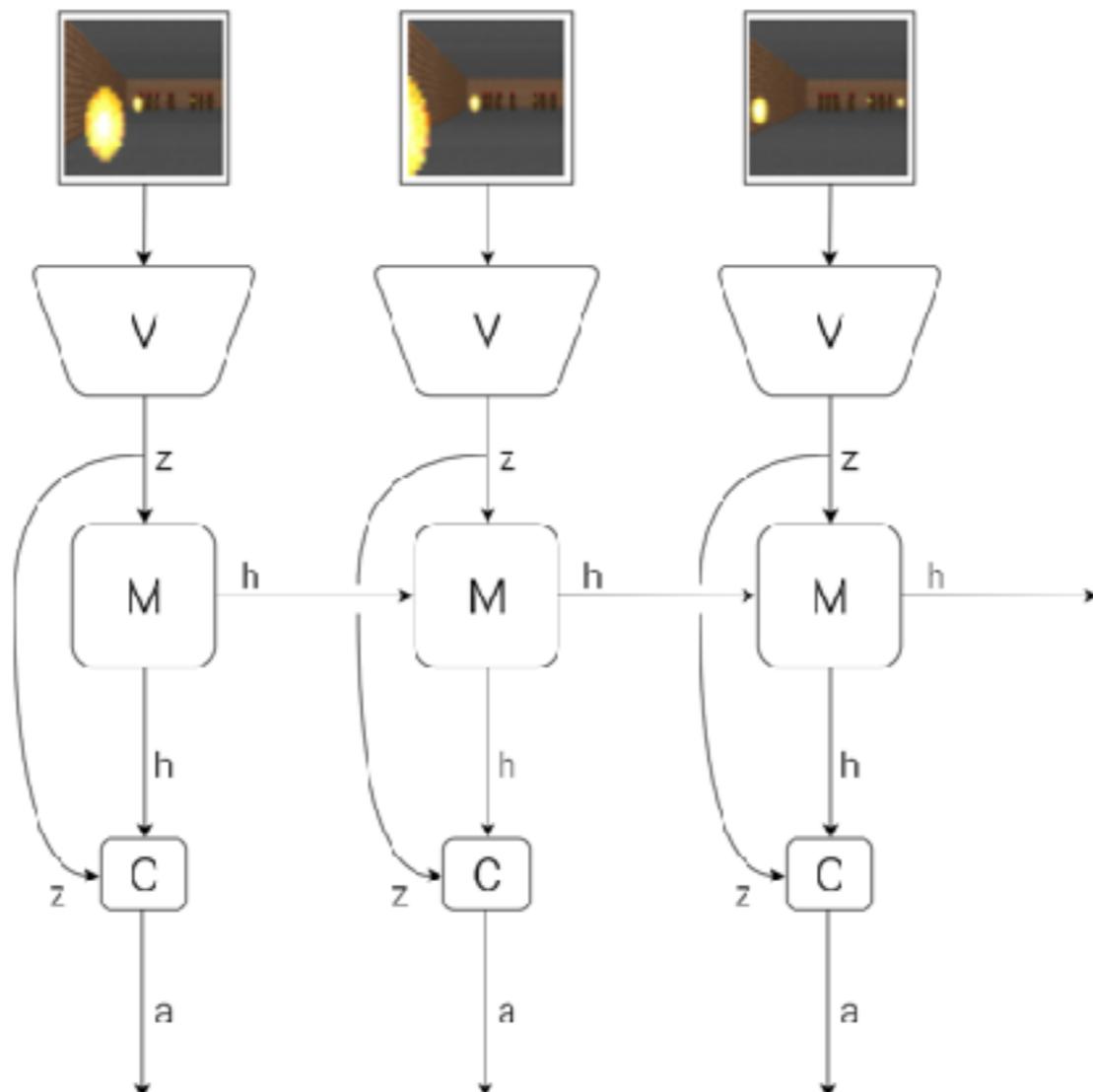
Simulation-based search



- Planning carries the promise of increasing performance just by increasing the computational budget for searching for actions.
- Search exploits temporality, in other words focuses on the current situation, by learning from this specific distribution of future experience, rather than learning from the distribution of all possible experience.

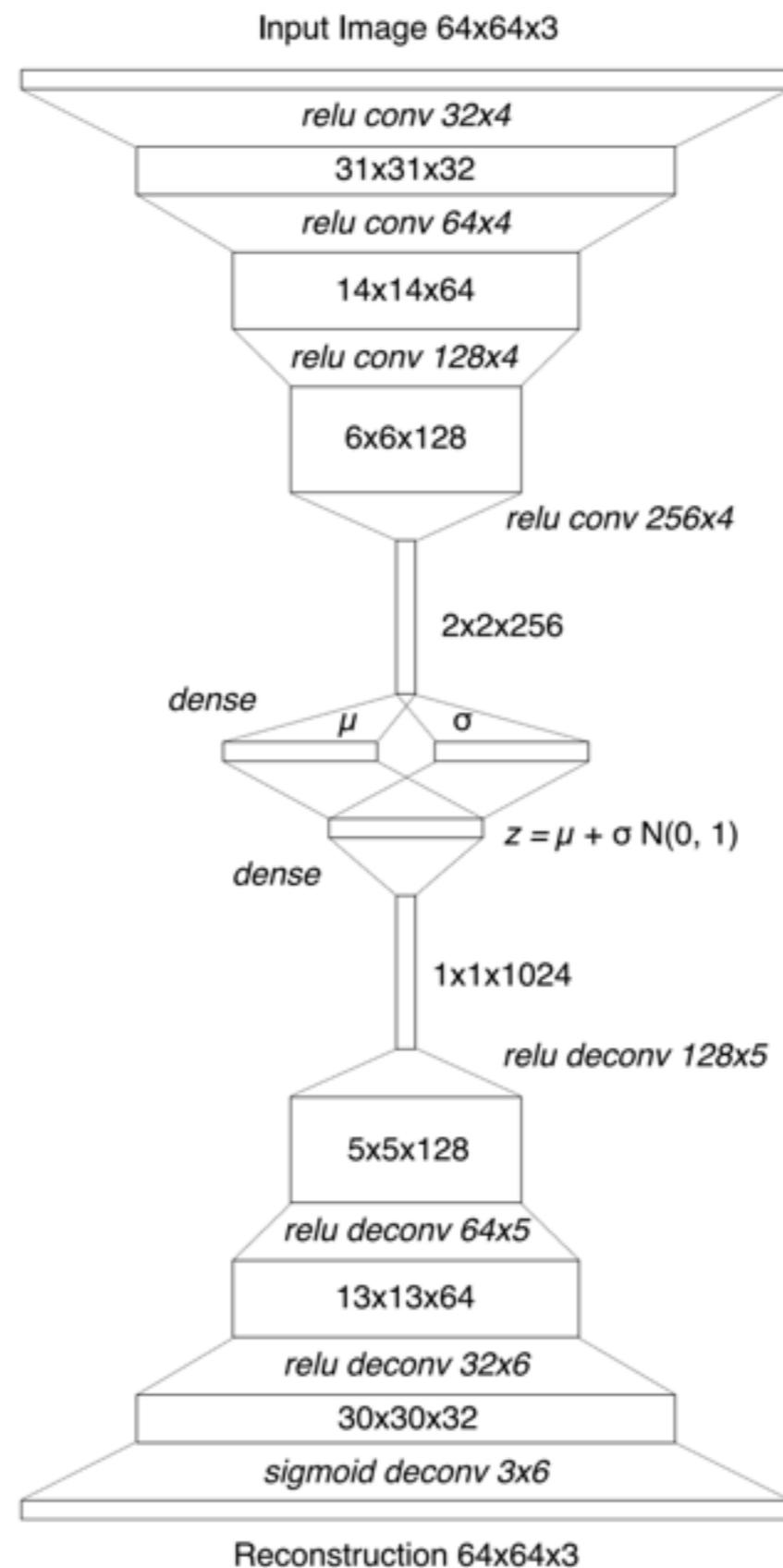
Solutions description and experiments

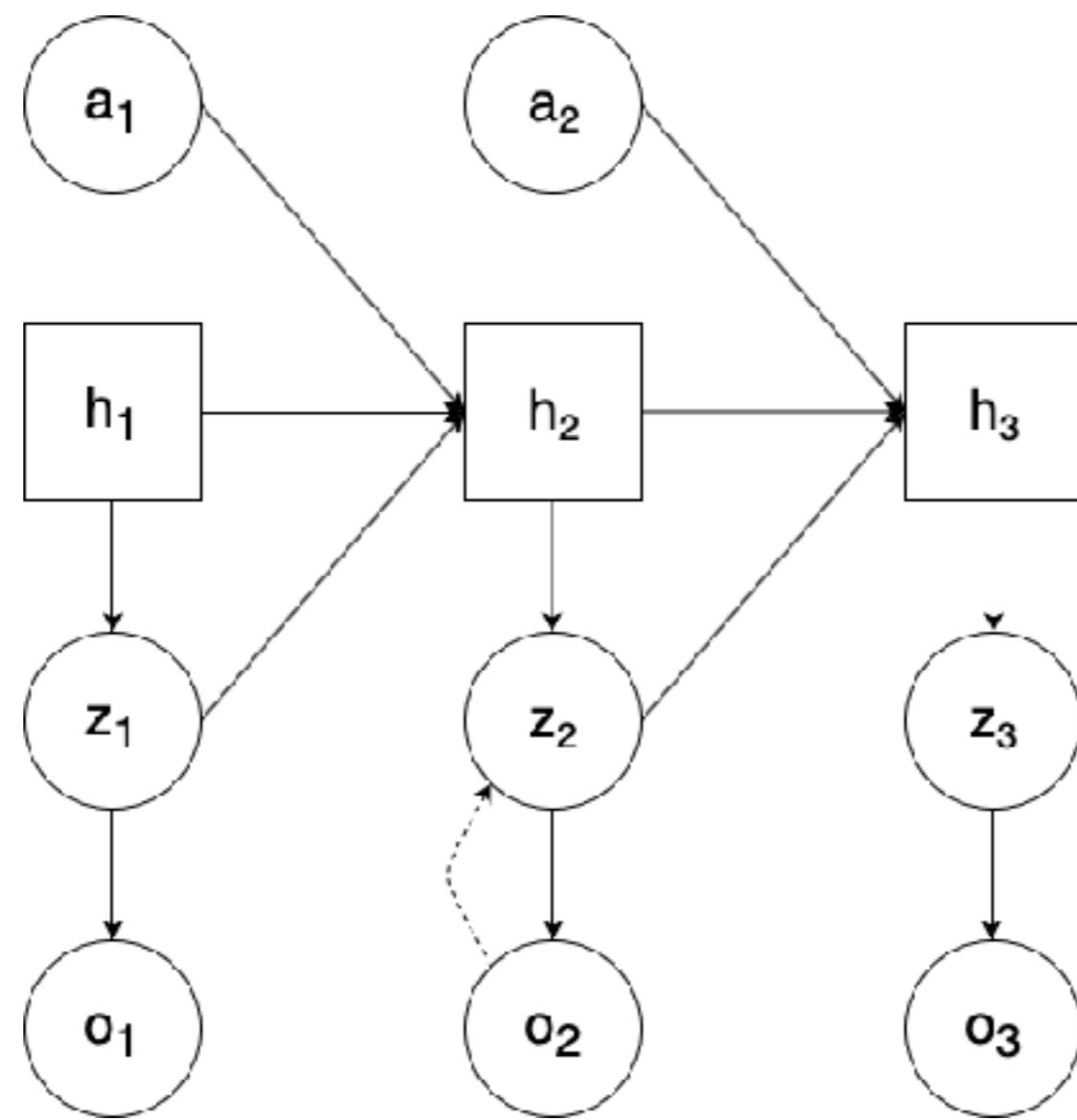
Original World Models (OWM)



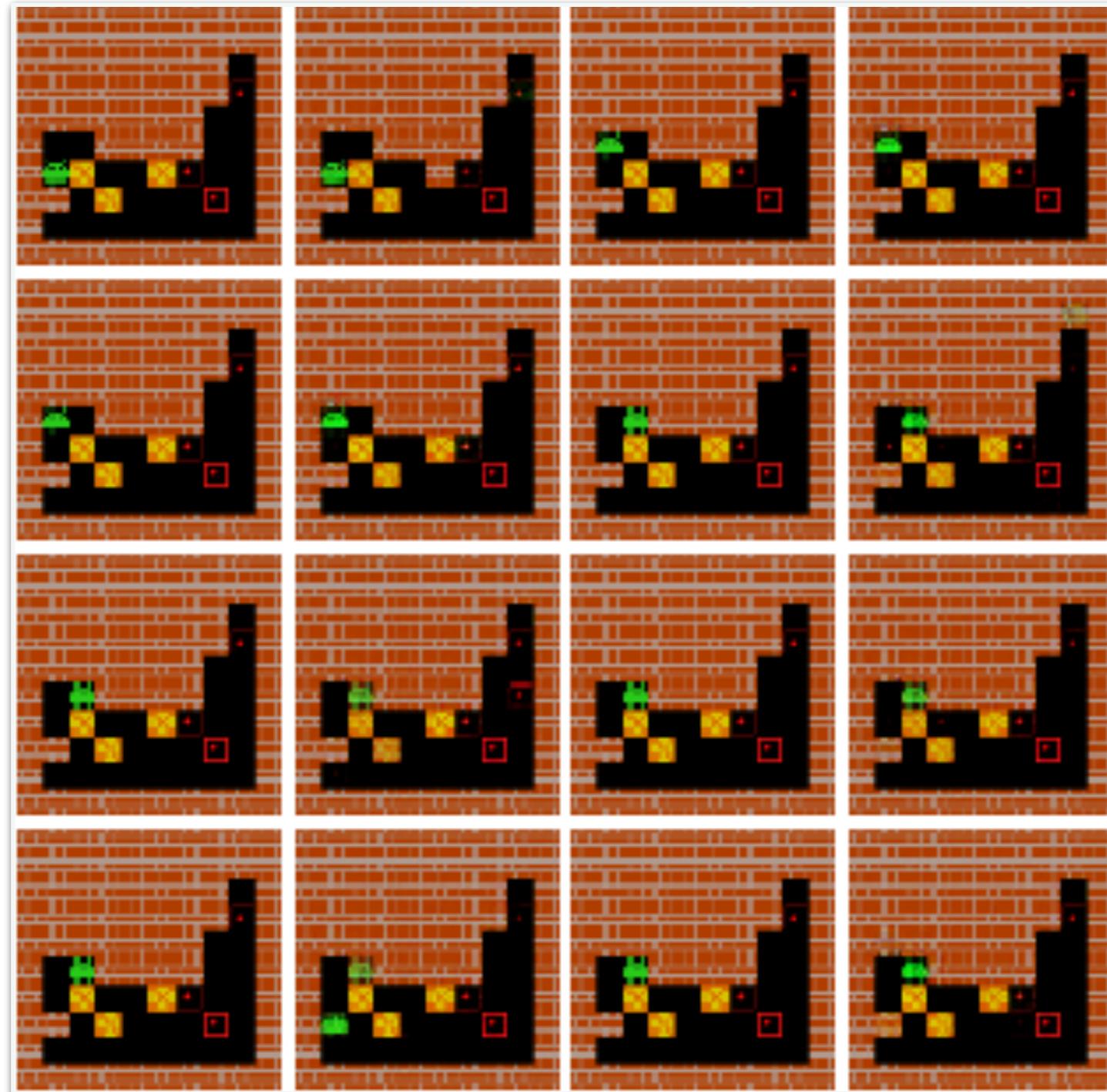
World Models agent consists of three components that work closely together: **Vision**, **Memory**, and **Controller**

vision



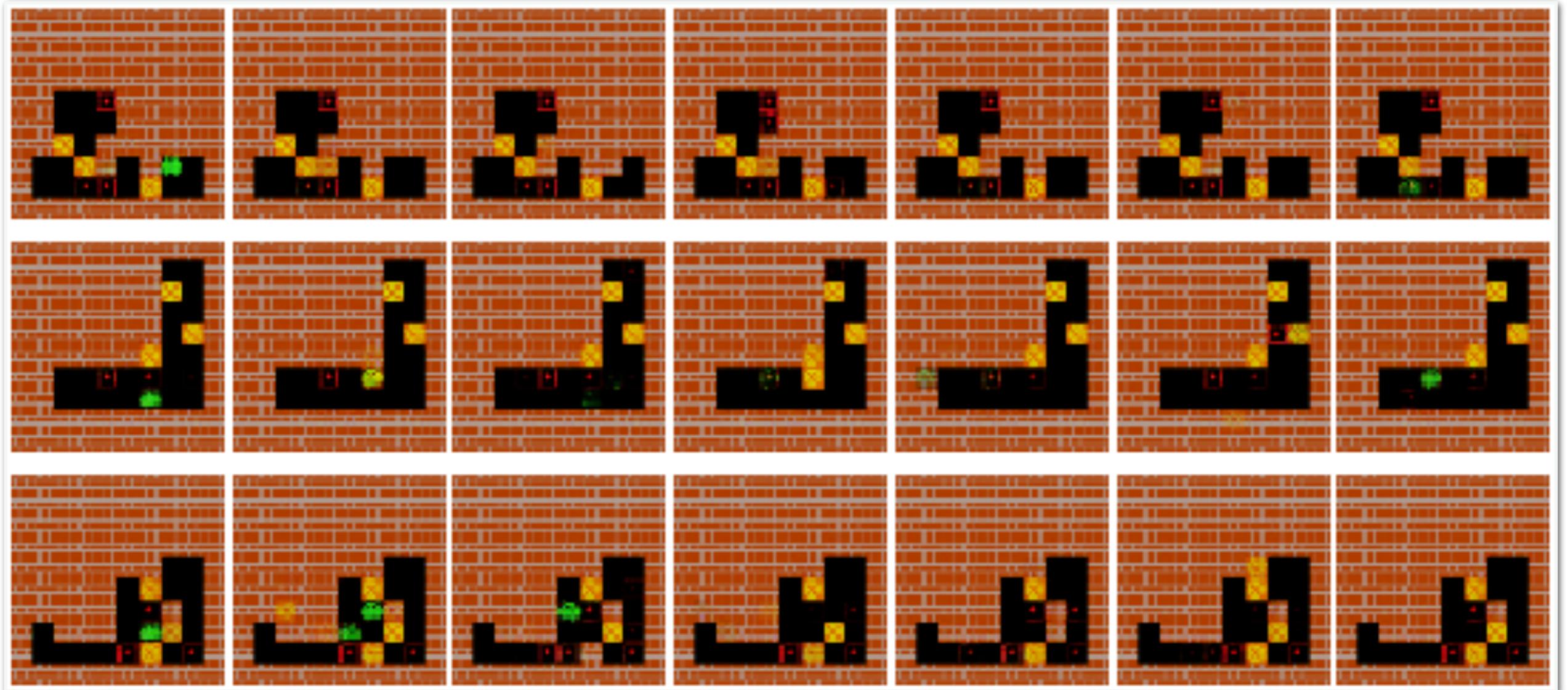


Memory



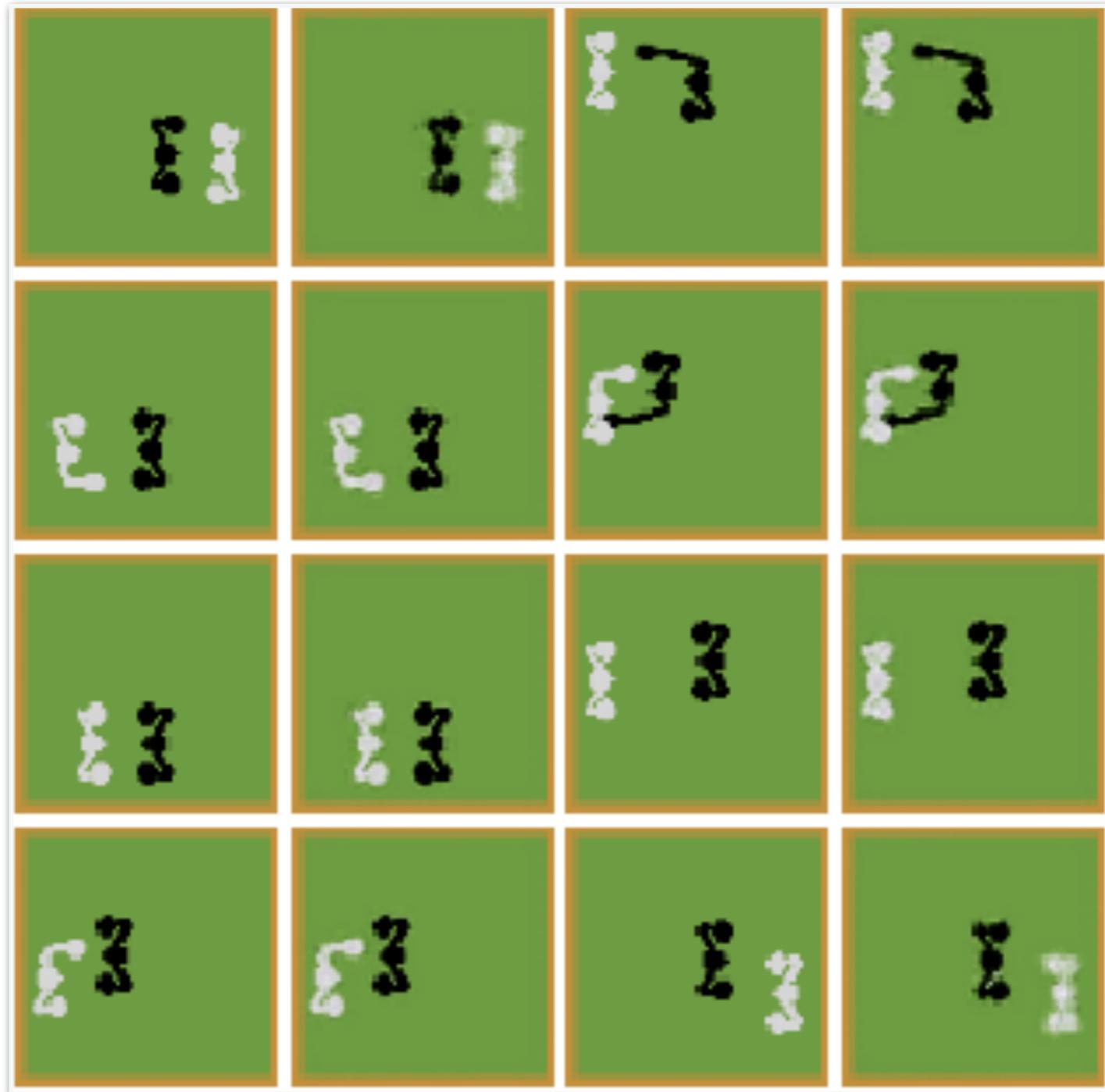
Vision

OWM for Sokoban



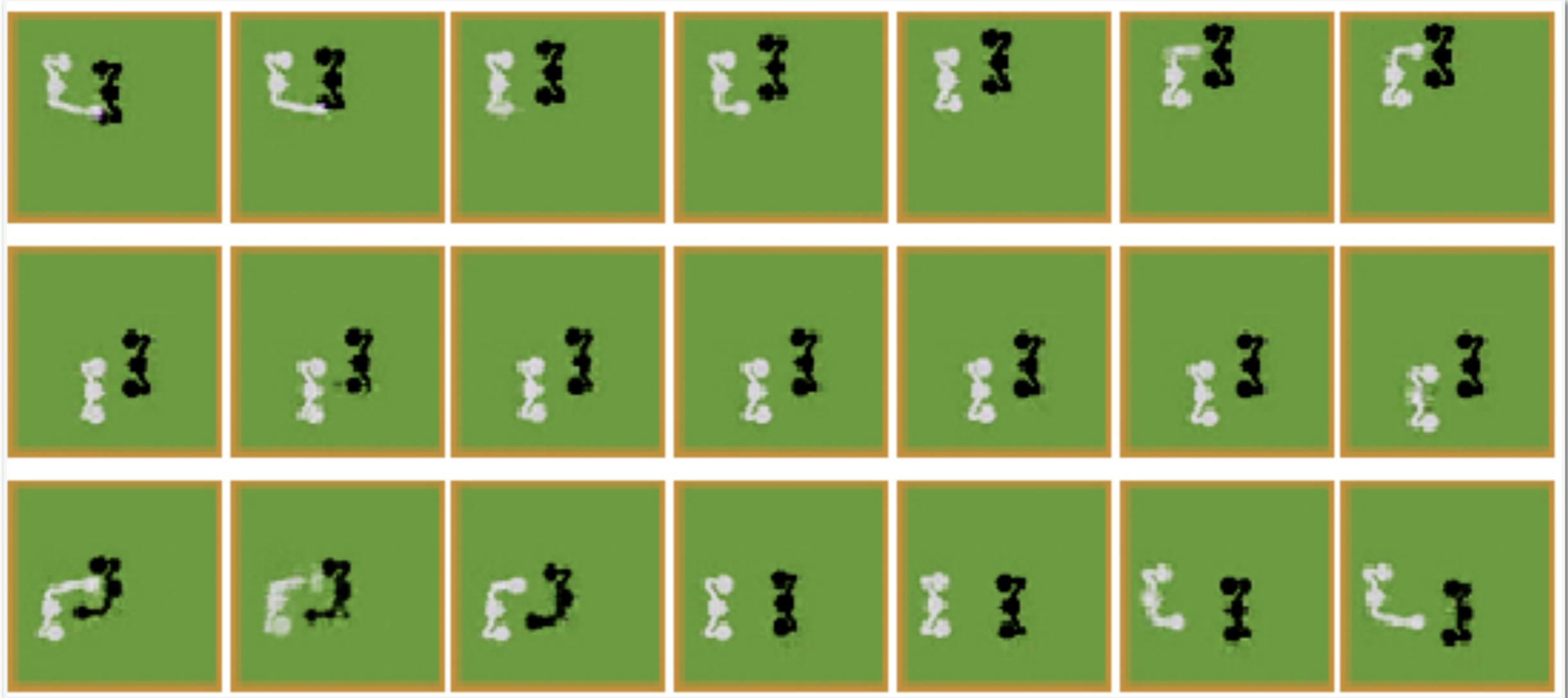
Memory

OWM for Sokoban



Vision

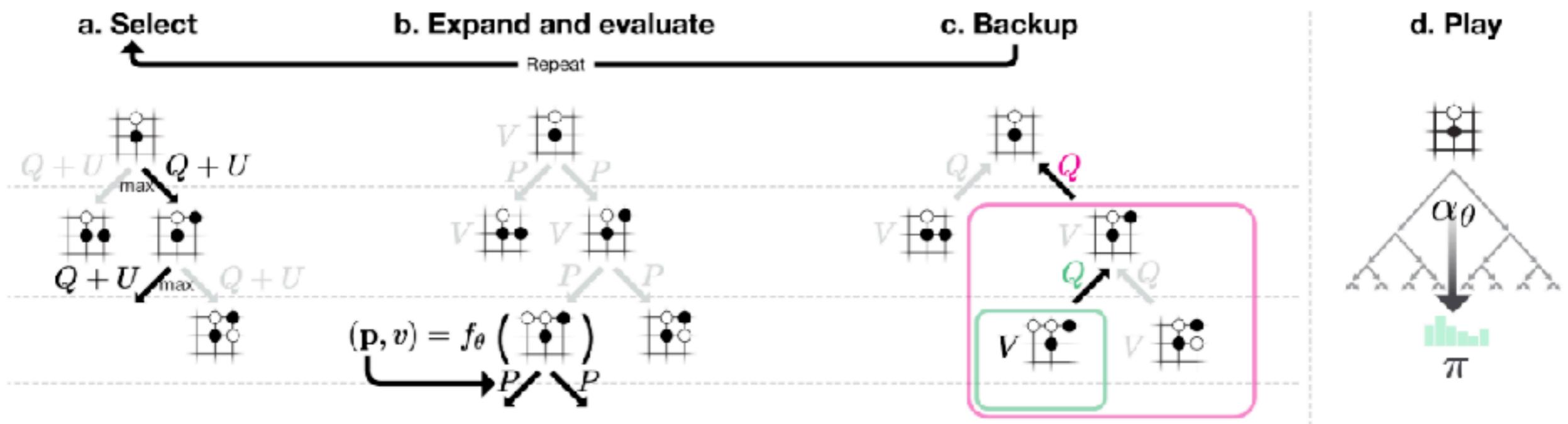
OWM for Boxing



Memory

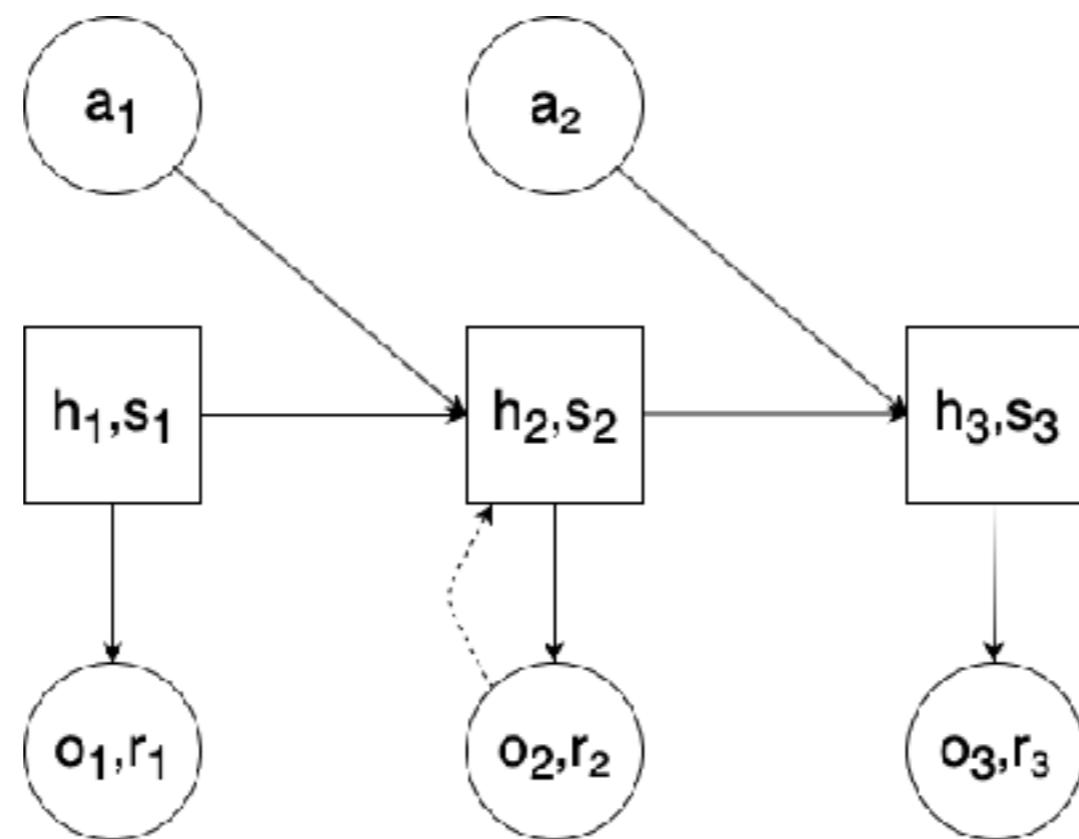
OWM for Boxing

World Models and AlphaZero (W+A)

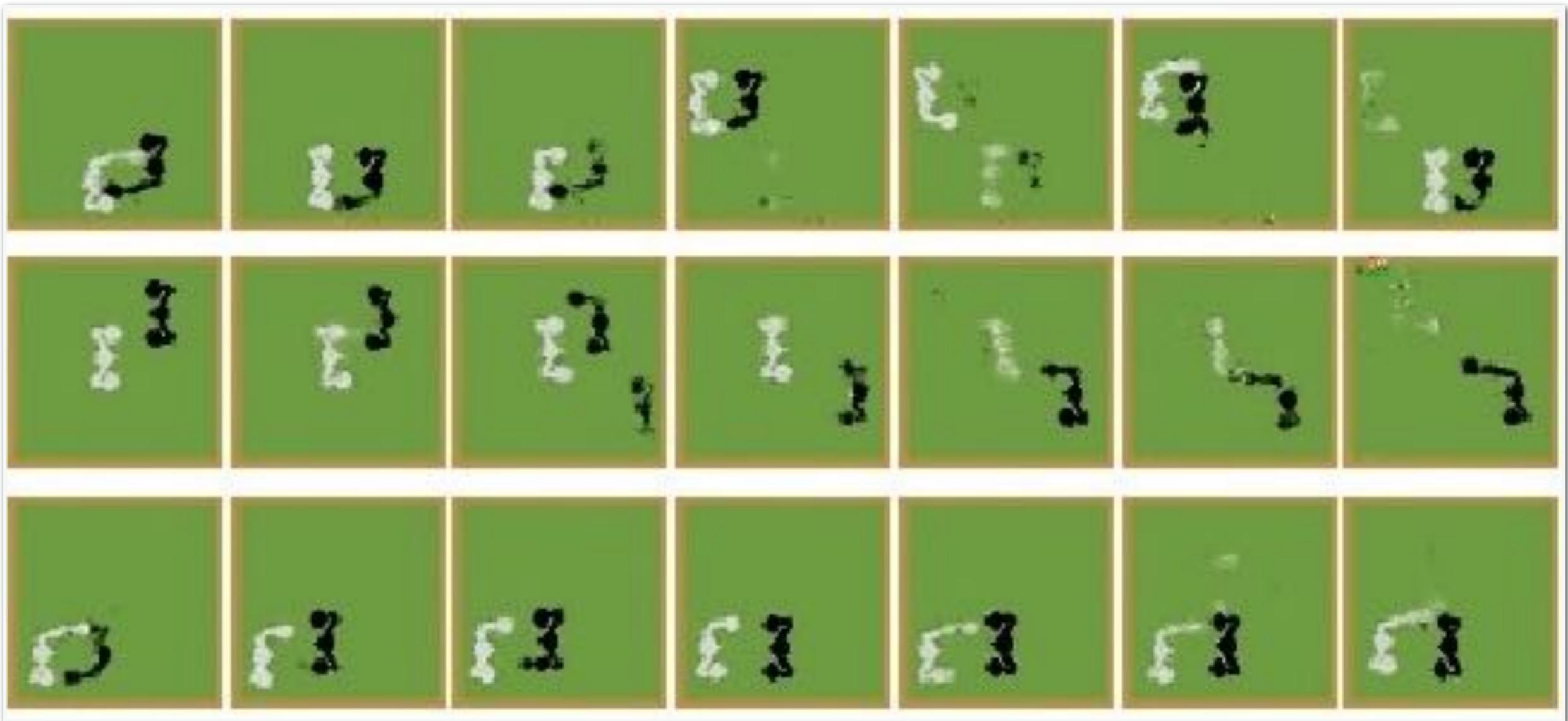


Controller

AlphaZero tree search algorithm



Memory

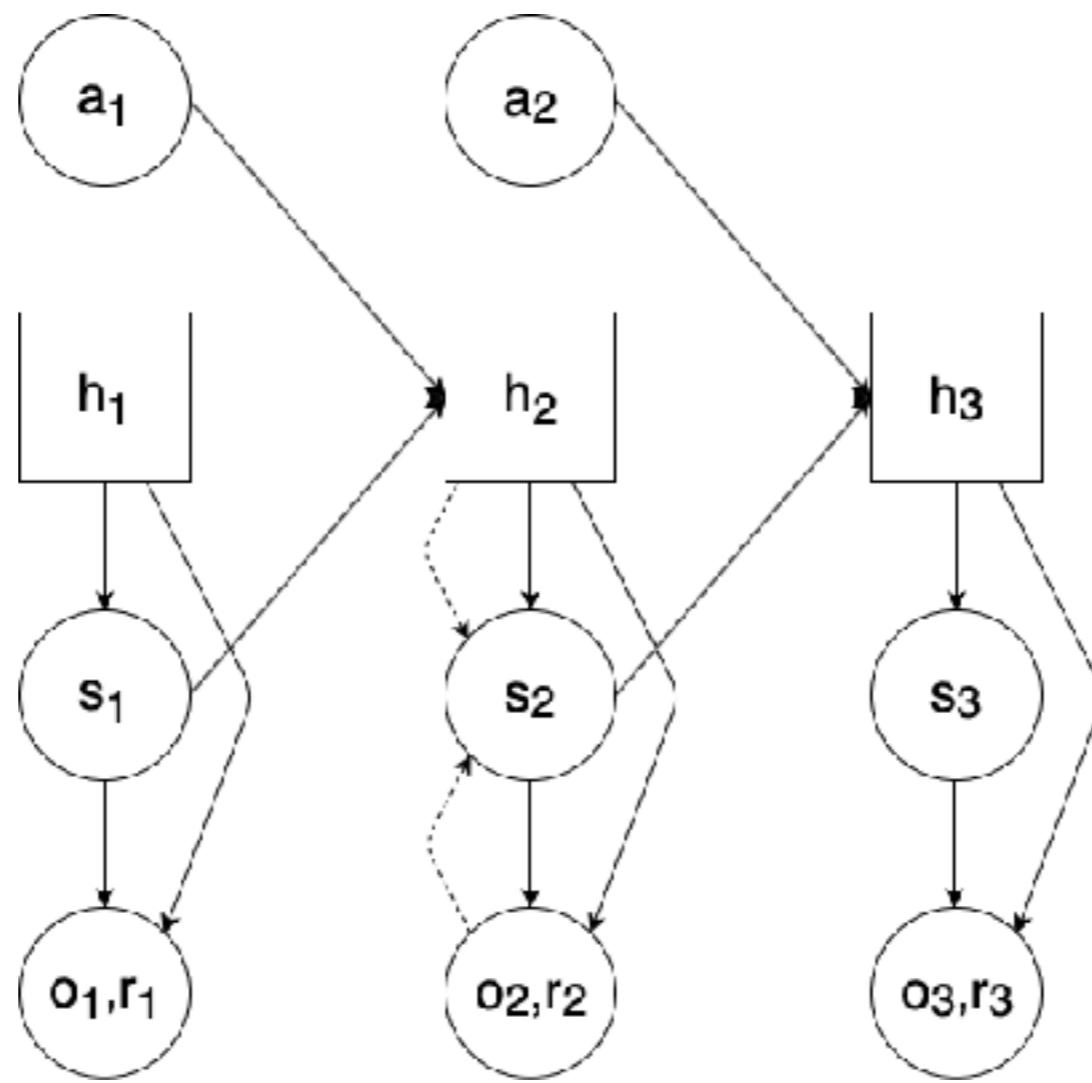


Memory
W+A for Boxing

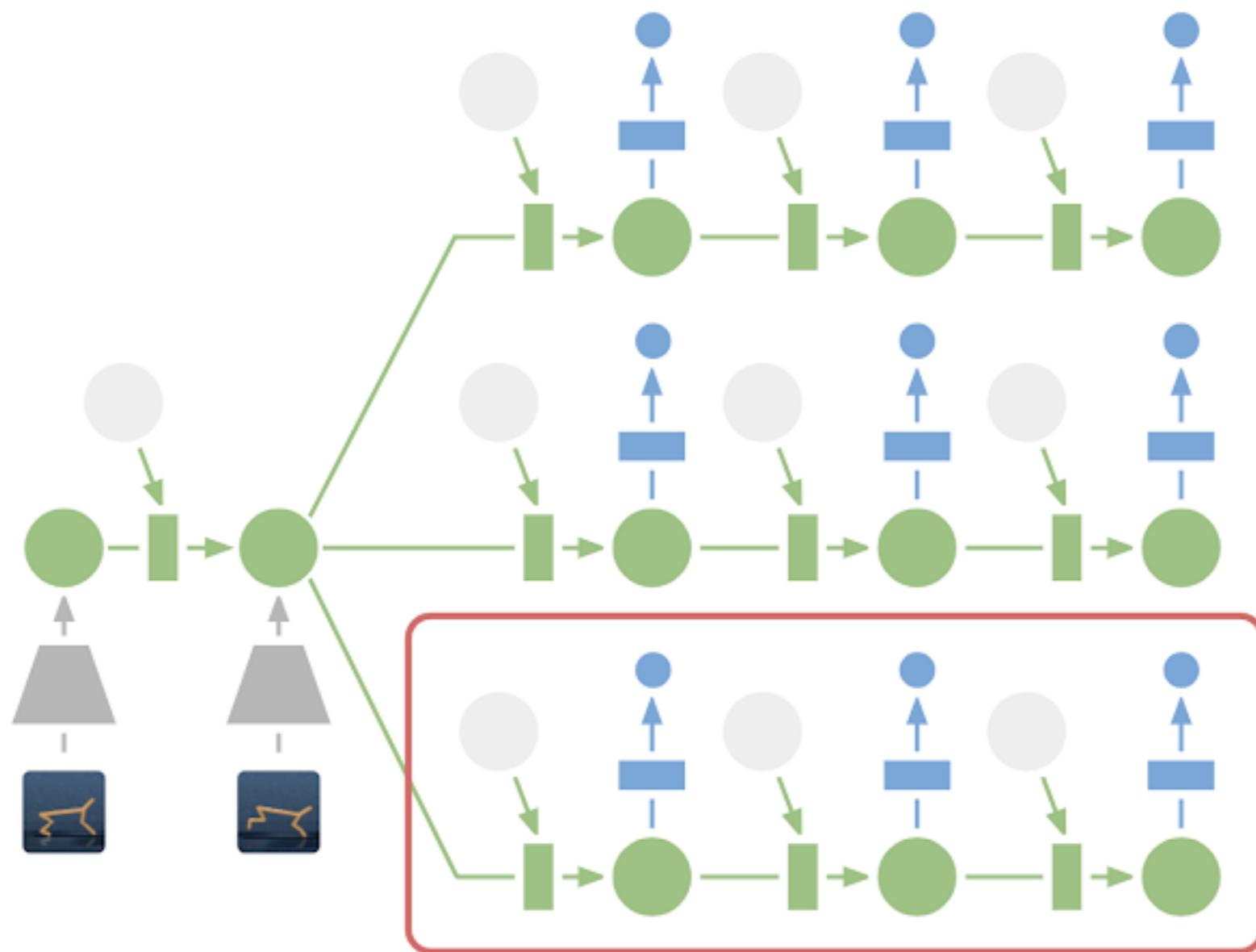
Discrete PlaNet (DPN)



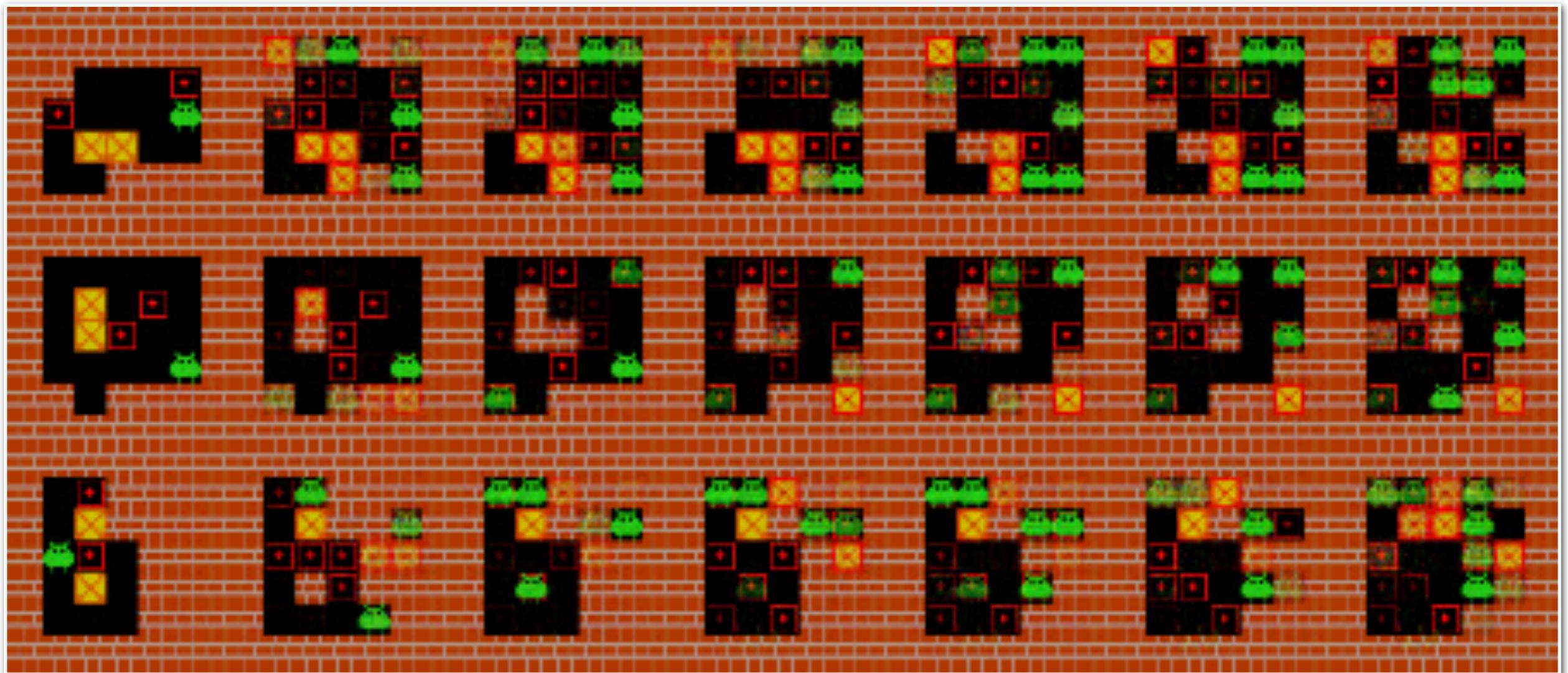
Survival of the fittest.



Memory



Planner



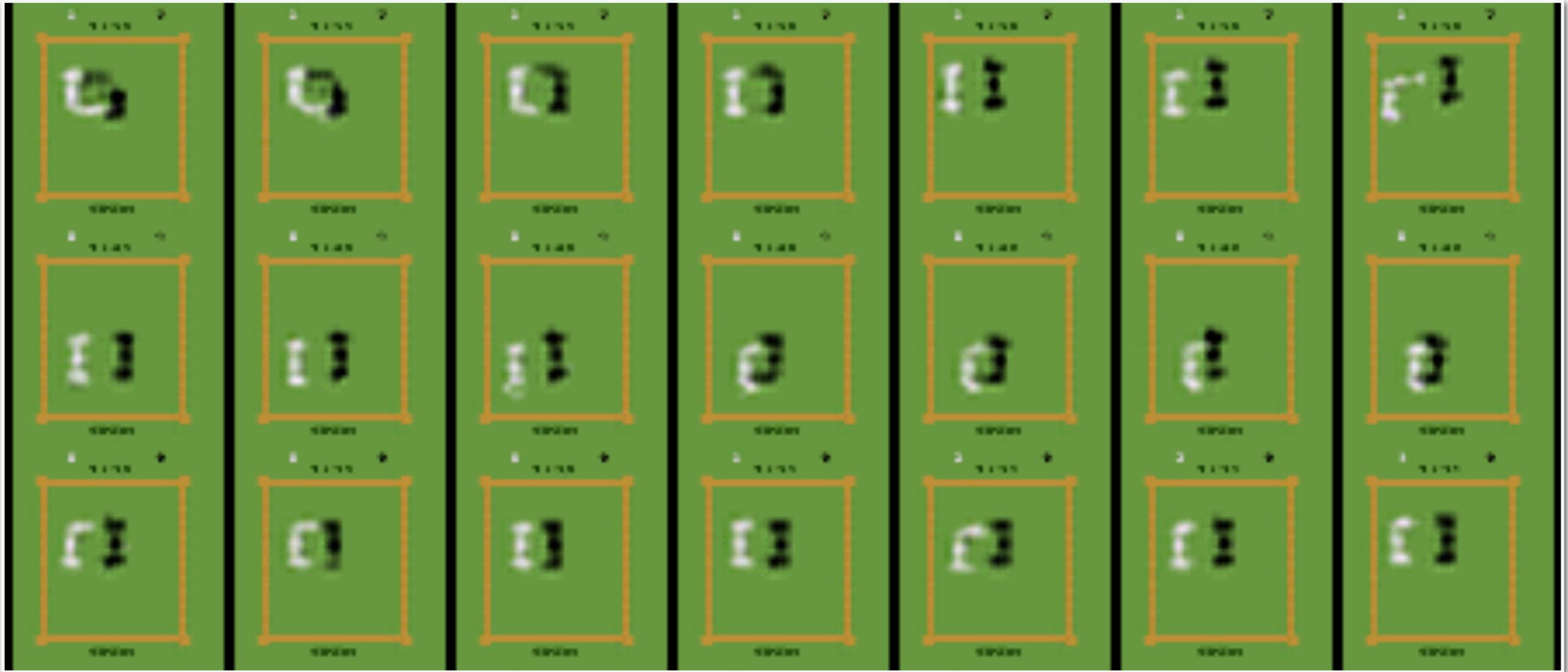
World model

DPN for Sokoban



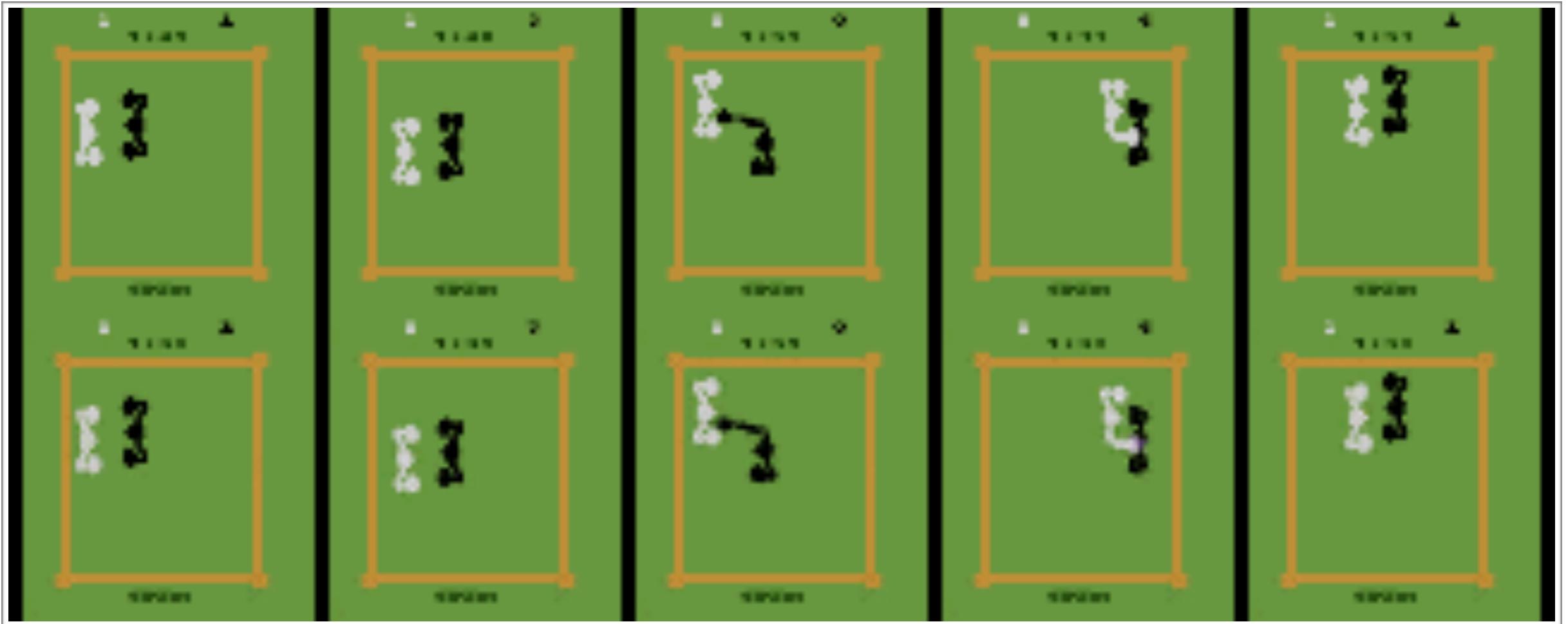
World model

DPN for Boxing



World model

Tuned DPN for Boxing



World model

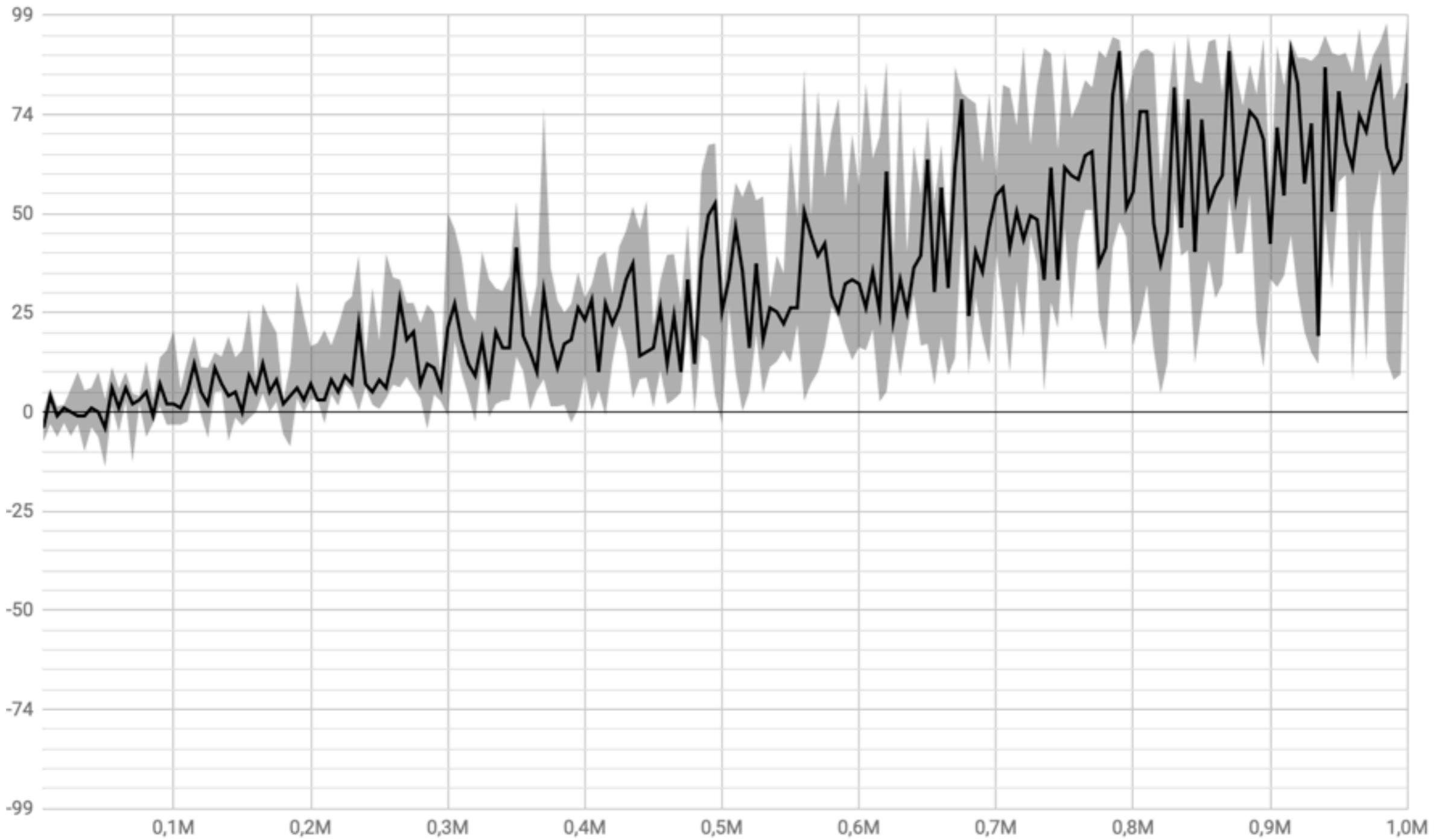
Tuned DPN for Boxing

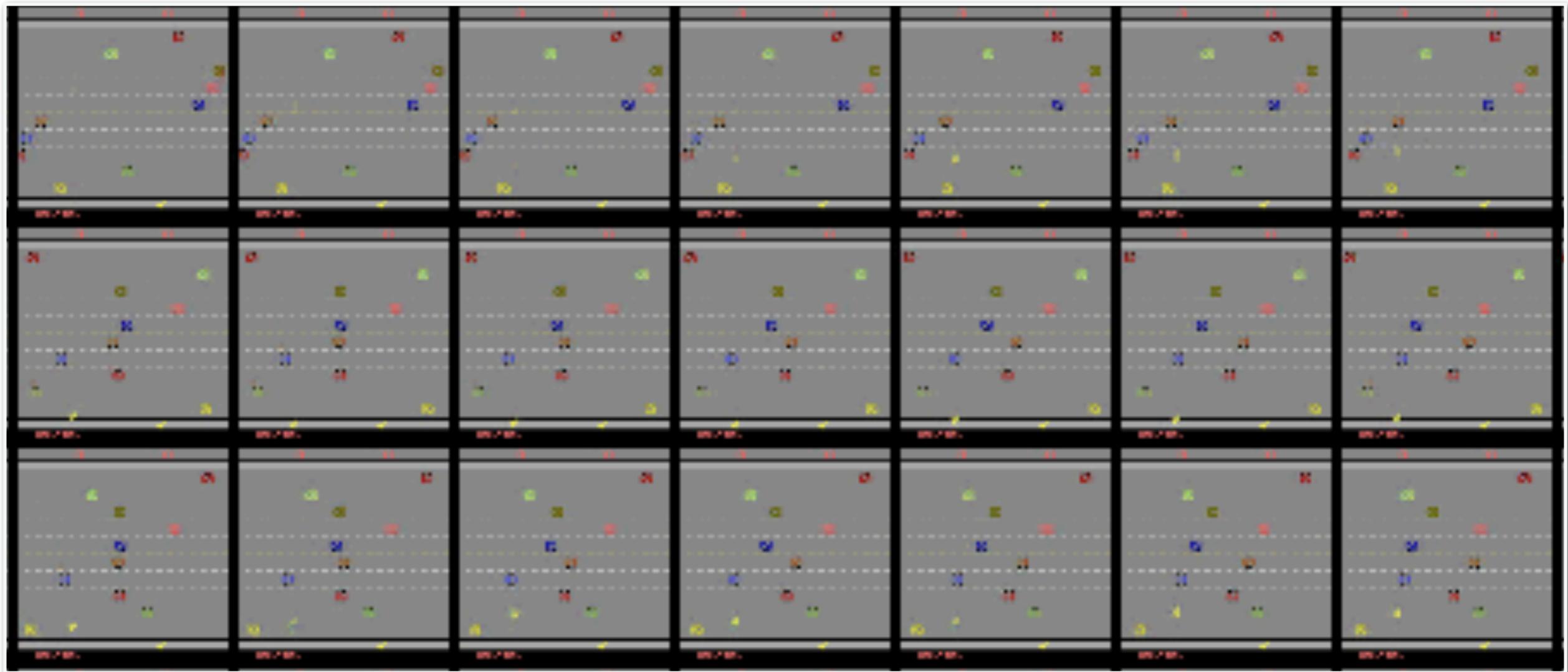


Planner

Tuned DPN for Boxing

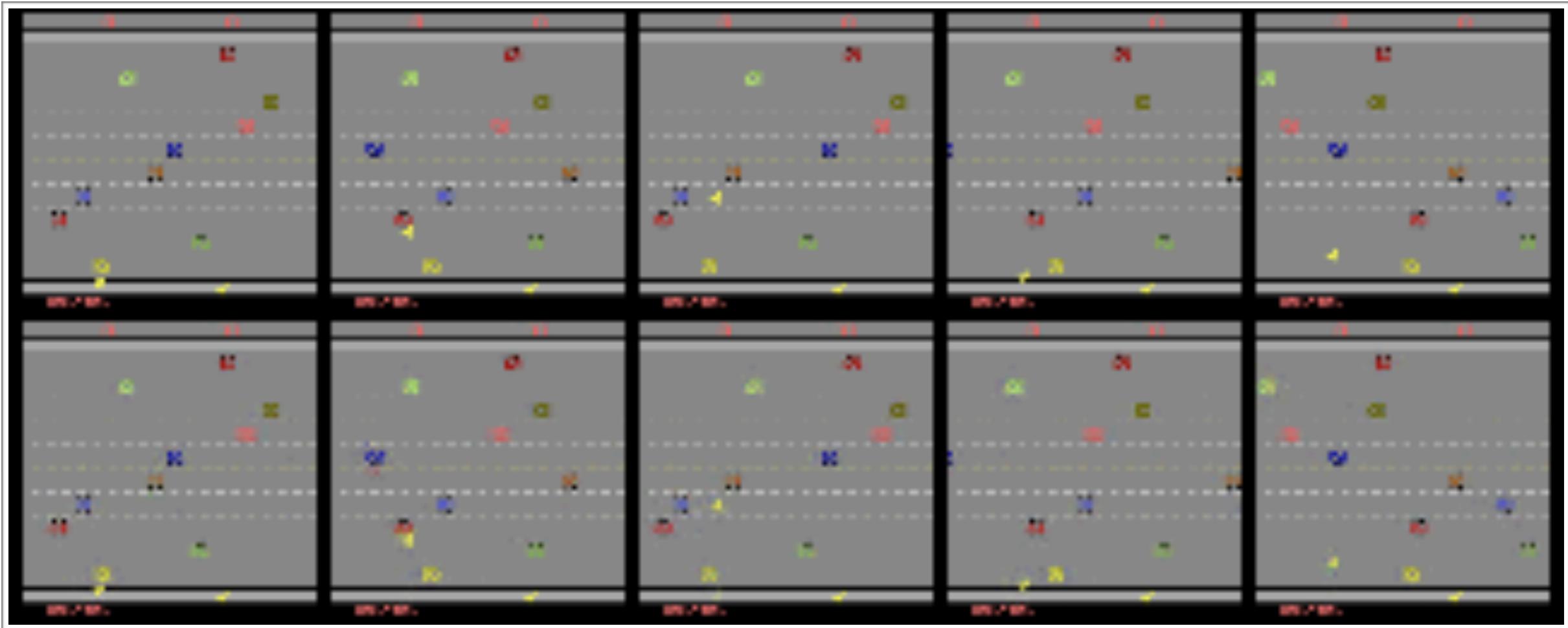
DPN for Boxing





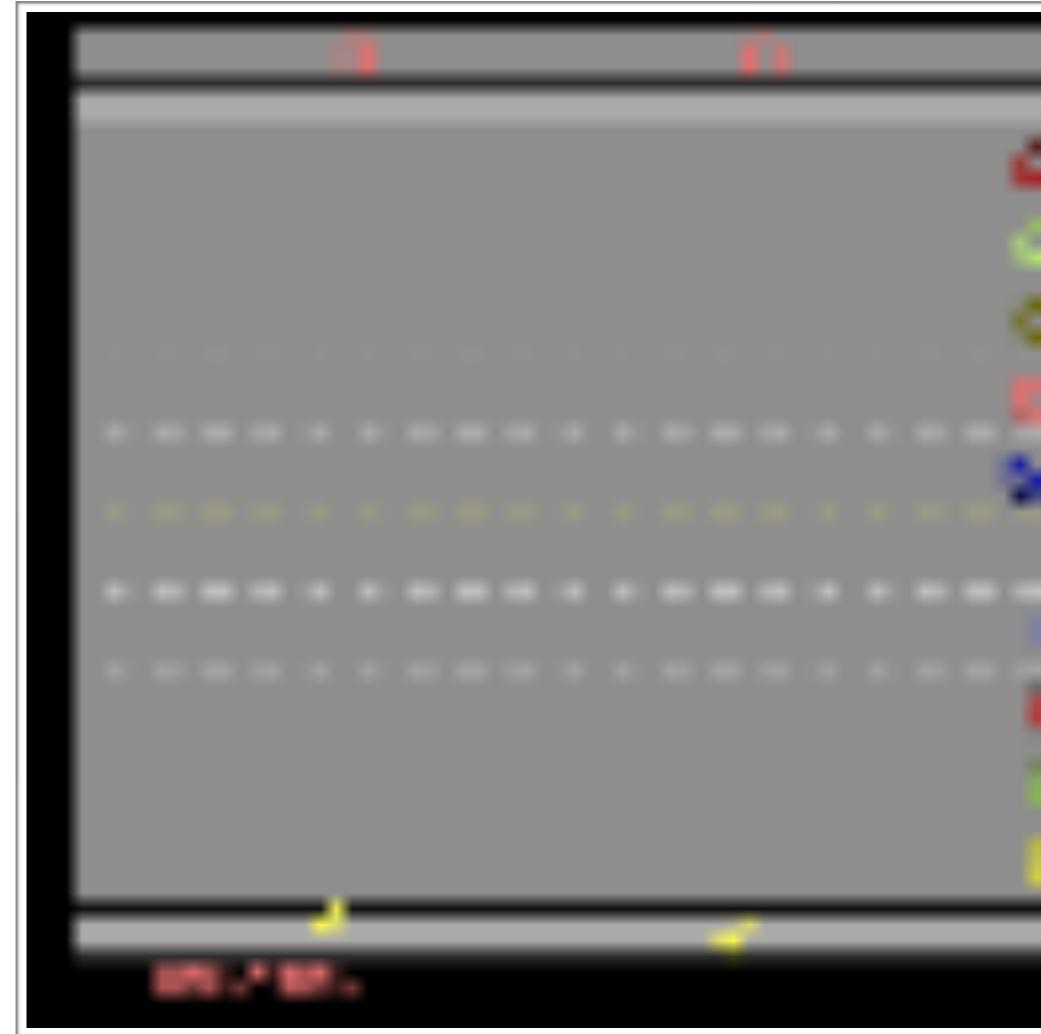
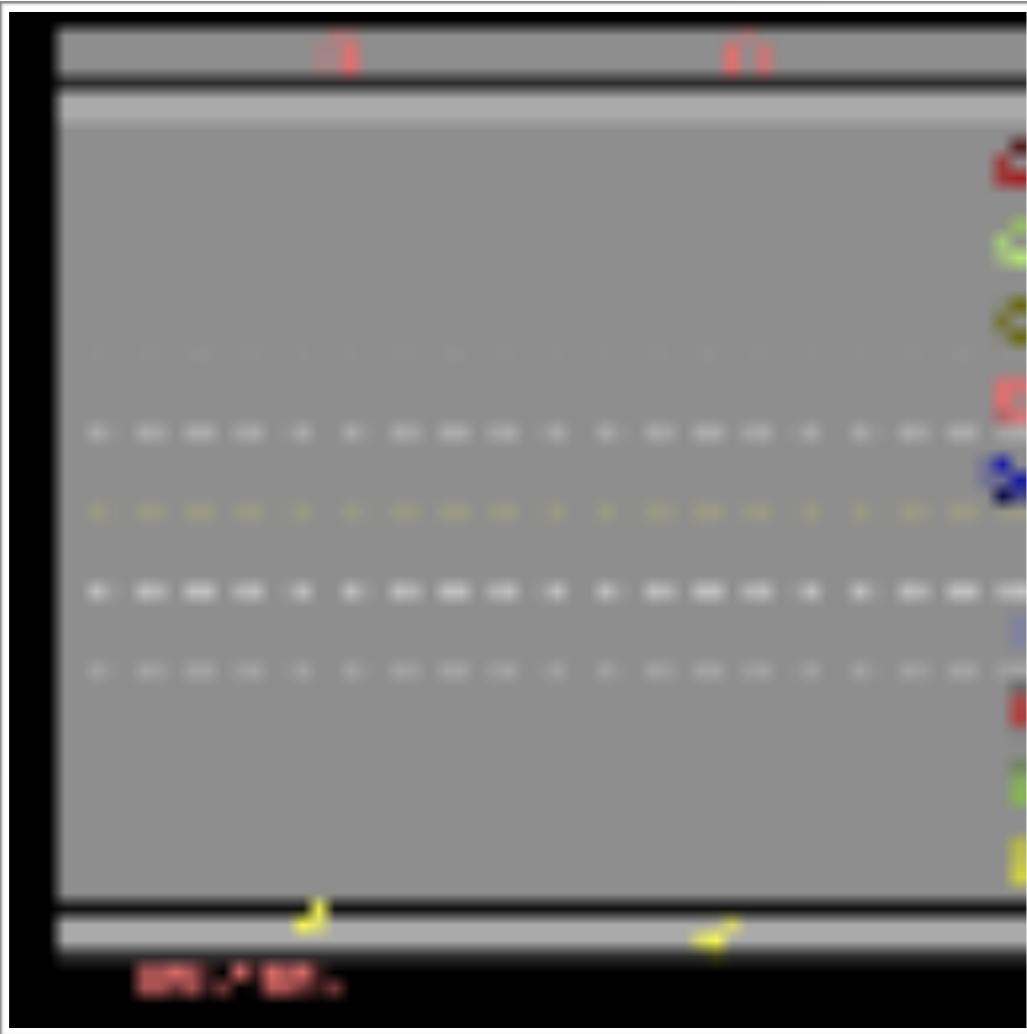
World model

DPN for Freeway



World model

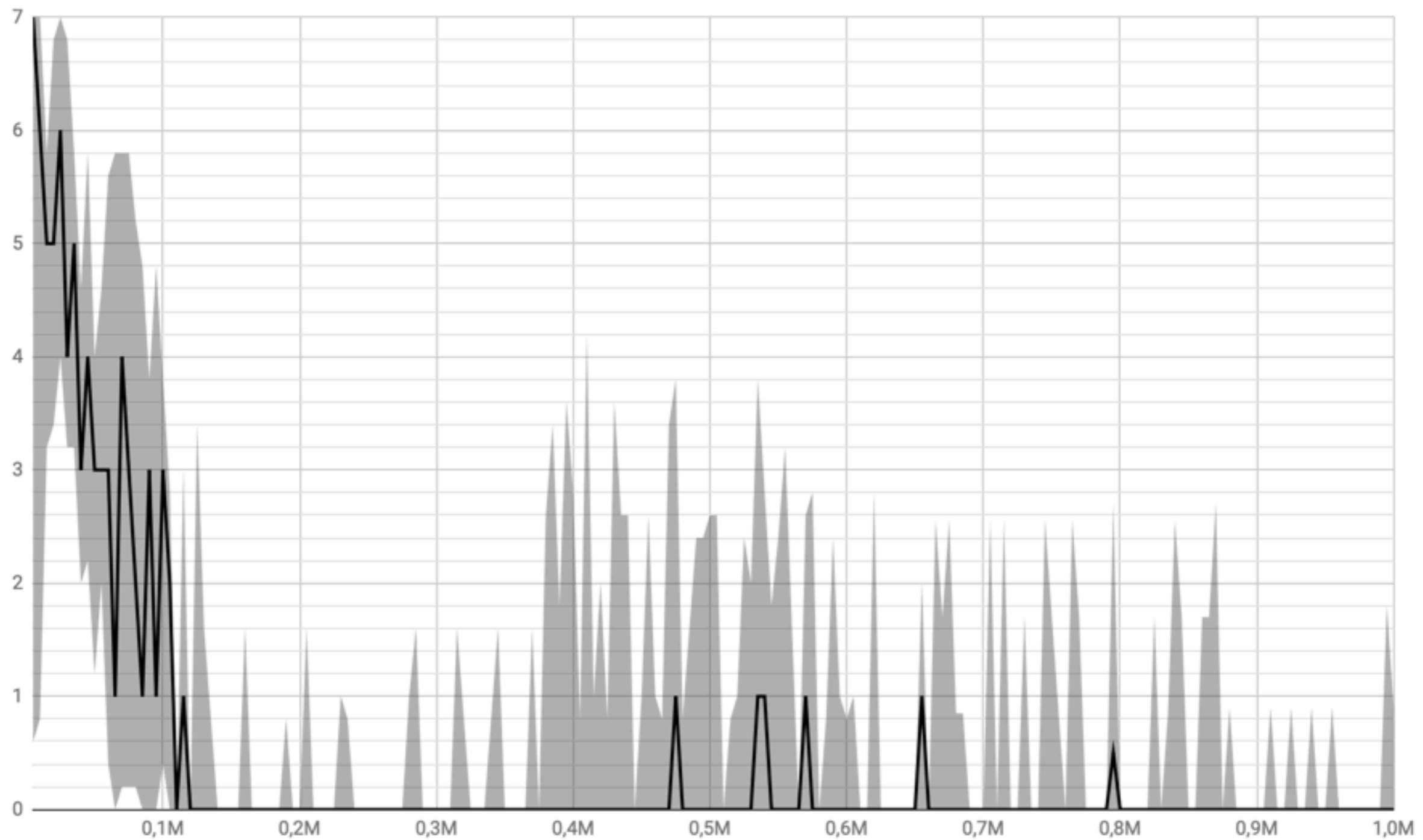
DPN for Freeway



Planner

DPN for Freeway

DPN for Freeway



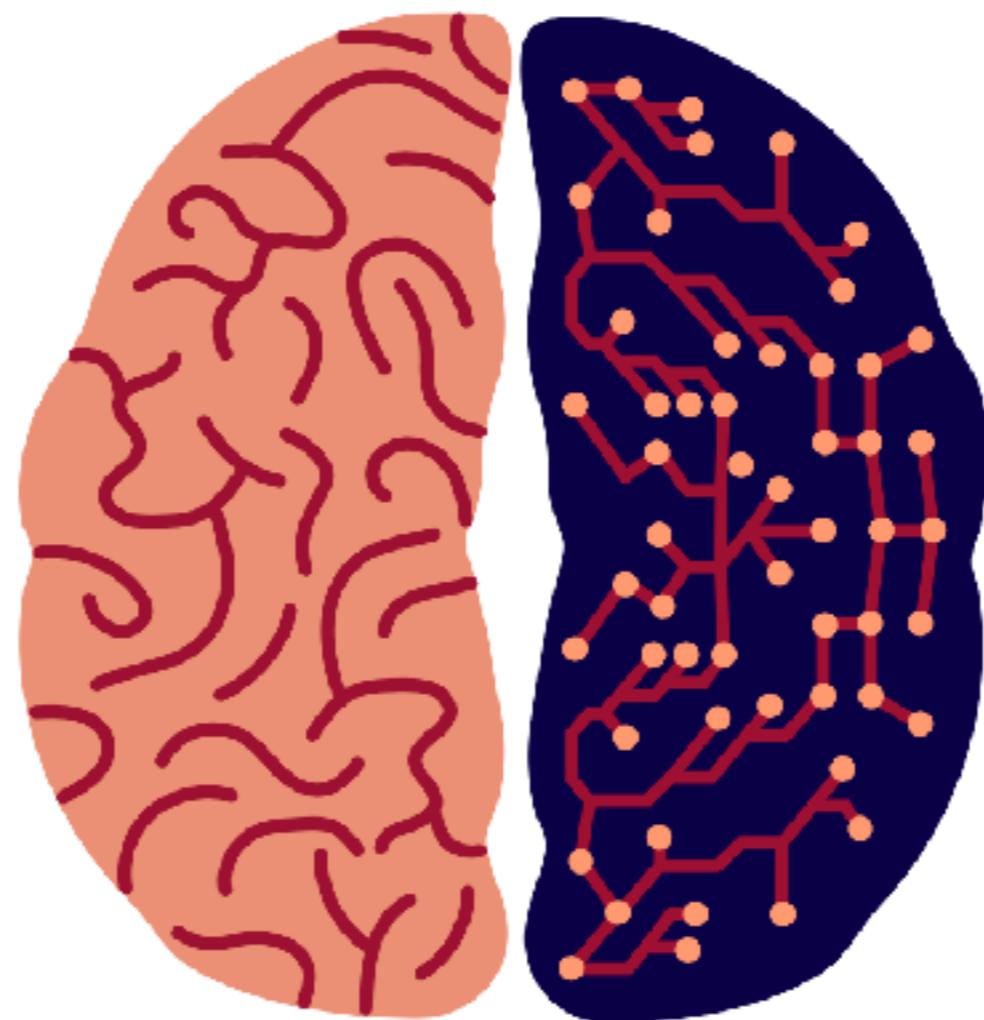
Final results for Boxing

Algorithm	100K	500K	1M
Ours	6,2 (10,7)	35,2 (8,3)	78,2 (19,1)
SimPLe	9,1 (8,8)	NDA	NDA
PPO	-3,9 (6,4)	3,5 (3,5)	19,6 (20,9)
Rainbow	0,9 (1,7)	58,2 (16,5)	80,3 (5,6)
Random		0.3	

Mean scores and standard deviations (in brackets) over five training runs of DPN in Boxing.

Conclusions and future work

Conclusions



Source: <https://course.elementsofai.com/>

Future work

- Future work could focus on extending this method to other challenging tasks like: sparse rewards environments, i.e. Freeway, and complex puzzle games with massive state-space sizes, i.e. Sokoban.
- DPN, unlike model-based baseline SimPLe, hold promise of increased performance with an increased computational budget for planning. This hypothesis could be put to the extensive test too.
- Furthermore, generalisation of the DPN world model to different tasks in the same or very similar environments could be explored.

Thank you for your attention

Questions?