

# We improve buffer sampling with uncertainty selection using bootstrap Neural Networks.

## MEET: A Monte Carlo Exploration-Exploitation Trade Off for Buffer Sampling

Julius Ott, 

Lorenzo Servadei, Jose Arjona-Medina, Enrico Rinaldi, Gianfranco Mauro, Daniela Sánchez-Lopera, Michael Stephan, Thomas Stadelmayer, Avik Santra, Robert Wille


### Background (Fig. 1)

- Off-Policy Reinforcement Learning (RL) stores and samples from a buffer to improve the policy.
- More relevant transitions should be sampled more often. How to determine relevance?
- Prioritized Sampling uses the TD error to estimate importance.

### Methods (Fig. 2)

- The Critic Neural Network uses multiple heads to predict the expected Q-Value. The active heads are selected at random.
- MEET uses the variation along the network heads to estimate the uncertainty in the prediction, and the average along the heads to estimate the expected performance.
- Samples with high uncertainty are useful for training (active learning).
- Samples with low performance and low uncertainty are harmful for training (local optimum).
- Importance on the uncertainty first and performance afterwards.

### Results

- MEET improves the stability and performance of Off-Policy RL algorithms (Soft Actor-Critic). (Fig. 3). 
- MEET outperforms the Prioritized Experience Replay and Uniform Replay.
- Prioritized replay can be harmful for training.

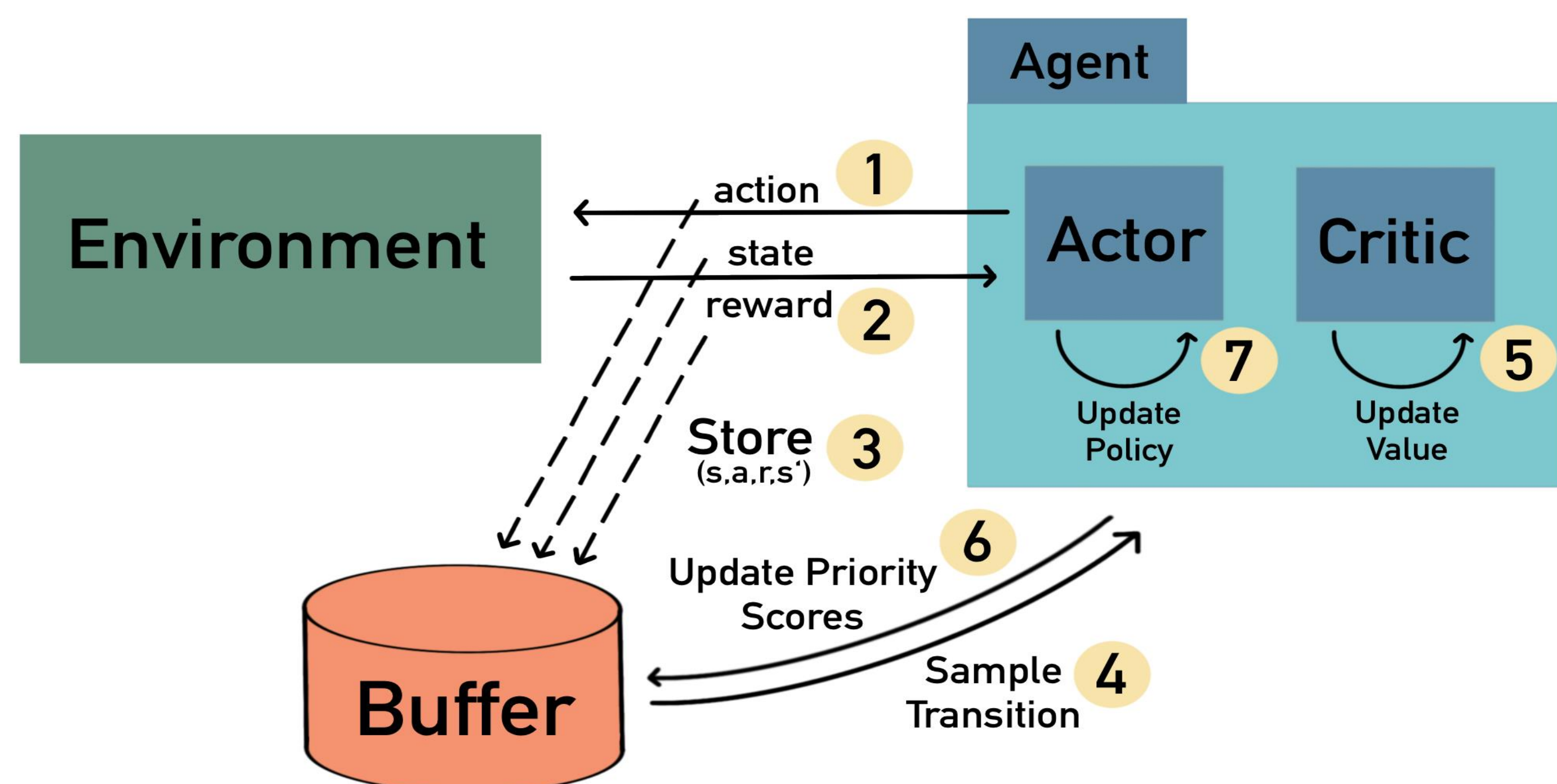


Fig. 1 : Building Blocks of Off-Policy Reinforcement Learning

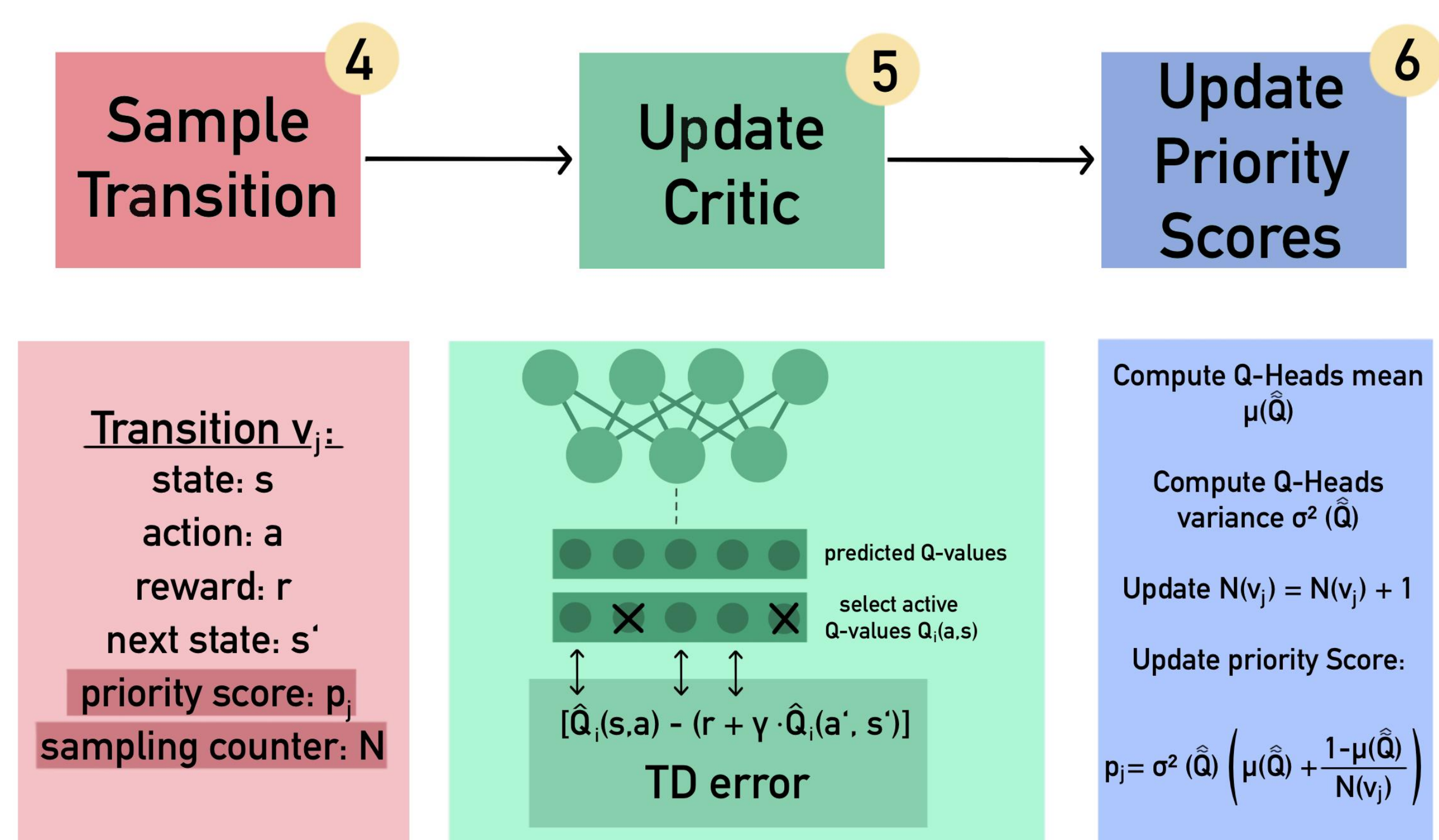


Fig. 2 : Methodology of MEET with bootstrap Neural Networks.

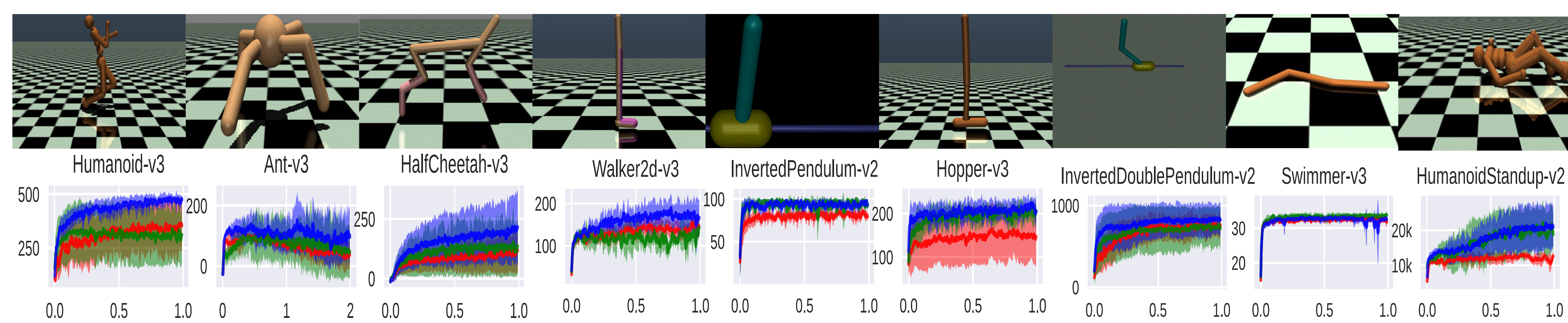


Fig. 3 : Evaluation of MEET (blue) against Uniform (green) and Prioritized (red) sampling on the MuJoCo benchmarking suite.

