Robustness Certification of Deep Learning

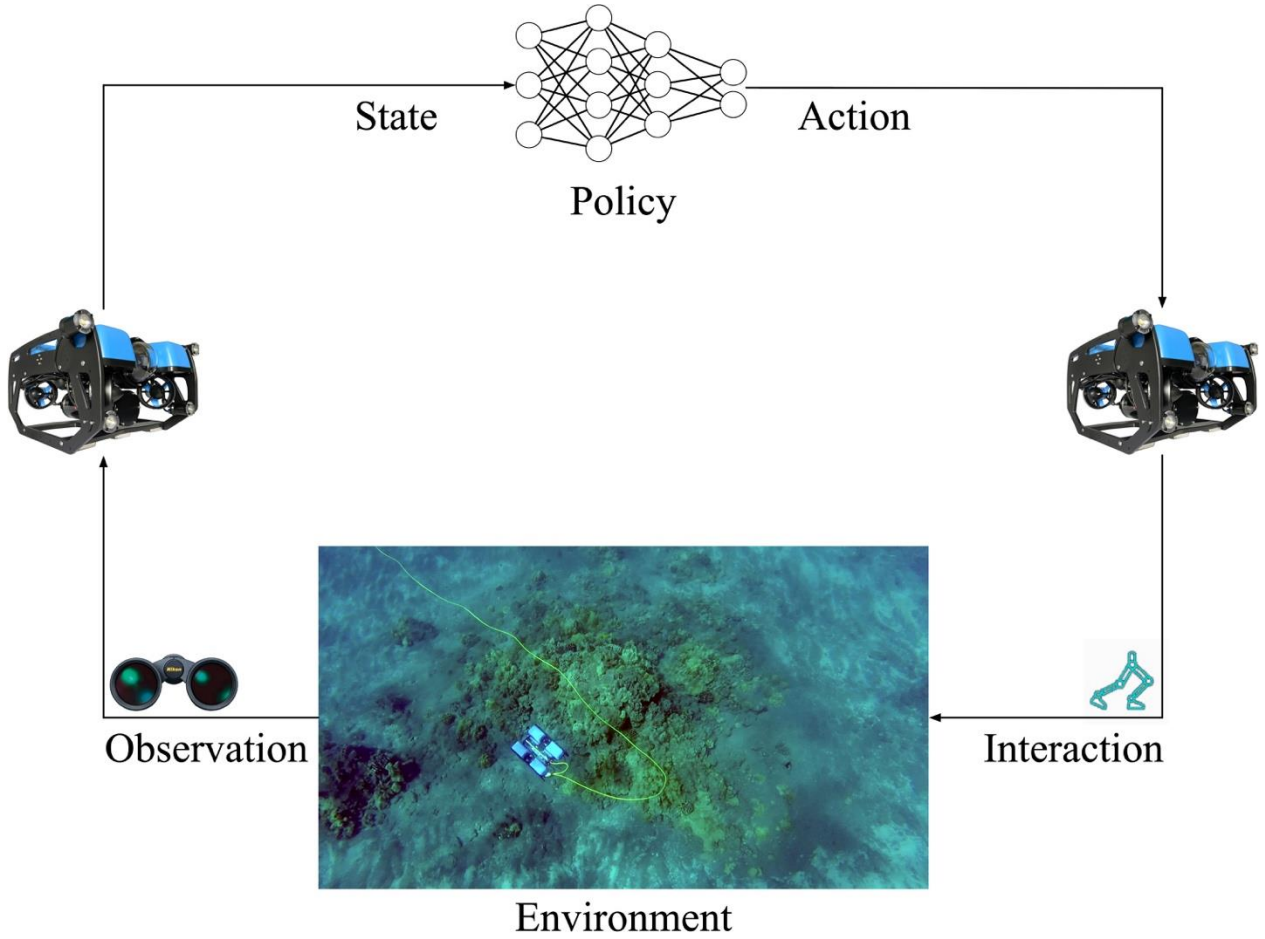
09/06/2025

Dr. Yi Dong & Prof. Xiaowei Huang

Outline

- Probabilistic Verification of Deep Reinforcement Learning
- Reachability Verification of Deep Reinforcement Learning
- Privacy-preserving Distributed Learning

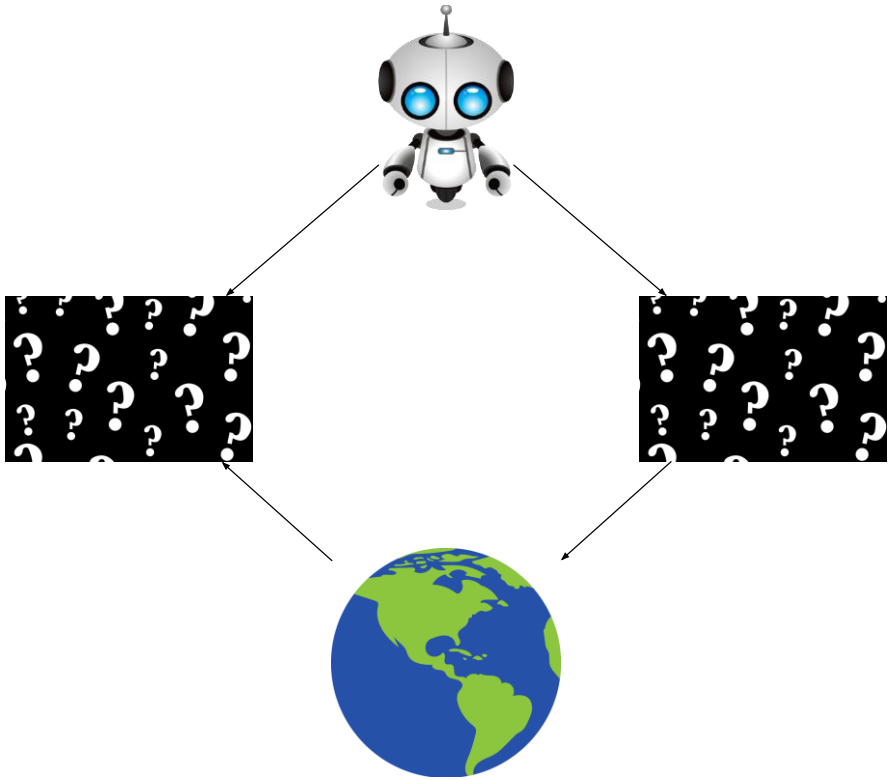
Deep Reinforcement Learning

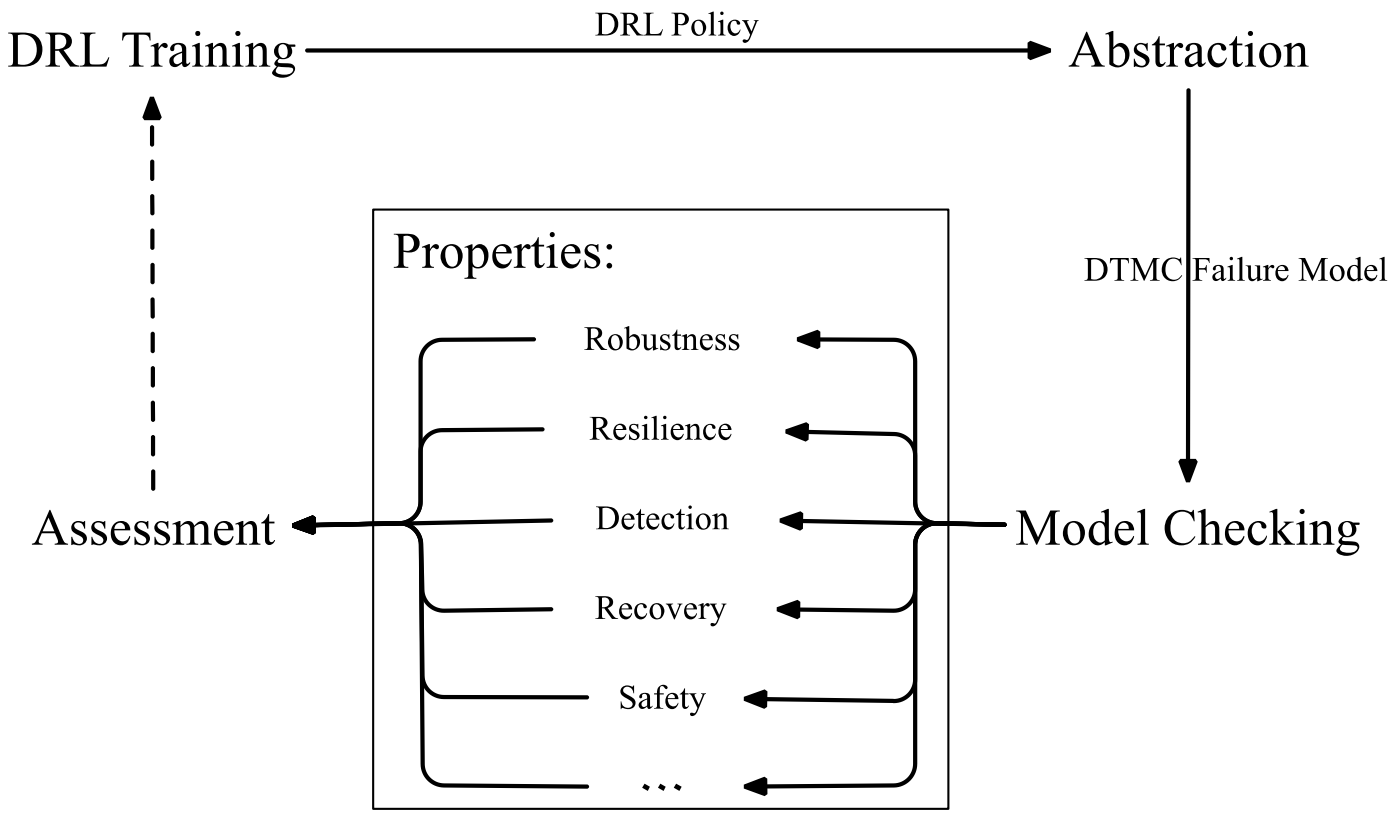


UNIVERSITY OF LIVERPOOL

EnnCore

Case 1: Unknown Environment





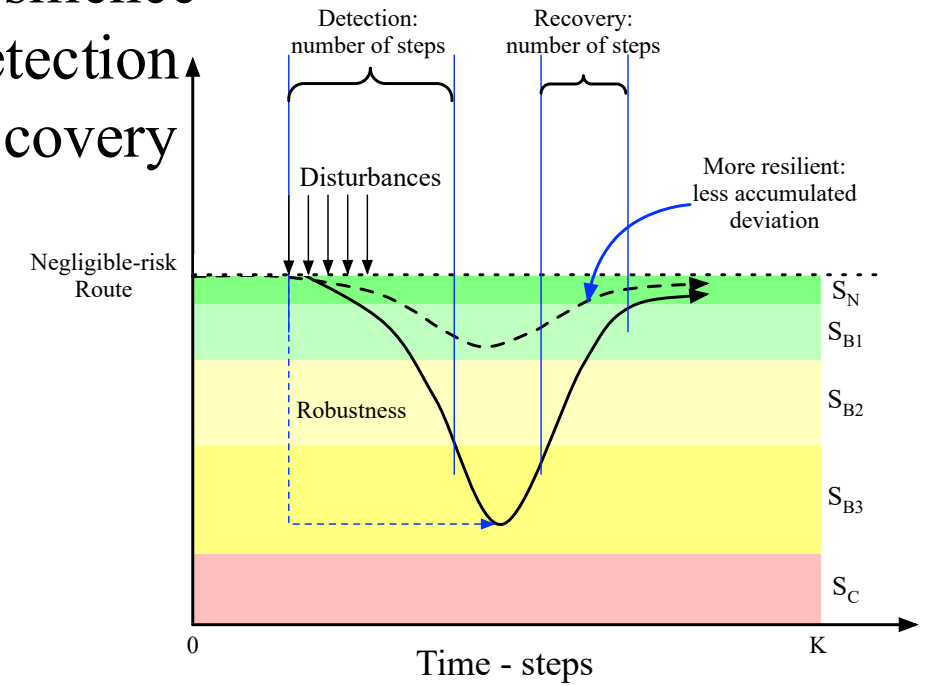
UNIVERSITY OF LIVERPOOL

[1] Y. Dong, X. Zhao and X. Huang, "Dependability Analysis of Deep Reinforcement Learning based Robotics and Autonomous Systems through Probabilistic Model Checking," *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*

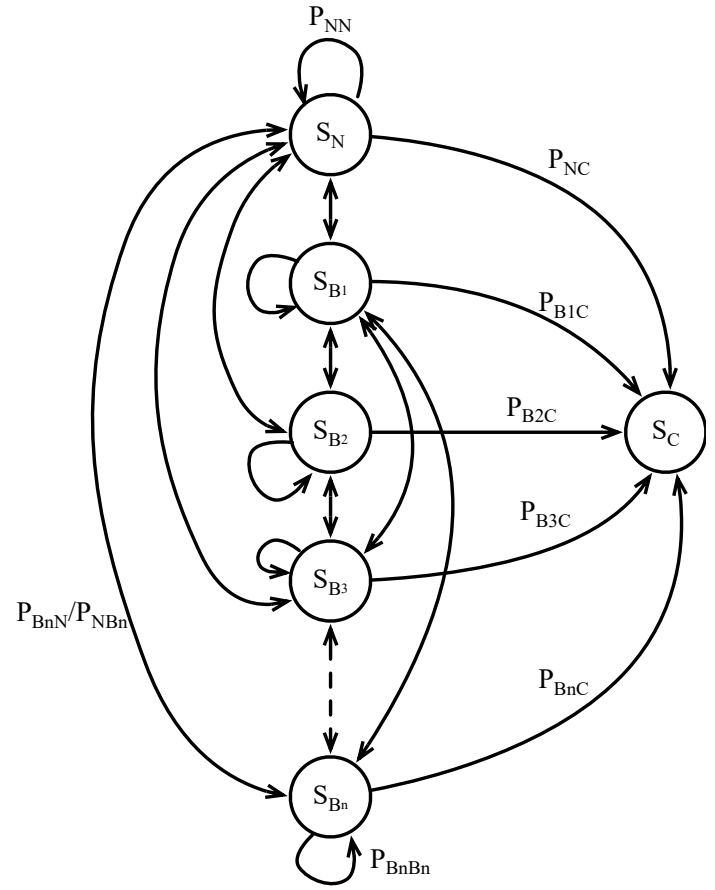
EnnCore

Our Solutions

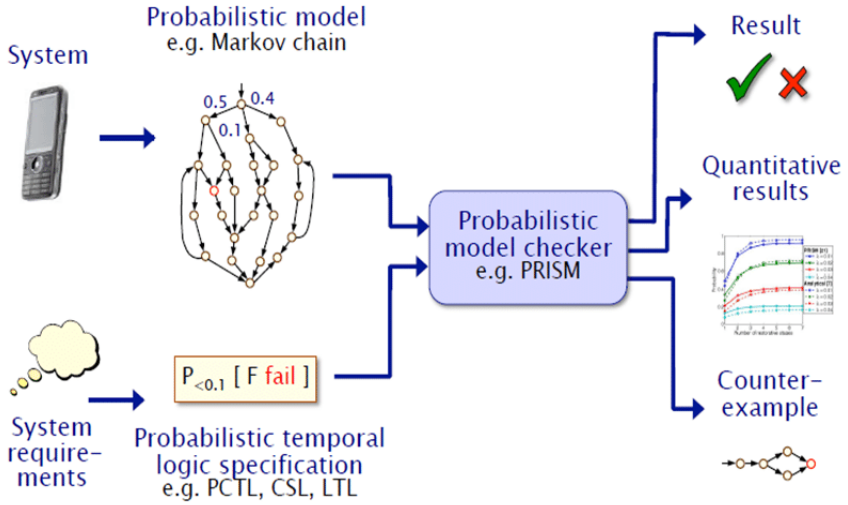
1. Safety Properties:
- Robustness
- Resilience
- Detection
- Recovery
- …



2. DTMC model:



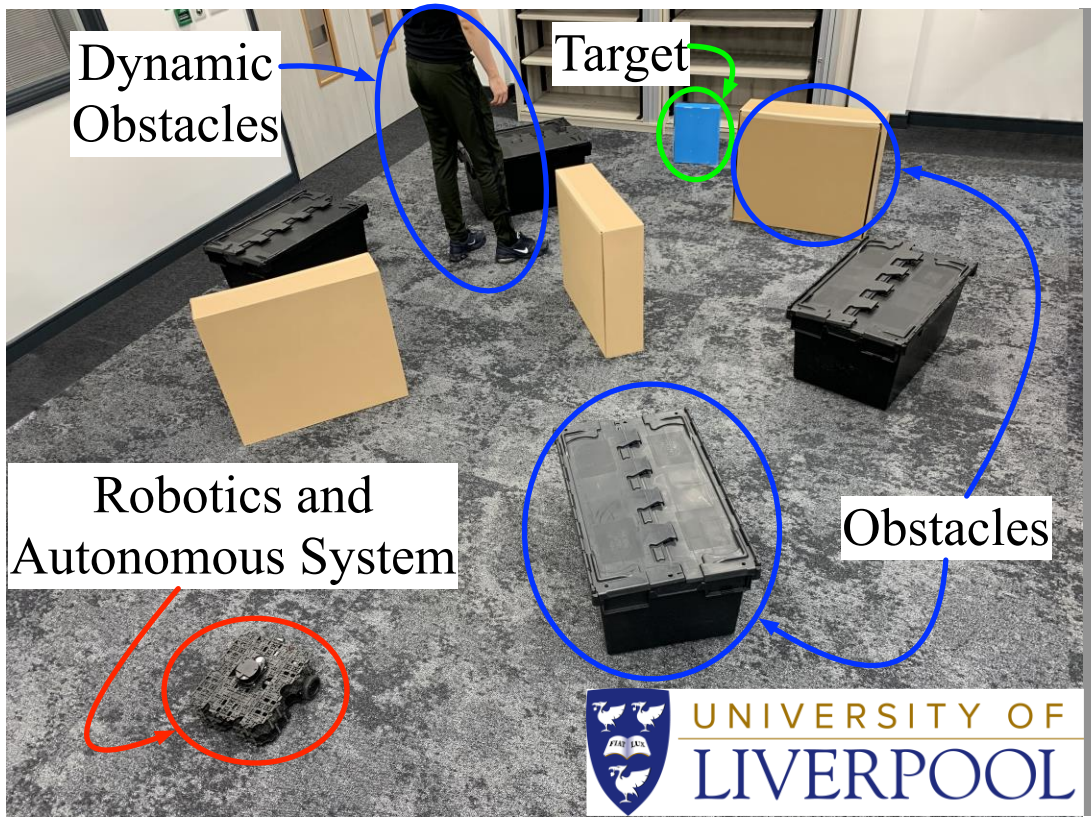
3. Probabilistic Model Checker:

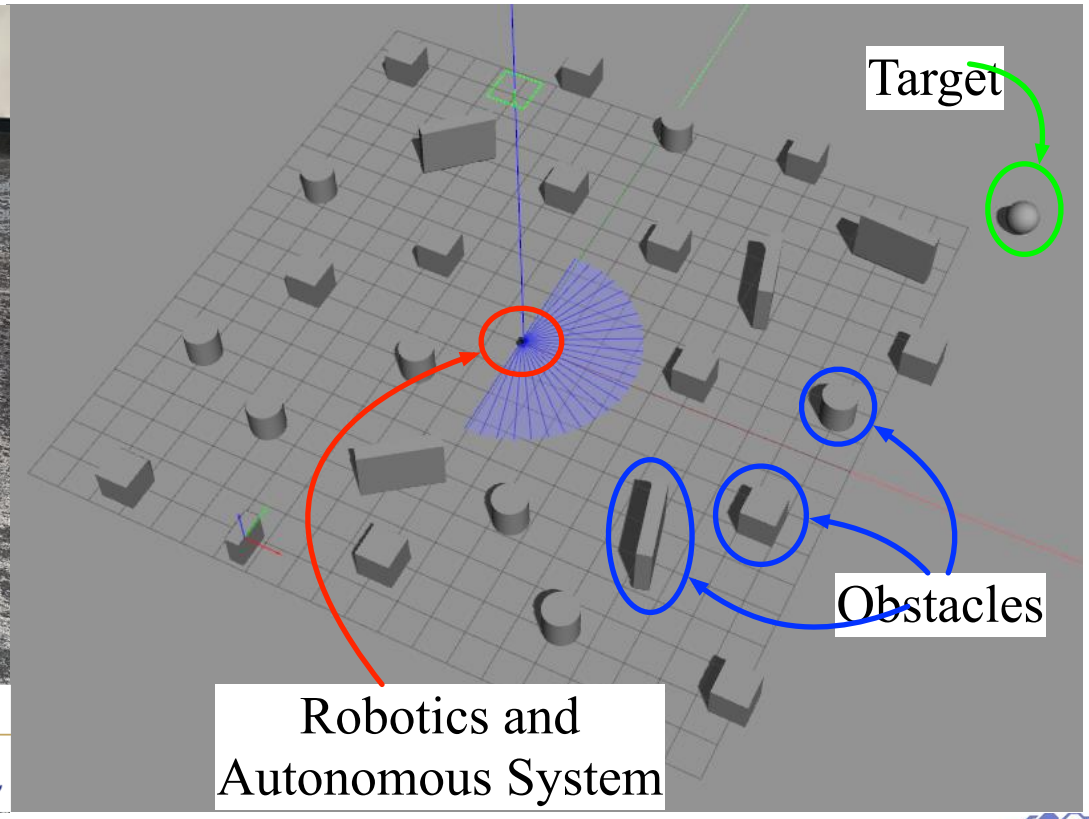


UNIVERSITY OF LIVERPOOL

EnnCore

Experiments



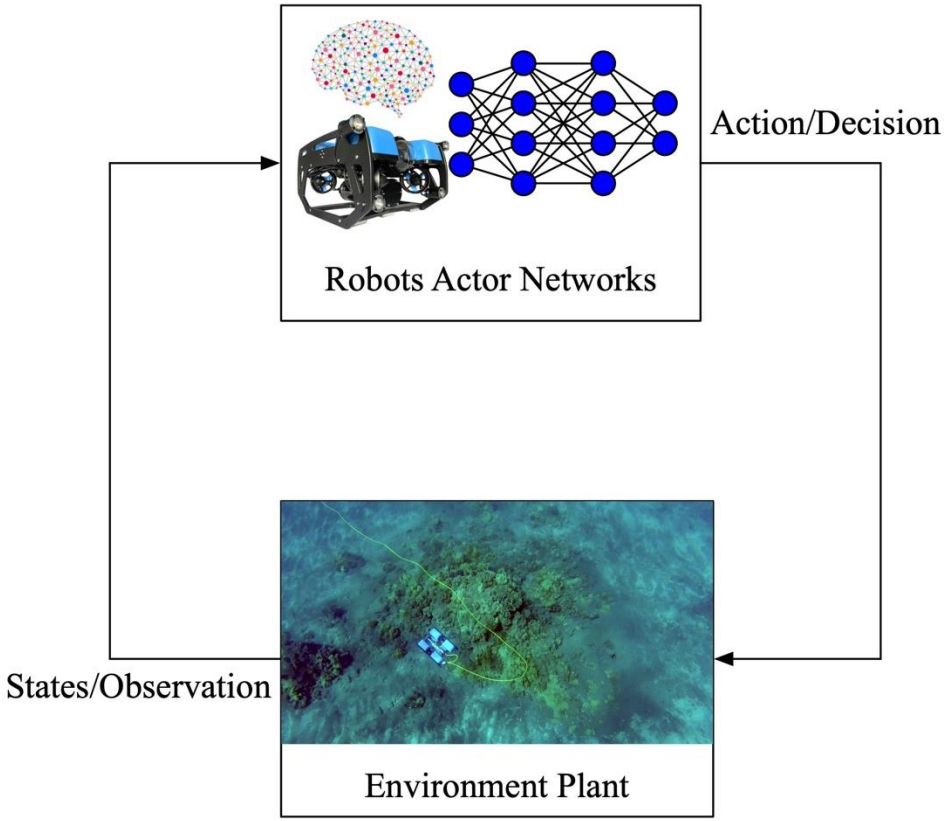


UNIVERSITY OF LIVERPOOL

EnnCore

Case 2: White-box Verification

Target: Probability of crash per random initial-condition (*pci*)



Robot Actor Networks:

- Number of Layers
- Number of Neurons
- Weights and Bias

Environment Plant:

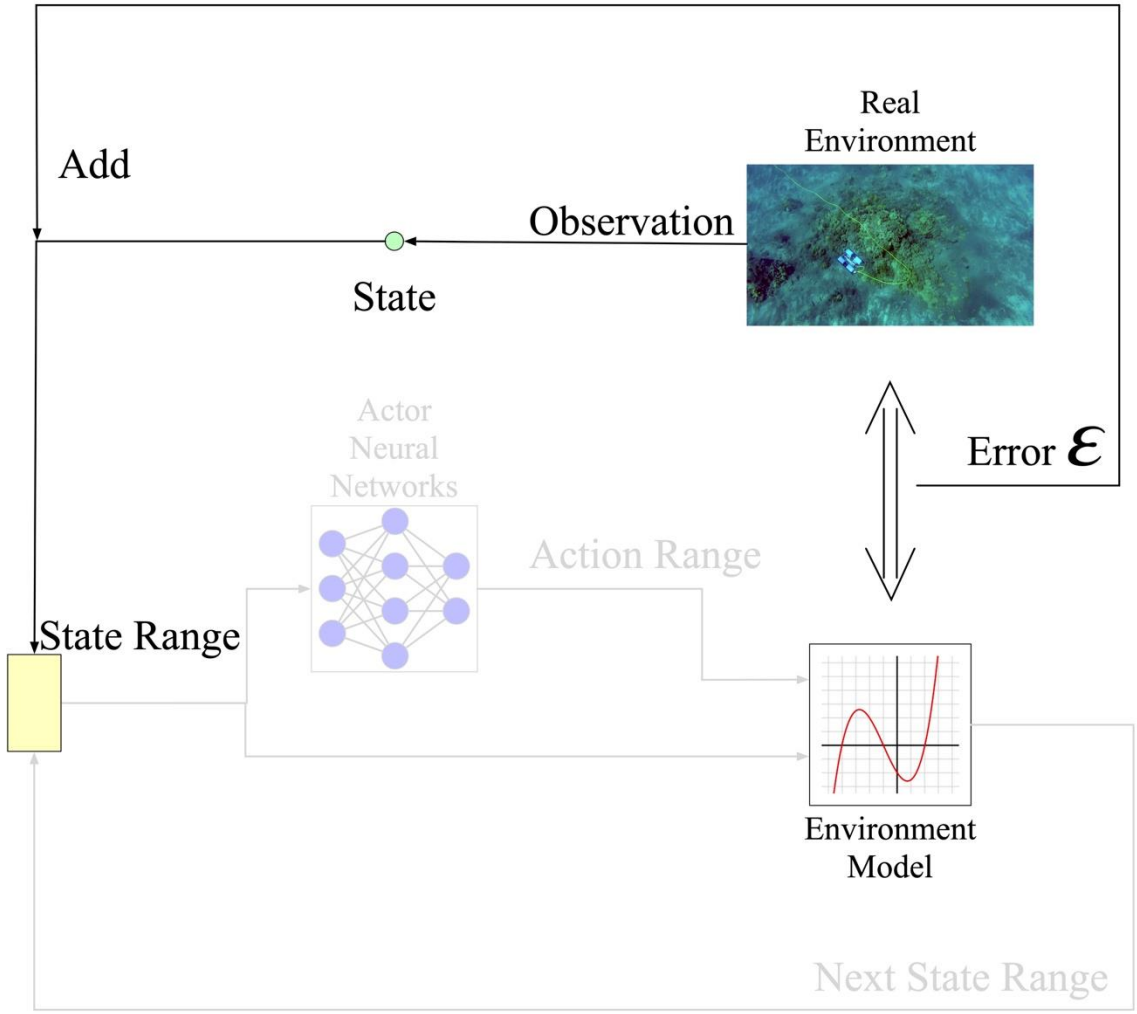
- Transition Probability
- Kinetic Model

[2] Y. Dong, X. Zhao, S. Wang and X. Huang, "Reachability Verification Based Reliability Assessment for Deep Reinforcement Learning Controlled Robotics and Autonomous Systems," in *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3299-3306, April 2024

UNIVERSITY OF LIVERPOOL

EnnCore

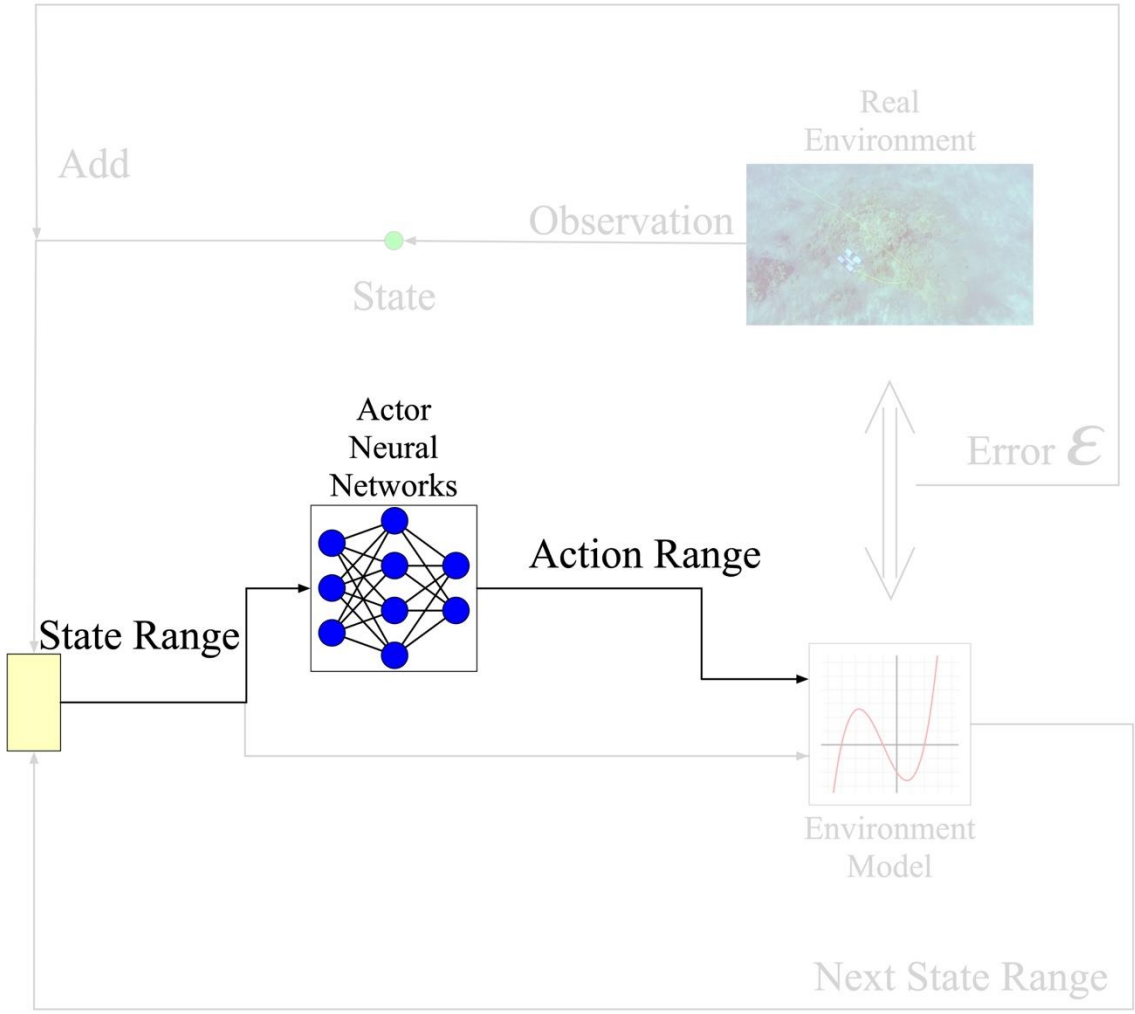
Reachability Verification



UNIVERSITY OF LIVERPOOL

EnnCore

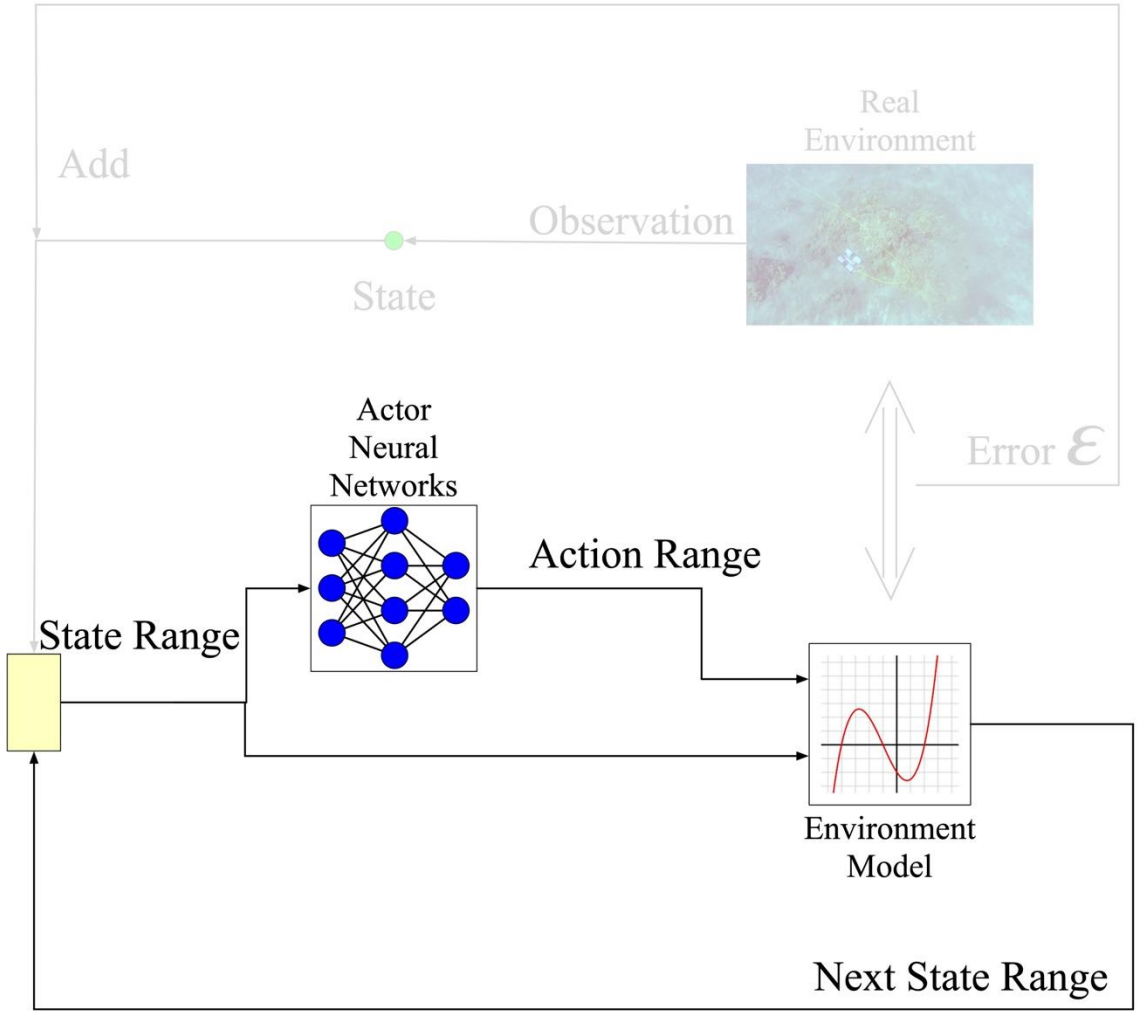
Reachability Verification



UNIVERSITY OF LIVERPOOL

EnnCore

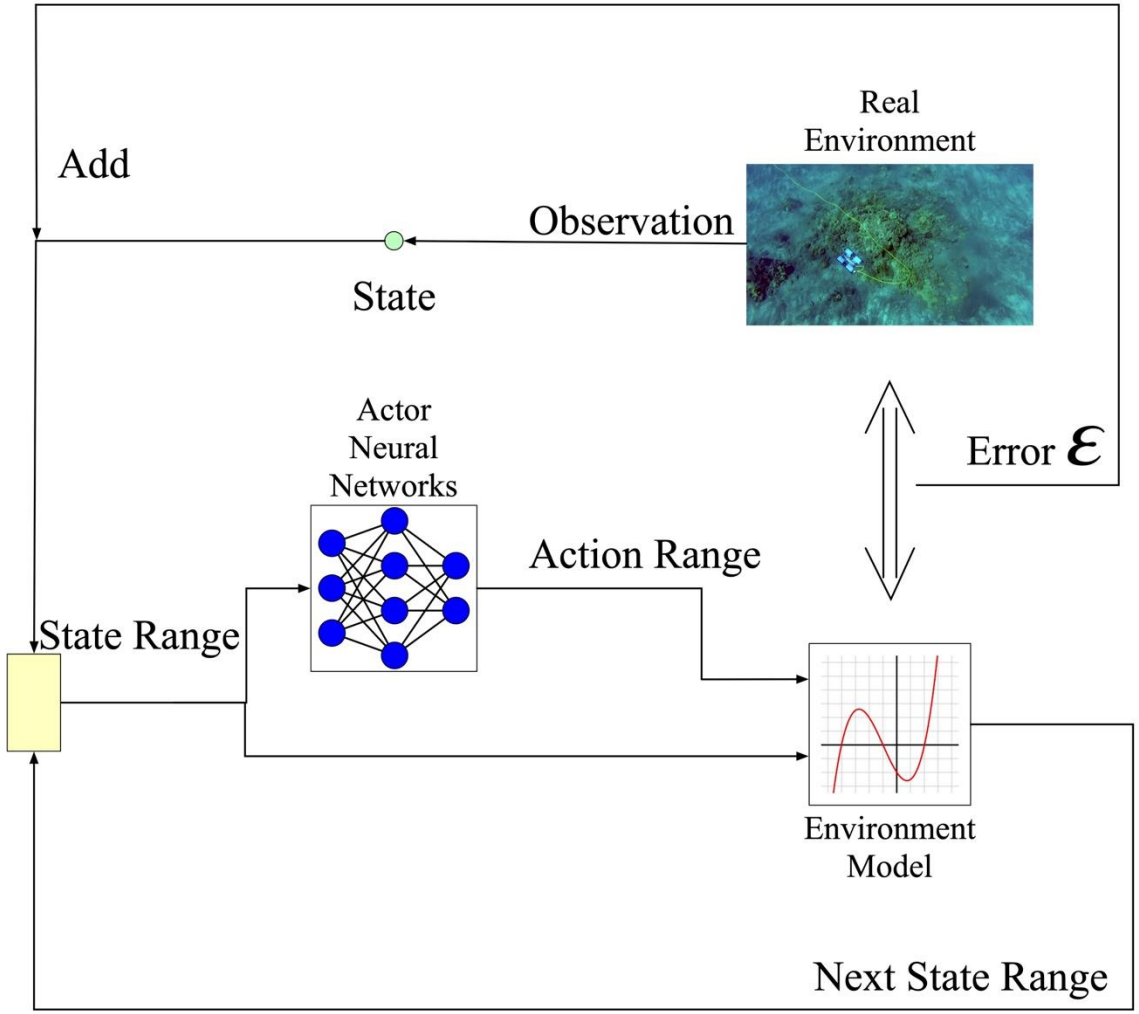
Reachability Verification

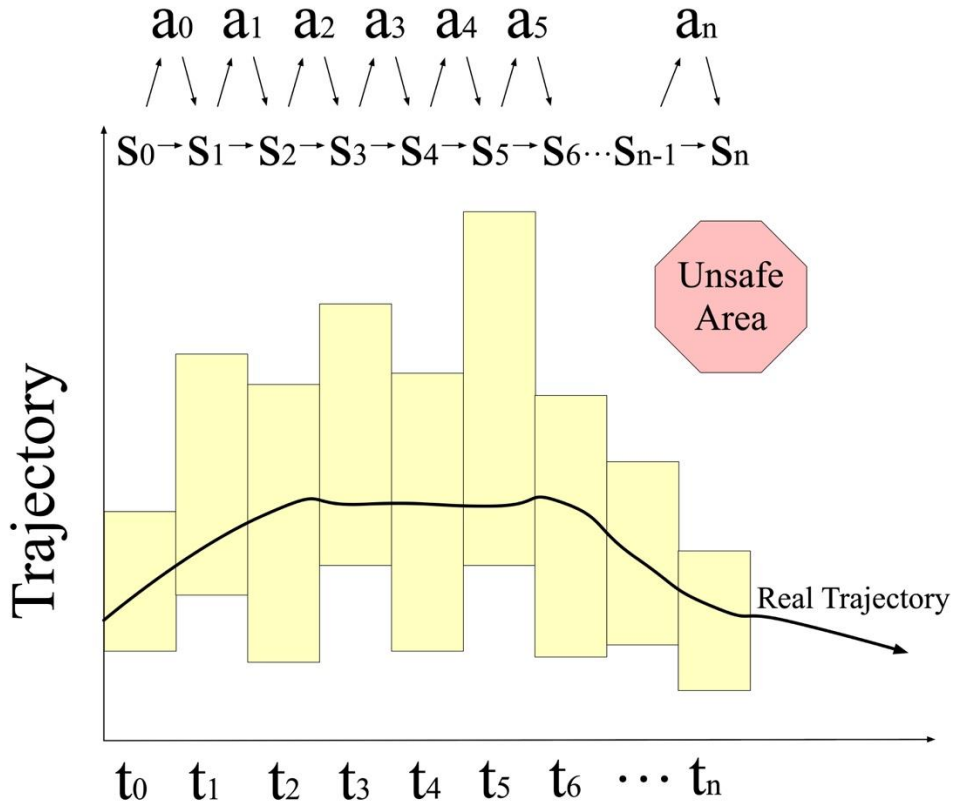


UNIVERSITY OF LIVERPOOL

EnnCore

Reachability Verification

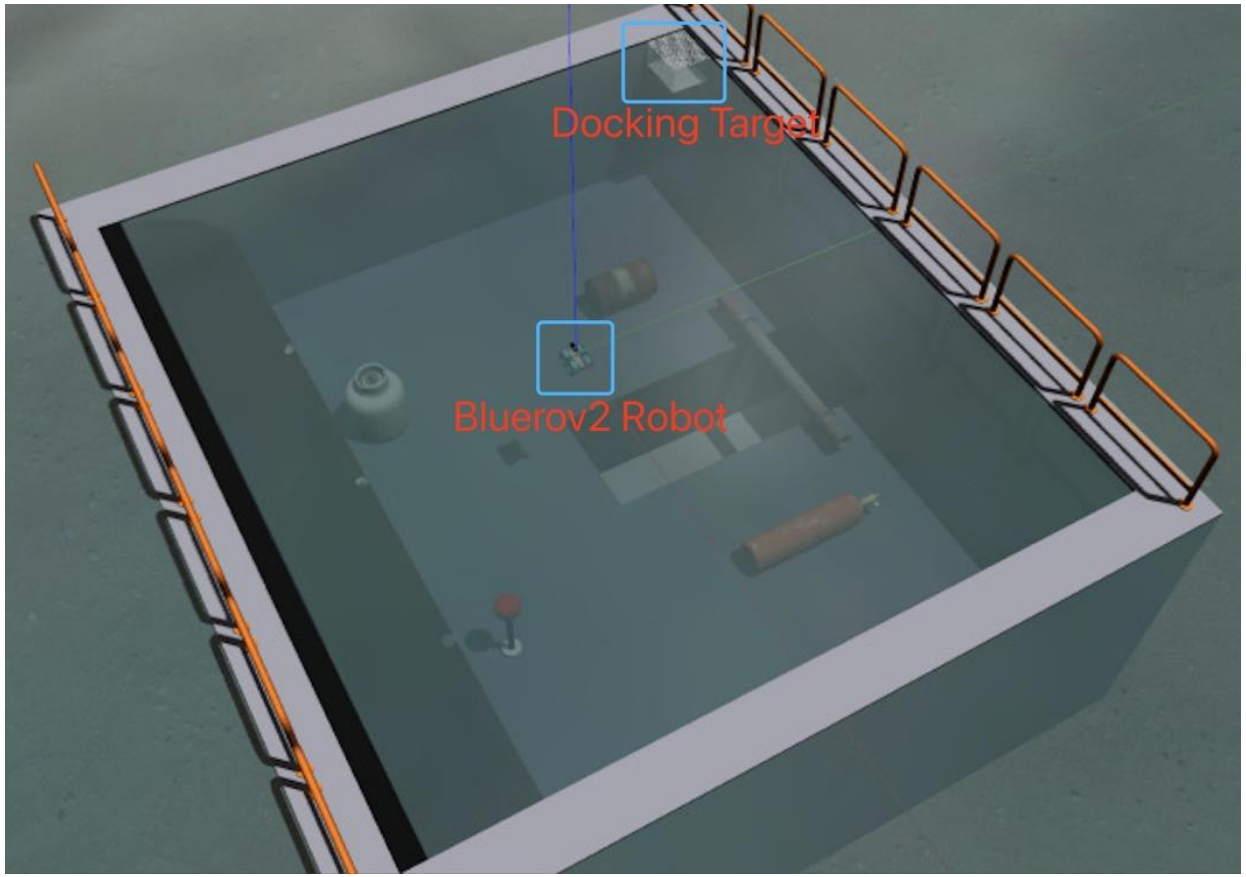




UNIVERSITY OF LIVERPOOL

EnnCore

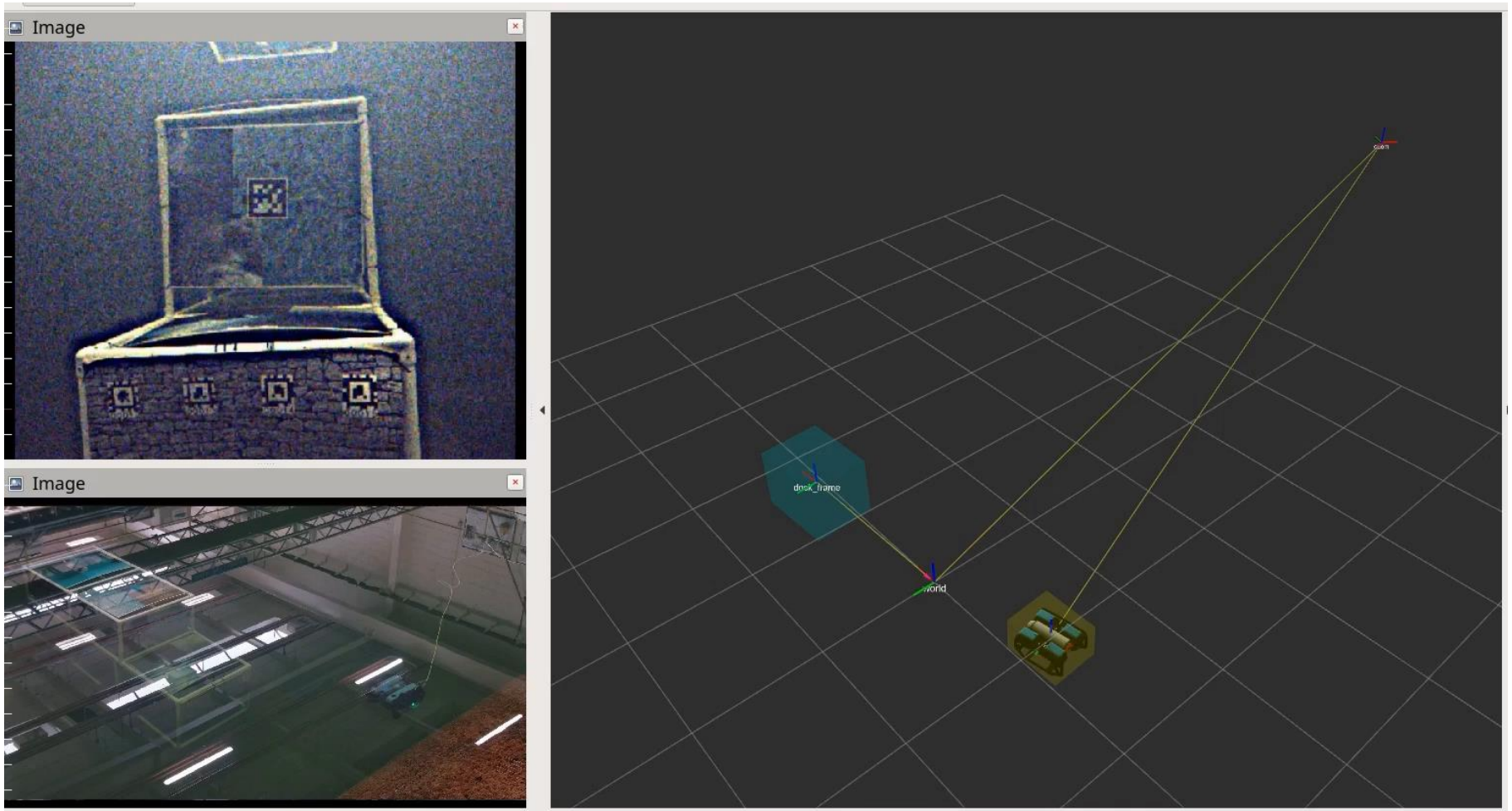
Environment Setup



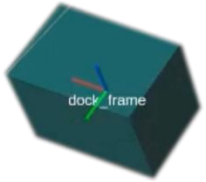
UNIVERSITY OF LIVERPOOL

EnnCore

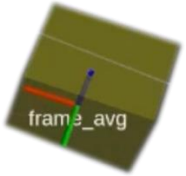
Demo a Safe Route



- Docking Cage



- Reachable Range of UUV

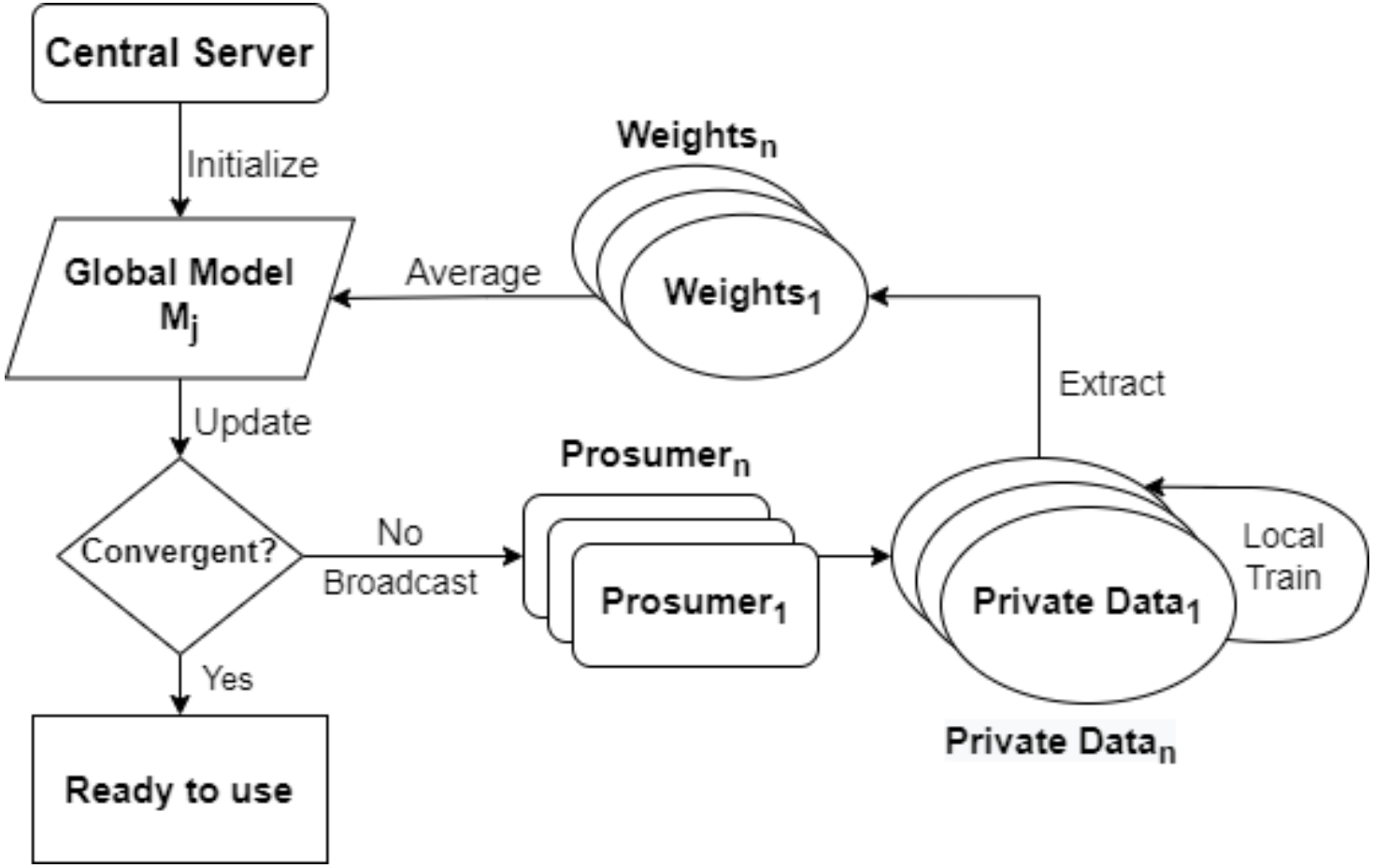


- World Frame (Static)

UNIVERSITY OF LIVERPOOL

EnnCore

Case 3: Privacy-preserving Distributed Learning



Research Challenges:

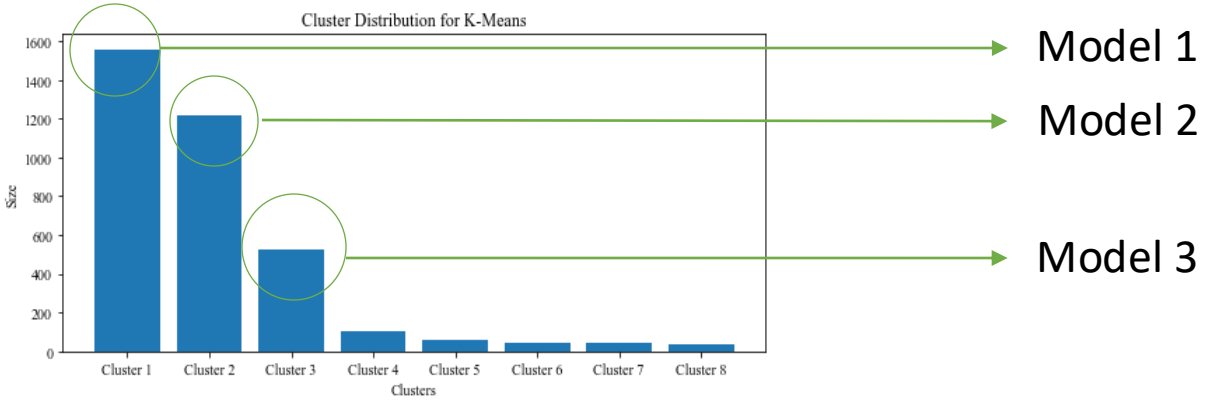
- Scalability
- Accuracy
- Robustness
- Privacy

UNIVERSITY OF LIVERPOOL

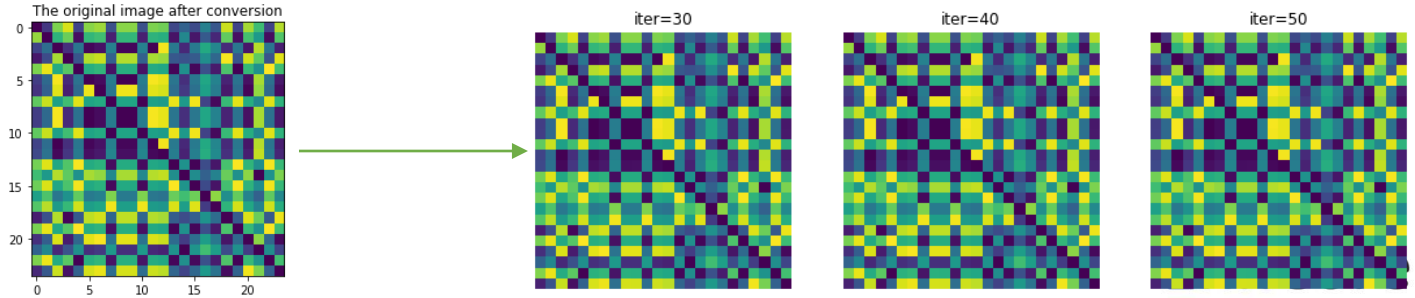
[3] Y. Dong, Y. Wang, M. Gama, M. A. Mustafa, G. Deconinck and X. Huang, "Privacy-Preserving Distributed Learning for Residential Short-Term Load Forecasting," in *IEEE Internet of Things Journal*, vol. 11, no. 9, pp. 16817-16828, 1 May1, 2024

EnnCore

Our Solution

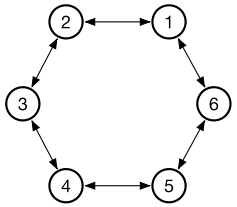


- Scalability: K-means clustering to handle the non-IID dataset.
- Accuracy: 4 types of NN models (DNN, LSTM, CNN, WaveNet)
- Robustness: Damaged training data.
- Privacy: Deep Leakage from Gradients.

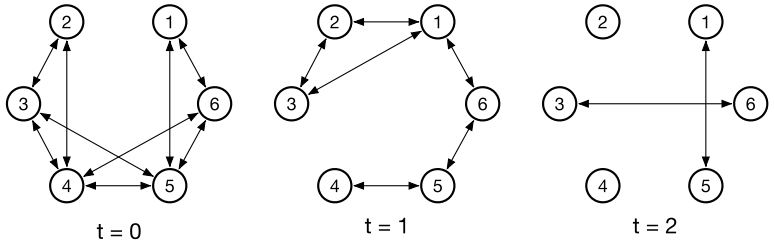


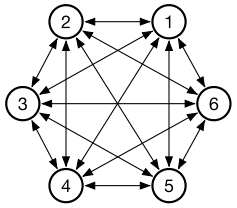
UNIVERSITY OF LIVERPOOL

Our Solution



Ring Topology

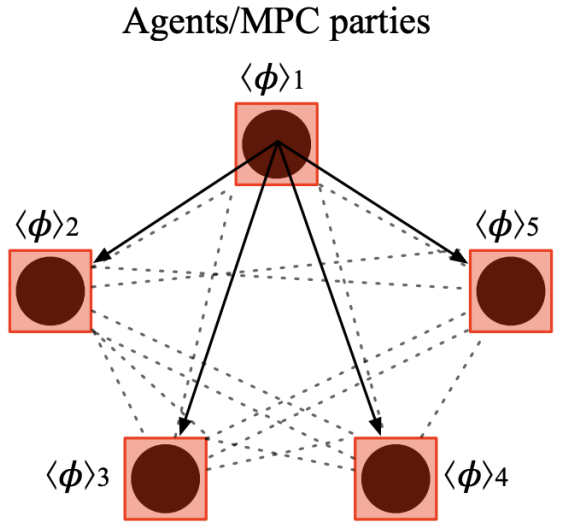




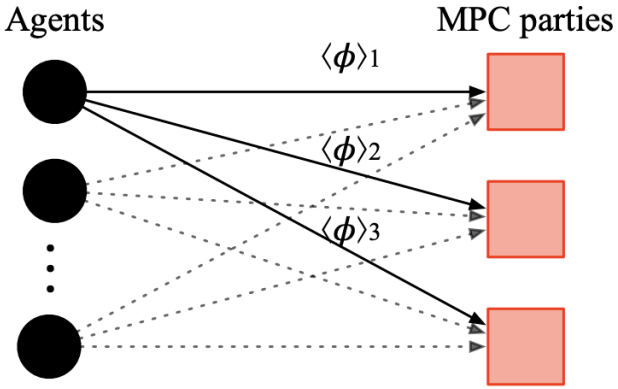
Fully Connected Topology

Markovian Switching Topologies

Distributed Frameworks



(a) MPC parties played by the agents.



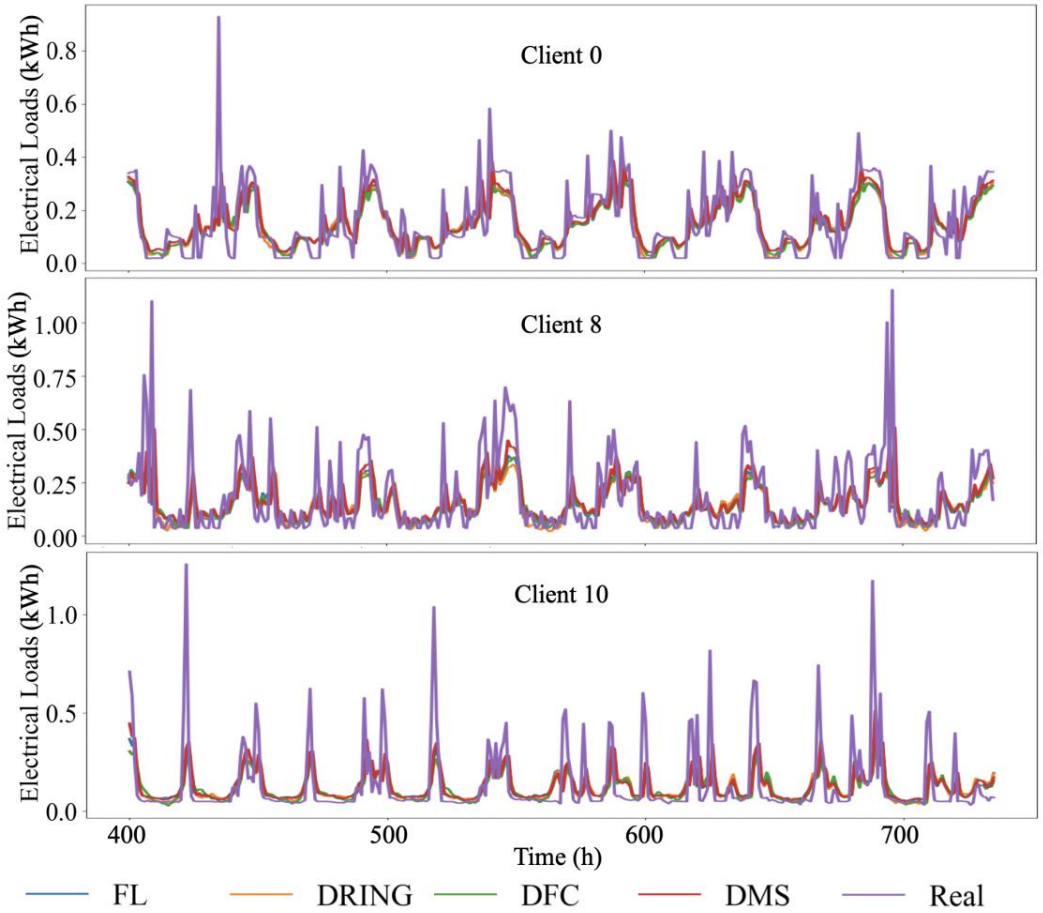
(b) MPC parties played by external servers.

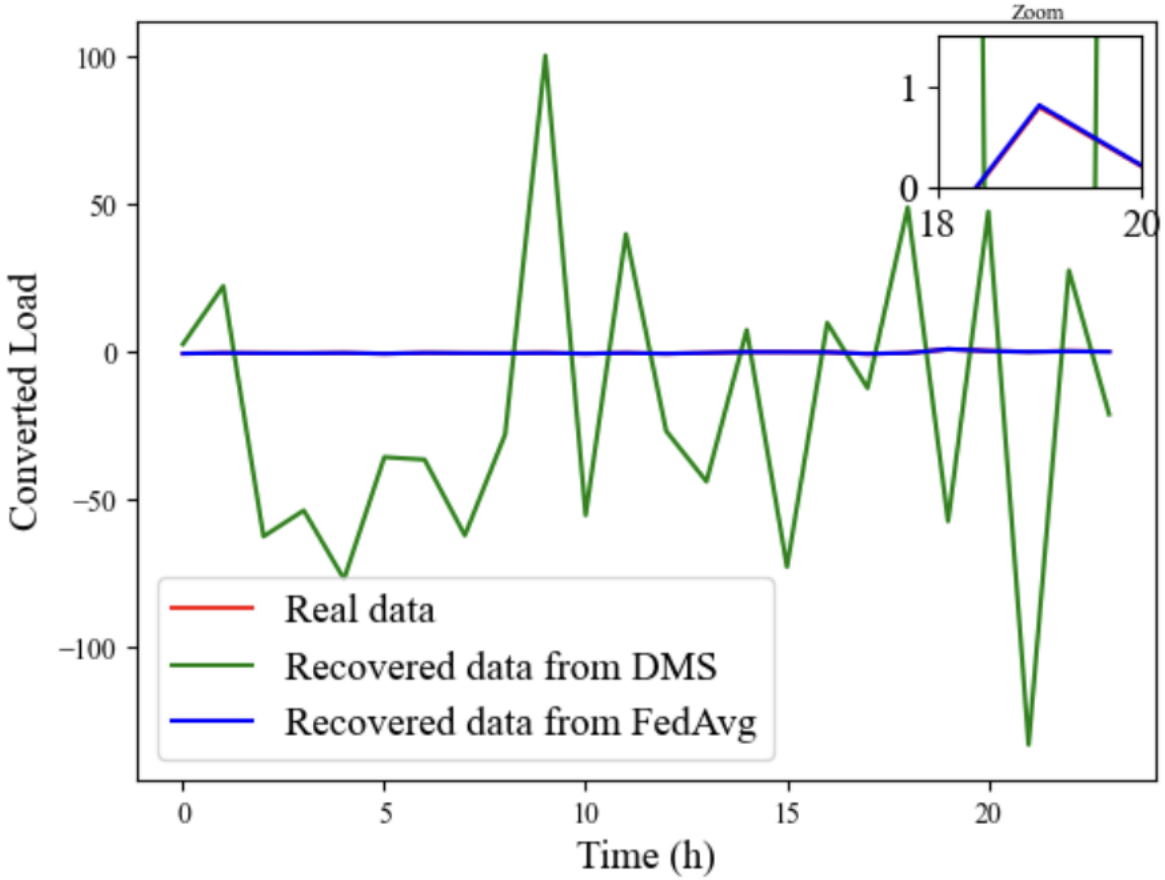
Agent secret sharing a gradient ϕ with the MPC parties.

UNIVERSITY OF LIVERPOOL

EnnCore

Experiment Results





Recovered data from DLG attack.

UNIVERSITY OF LIVERPOOL

EnnCore

Thank You

- Please refer to our EnnCore project website for more details:
- https://enncore.github.io/
- yi.dong@liverpool.ac.uk

UNIVERSITY OF LIVERPOOL

EnnCore