Multi-Agent Domain Calibration with a Handful of Offline Data

Tao Jiang[1,2,3*], **Lei Yuan**[1,2,3*], **Lihe Li**[1,2], **Cong Guan**[1,2], **Zongzhang Zhang**[1,2†], **Yang Yu**[1,2,3]

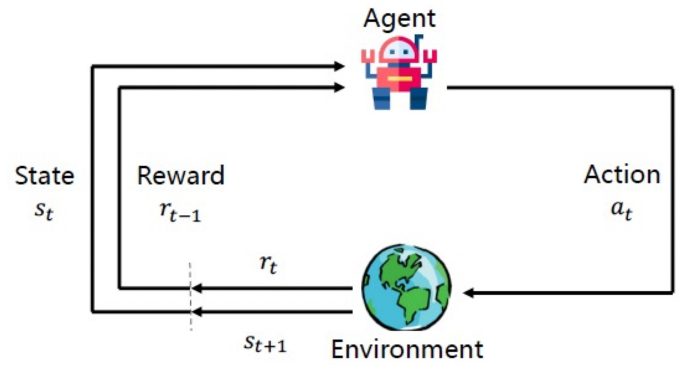
[1] National Key Laboratory of Novel Software Technology, Nanjing University
[2] School of Artificial Intelligence, Nanjing University
[3] Polixir Technologies

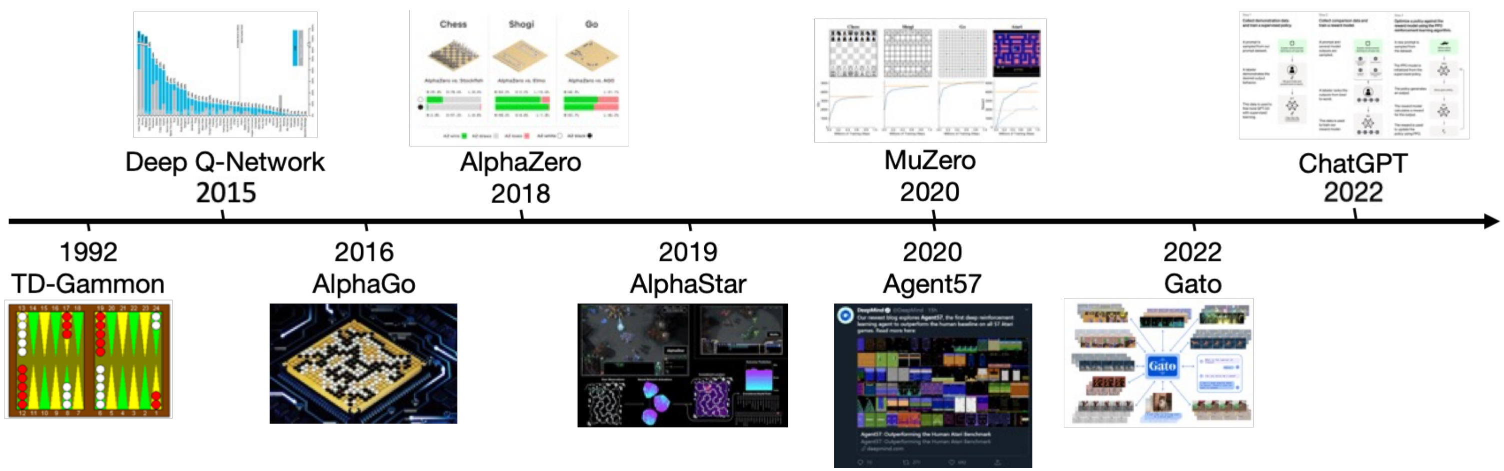
Presenter: Tao Jiang

2024/11/12

Background

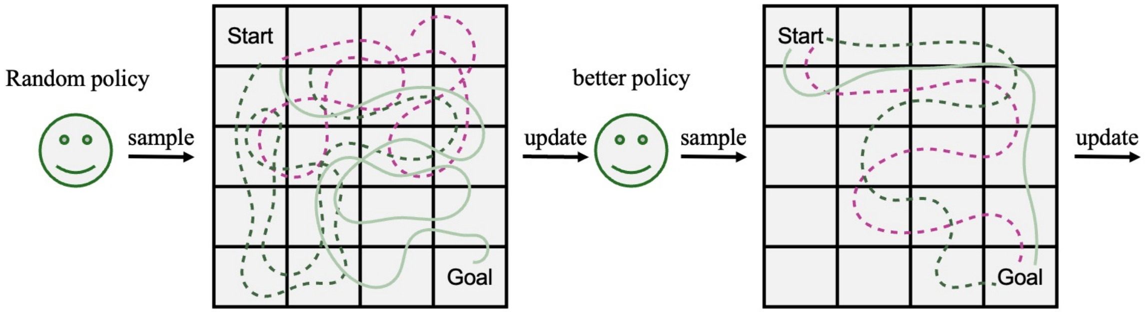
- Reinforcement Learning (RL) continuously optimizes a policy to maximize the expected cumulative reward:
  - **Interact** with the environment
  - Learn from **trial and error**
  - Demonstrate significant **potential and progress** across various industries



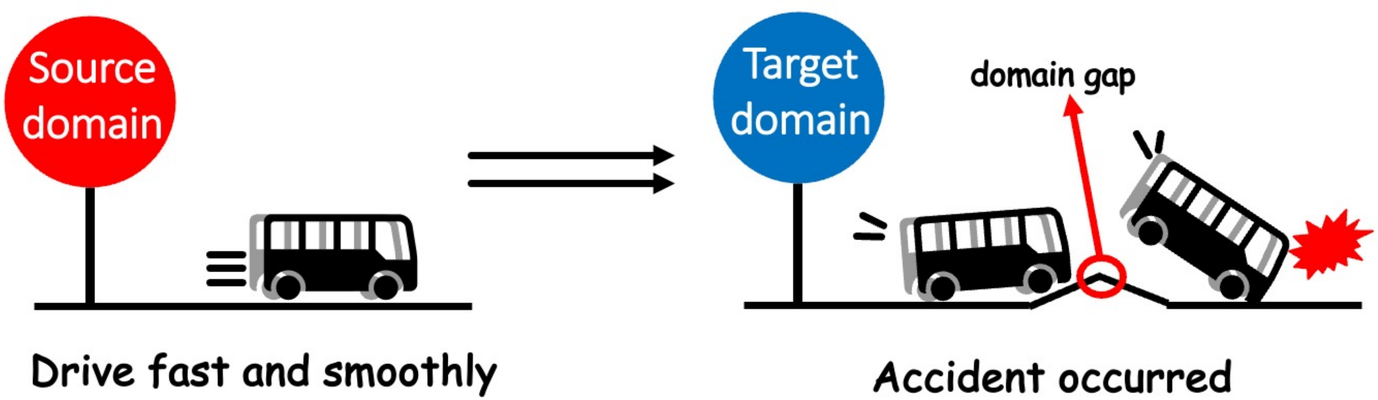


Background

- The trial-and-error learning process

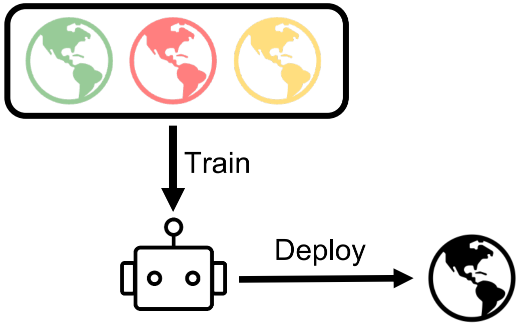


  - may be unrealistic in safety-critical areas like autonomous driving
  - One solution is to train the policy in a surrogate source domain and deploy it in the downstream target domain, but it fails due to the domain gap

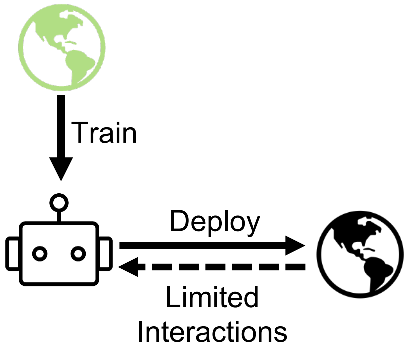


Background

- **Domain transfer** methods aim to close the domain gap for policy transfer without sacrificing significant performance:
  - **Domain randomization** trains a generalizable policy across a series of randomized domains
    - require manually setting the randomized physics parameter distribution
    - trade optimality for robustness



  - **Domain adaptation** refines the trained policy through limited interactions in the target domain
    - still need online interactions with the target domain
    - entail prohibitive costs and safety risks

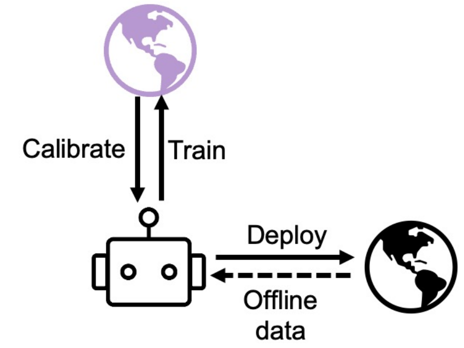


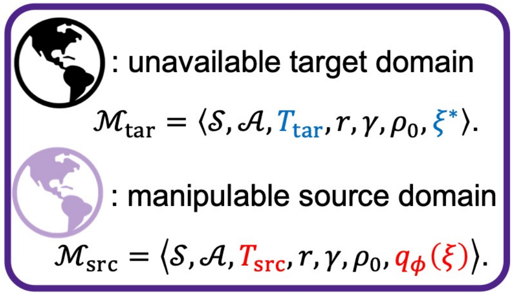
  - A more reasonable method is needed

Sadeghi, Fereshteh, and Sergey Levine. "CAD2RL: Real Single-Image Flight Without a Single Real Image." *Robotics: Science and Systems XIII* (2017).
Eysenbach, Benjamin, et al. "Off-dynamics reinforcement learning: Training for transfer with domain classifiers." International Conference on Learning Representations.

Background

- Offline domain calibration:
  - Utilize offline data from the target domain to calibrate the physics parameters of the source domain to align with the target domain
    - No manual setup required
    - No online interactions needed
  - Previous methods use evolution algorithms (EA) or RL to calibrate the parameters distribution $q_\phi(\xi) = \mathcal{N}(\mu, \Sigma)$
    - μ, Σ are both N-dimensional vectors
    - Work well when N is small, but work poorly when N is large due to low sample efficiency
      - Single-agent methods struggle to correctly evaluate the utility of all parameters simultaneously

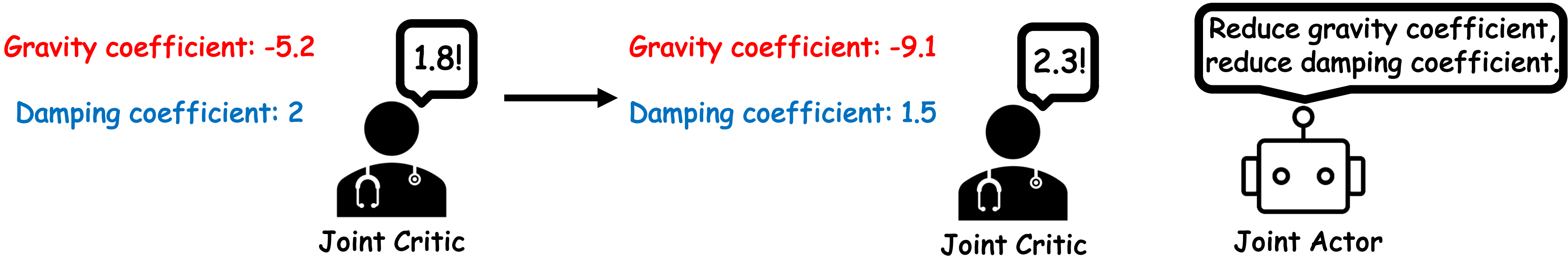




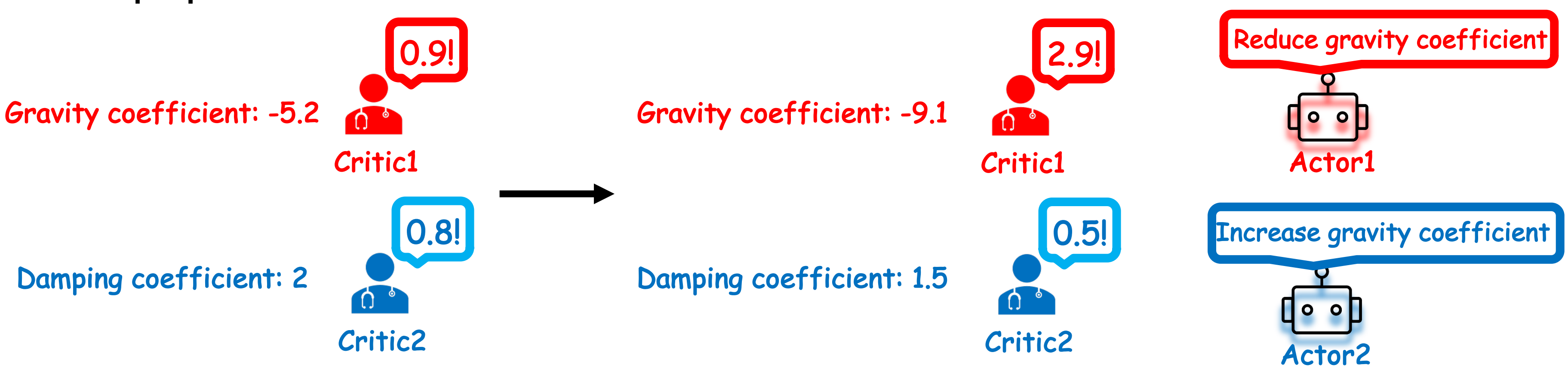
Tsai, Ya-Yen, et al. “DROID: Minimizing the reality gap using single-shot human demonstration.” IEEE Robotics and Automation Letters 6.2 (2021): 3168-3175.
Gur, Izzeddin, Ofir Nachum, and Aleksandra Faust. ”Targeted environment design from offline data." Deep RL Workshop NeurIPS 2021.

Motivation

- The action space of single-agent method is $|A|^N$.

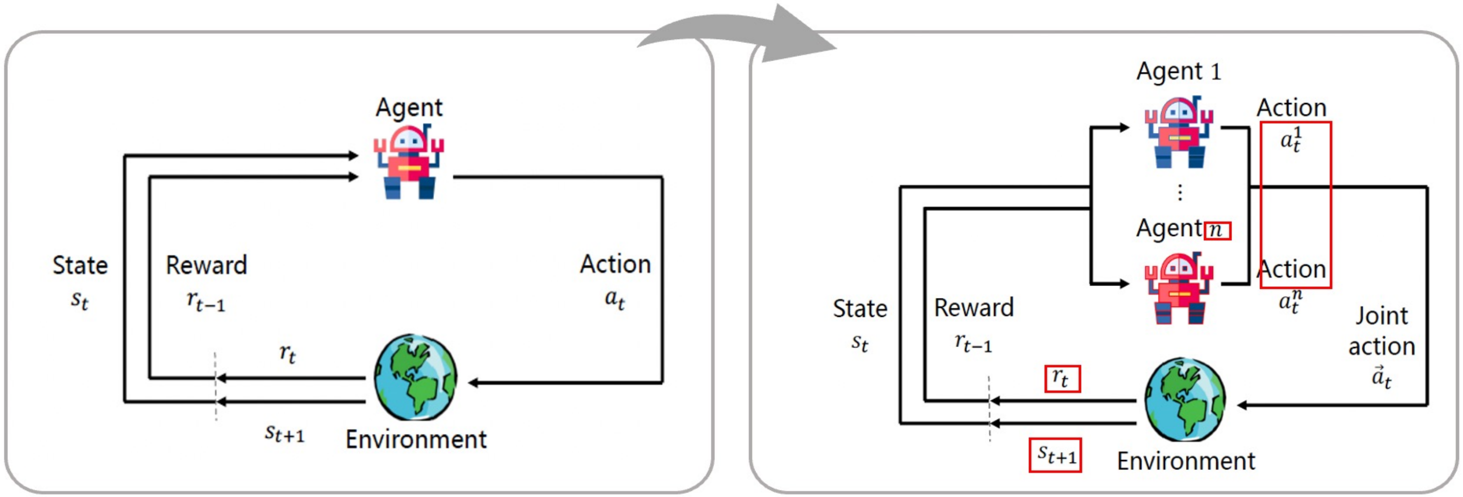


- The action space for each individual agent of multi-agent method is $|A|$.



Motivation

- When the dimension of physics parameters N is large, different physics parameters contribute to different aspects of the calibration process.
  - Formulate the problem into Multi-Agent System (MAS), where each agent calibrates a group of domain parameters with similar effects on the dynamics may be a wise choice



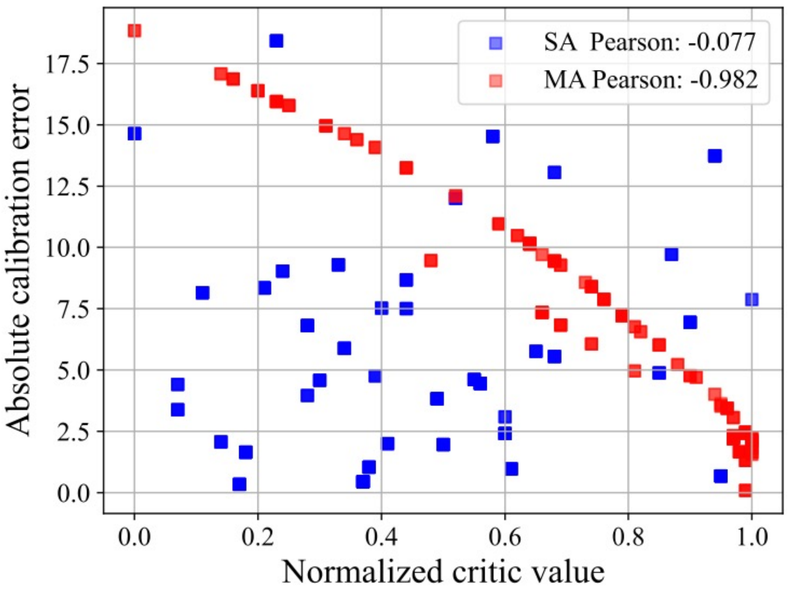
  - All the agents coordinate to reduce the domain gap, becoming a cooperative Multi-Agent Reinforcement Learning (MARL) problem

Albrecht, Stefano V., Filippos Christianos, and Lukas Schäfer. Multi-agent reinforcement learning: Foundations and modern approaches. MIT Press, 2024.

Motivation

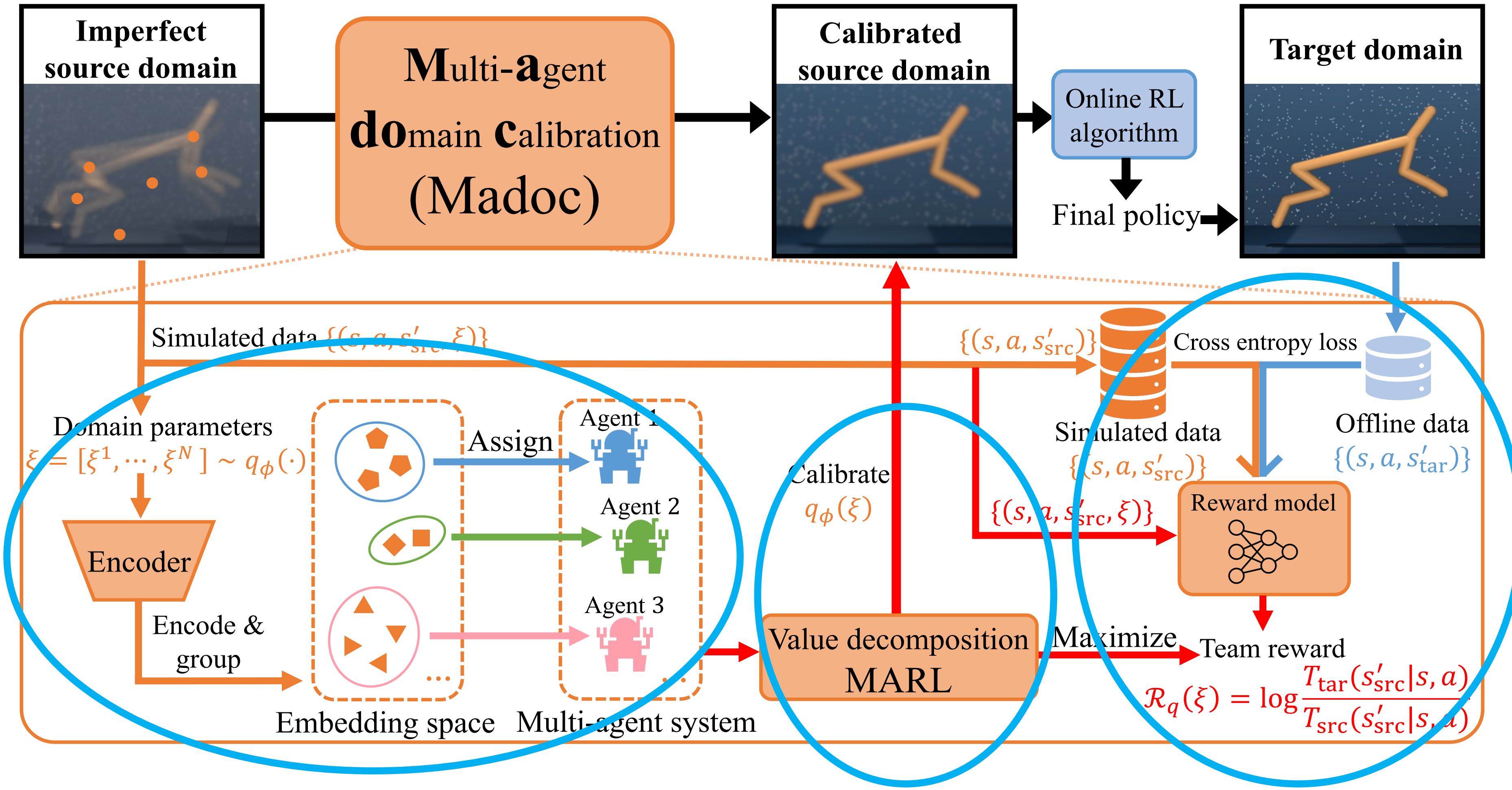
Learning And Mining from DatA
http://www.lamda.nju.edu.cn

- We conduct an experiment to investigate the correlation between the critic value and the absolute calibration error:
  - A good critic should output a high value when the parameter's absolute calibration error is low, and vice versa
  - In the figure, the single-agent (SA) method with a shared critic fails to achieve this, while the multi-agent (MA) method succeeds



- Therefore, our method formulates domain calibration as a cooperative multi-agent reinforcement learning (MARL) problem, improving fidelity and efficiency, even with a handful of offline data.

Framework


Imperfect source domain
Multi-agent domain calibration (Madoc)
Calibrated source domain
Online RL algorithm
Final policy
Target domain
Simulated data $\{(s, a, s'_{src}, \xi)\}$
$\{(s, a, s'_{src})\}$
Cross entropy loss
Domain parameters
$\xi = [\xi^1, \cdots, \xi^N] \sim q_\phi(\cdot)$
Assign
Agent 1
Agent 2
Agent 3
Encoder
Encode & group
Embedding space
Multi-agent system
Calibrate $q_\phi(\xi)$
Simulated data $\{(s, a, s'_{src})\}$
Offline data $\{(s, a, s'_{tar})\}$
$\{(s, a, s'_{src}, \xi)\}$
Reward model
Value decomposition MARL
Maximize
Team reward
$\mathcal{R}_q(\xi) = \log \frac{T_{tar}(s'_{src}|s,a)}{T_{src}(s'_{src}|s,a)}$

Experiment

LAMDA
Learning And Mining from DatA
http://www.lamda.nju.edu.cn

Offline domain calibration methods

Hybrid offline-and-online RL methods

Pure offline RL methods

Single-agent version

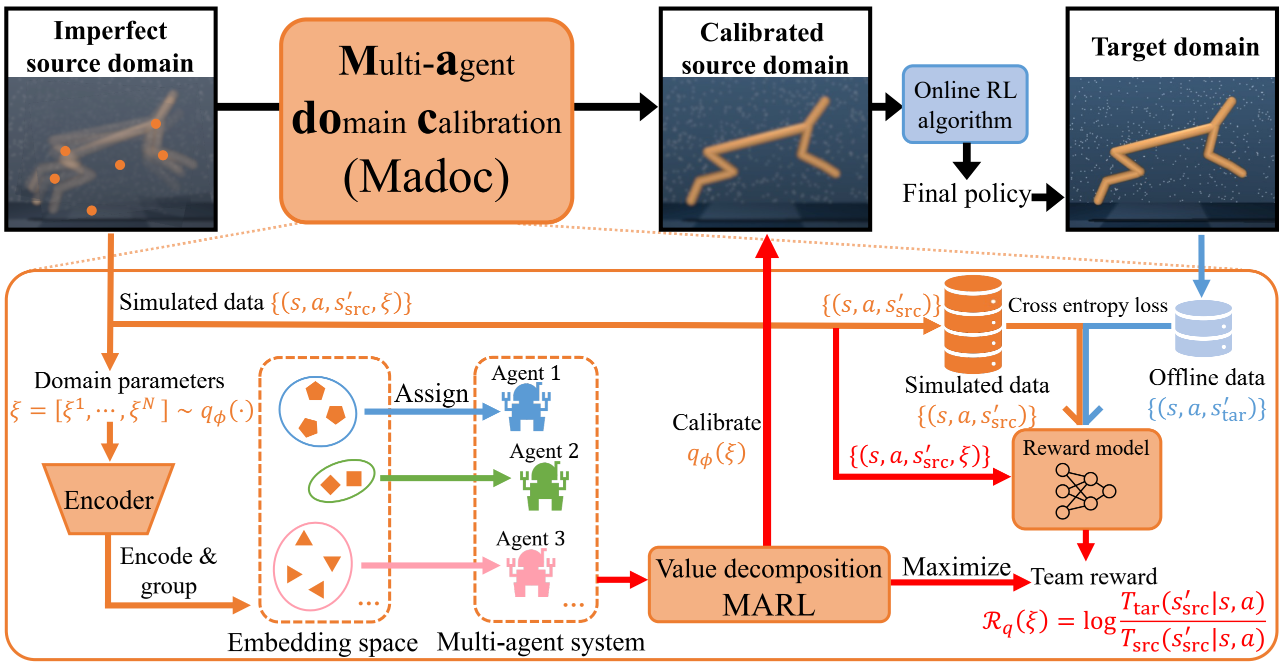
Our methods

Task	DROPO	DROID	OTED	H2O	DR+BC	CQL	MOREC	Madoc-S	Madoc
hfctah-med	41.0± 7.9	29.9± 7.5	76.1±11.7	57.3± 3.7	31.9±14.5	52.0± 3.0	73.9± 3.0	**93.3± 9.4**	91.9± 7.7
hfctah-med-rep	51.2±16.2	45.3± 8.0	67.5±13.4	50.4± 3.7	38.6± 4.4	48.9± 2.8	74.1± 2.8	84.7±18.6	**95.7± 9.9**
hfctah-med-exp	55.4±10.4	46.9±11.3	78.9±12.9	55.3± 4.1	40.5± 4.1	53.0± 2.9	72.0± 3.1	70.7±23.8	**96.9± 5.3**
hopper-med	59.1±34.6	73.9±10.4	49.8±21.4	83.7±19.0	43.2±24.7	50.3±18.8	**105.0± 1.0**	57.5±17.0	76.0±13.9
hopper-med-rep	43.0±18.7	45.3±17.6	65.4±26.1	84.1±12.3	43.5±16.9	70.0±13.9	28.7± 0.7	79.9±34.1	**90.2±11.7**
hopper-med-exp	80.4±21.3	21.5± 8.3	41.1±20.7	89.0±11.3	63.1±24.0	68.5±12.1	**106.0± 0.3**	47.7±13.0	81.5±18.6
walker-med	61.5±21.1	63.2±12.1	58.8±31.6	75.5± 8.7	57.2±12.1	4.5± 3.5	84.1± 0.8	69.9±19.8	**90.5±17.5**
walker-med-rep	19.8±16.6	16.8± 8.7	71.2±22.5	83.4± 1.3	43.7± 5.5	62.4±13.1	**85.4± 0.3**	60.6±33.1	**85.8±20.8**
walker-med-exp	60.0±13.8	73.8± 9.6	74.8±28.2	**91.7± 7.7**	61.1± 7.3	12.2± 8.3	86.2± 0.5	60.7±18.3	79.9±12.8
ant-med	16.4±12.2	20.8±17.8	65.3±41.8	60.0±26.6	29.2±12.6	58.0±20.6	64.9±41.2	76.5±29.6	**88.7±24.8**
ant-med-rep	64.1±31.9	64.4±35.0	62.4±41.9	**98.4±12.7**	34.8±15.0	43.8±33.7	6.1±14.1	58.8±42.9	81.2±16.5
ant-med-exp	76.7±34.0	64.2±41.1	70.0±35.5	66.5±22.8	30.3± 9.2	14.4±19.6	67.8±35.6	65.6±24.5	**101.0±21.5**
Average	52.4	47.2	65.1	74.6	43.1	44.8	71.2	68.9	**88.3**

Conclusion

Learning And Mining from DatA
http://www.lamda.nju.edu.cn

- We formulate offline domain calibration as a cooperative MARL problem to improve efficiency and fidelity



- Future work
  - Apply to high-dimensional vision tasks and real-world tasks

Thanks !