DFAC Framework: Factorizing the Value Function via Quantile Mixture for Multi-Agent Distributional Q-Learning

Wei-Fang Sun, Cheng-Kuang Lee, Chun-Yi Lee

Outline

- Background & Motivation
- Proposed Method: DFAC
- Experiment Results: Outperform all baselines

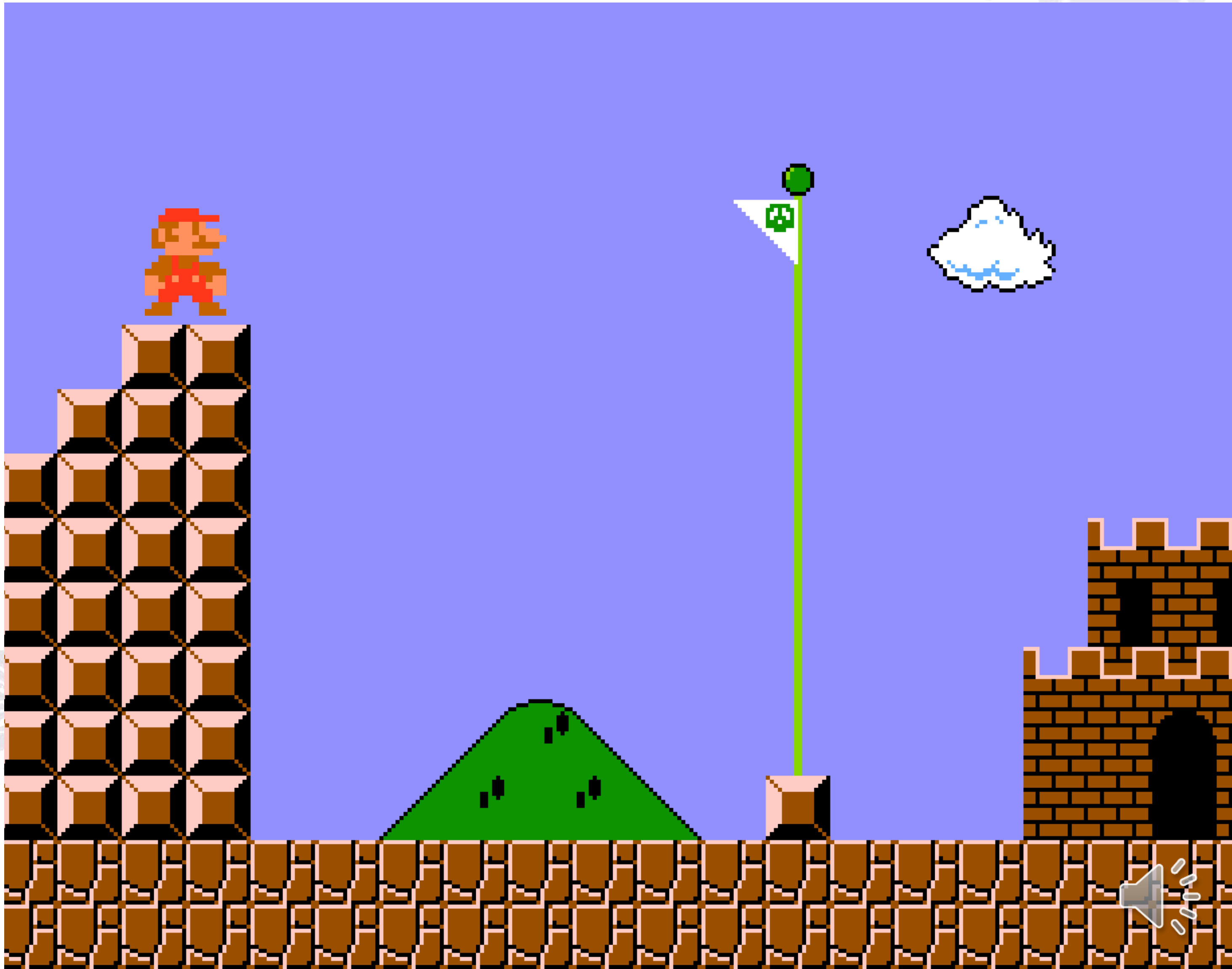
Outline

- **Background & Motivation**
- Proposed Method: DFAC
- Experiment Results: Outperform all baselines

Reinforcement Learning
(Focus on Q-Learning)

Reinforcement Learning
(Focus on Q-Learning)

Single-Agent RL
(SARL)

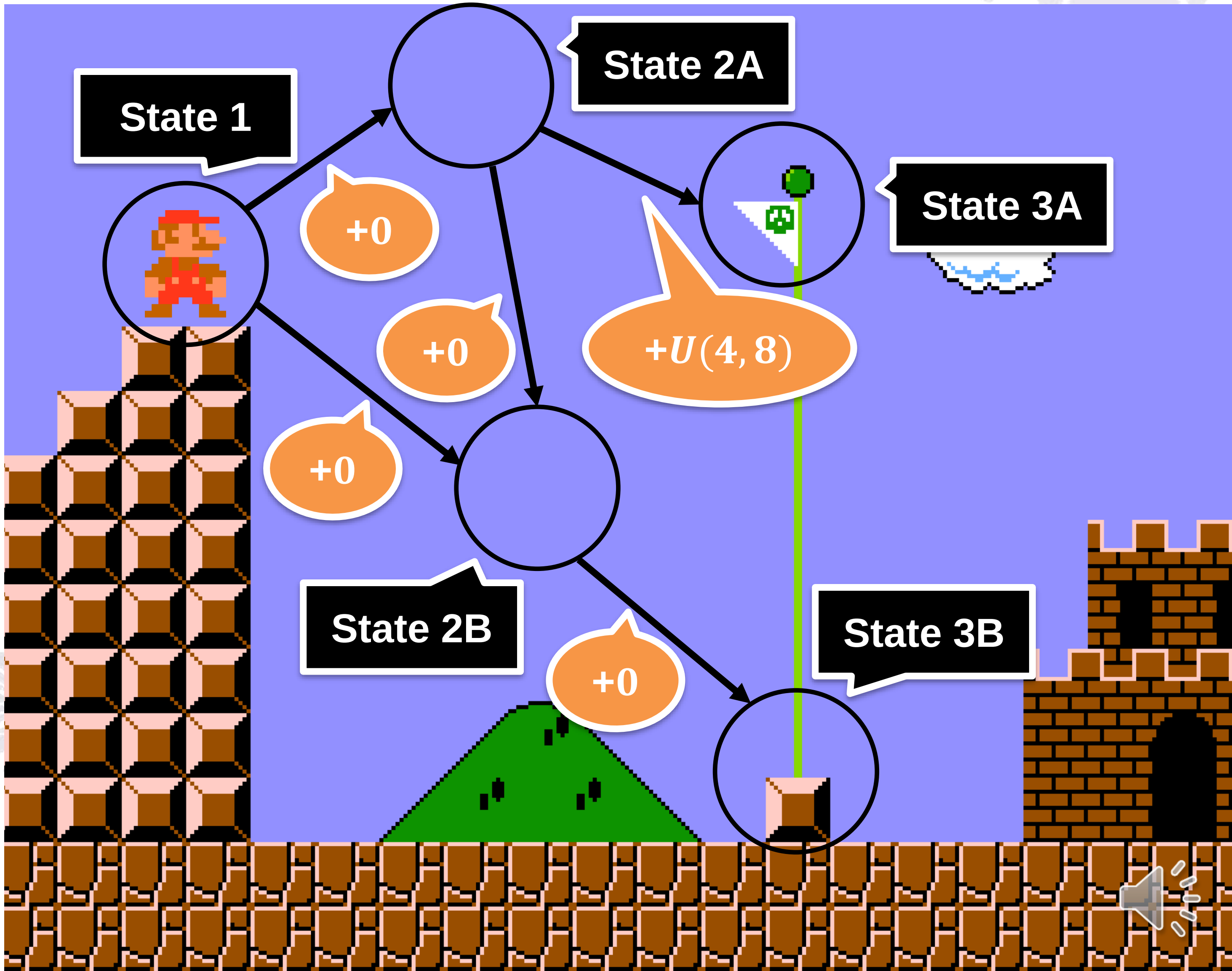


Markov Decision Process (MDP)

Reinforcement Learning
(Focus on Q-Learning)

Single-Agent RL
(SARL)

Optimization Goal:
Maximize the total reward within an episode

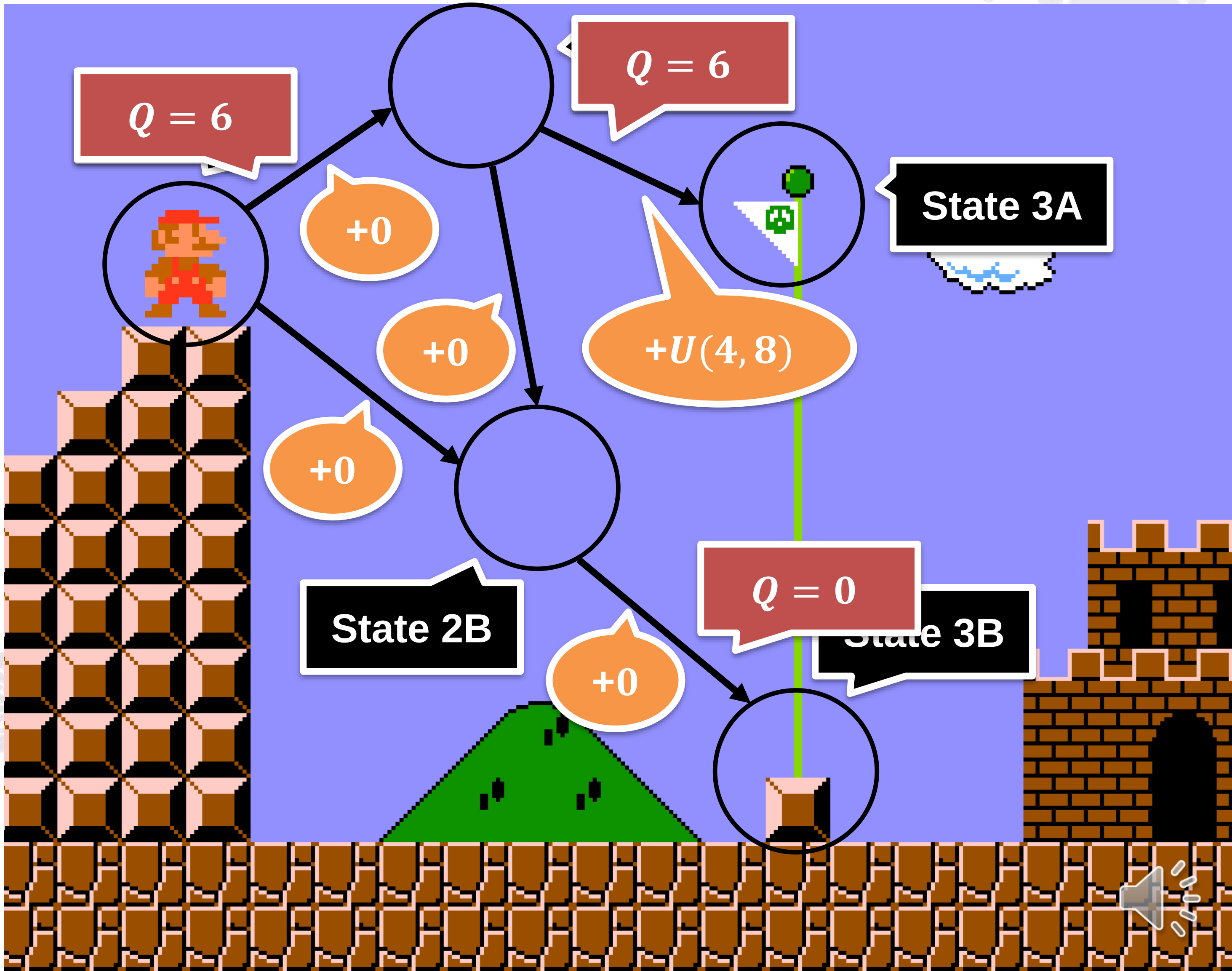


Q-Learning w/ Bellman Operator

Reinforcement Learning (Focus on Q-Learning)

Single-Agent RL (SARL)

Approximation Goal: Approximate the optimal expected reward of a state-action pair



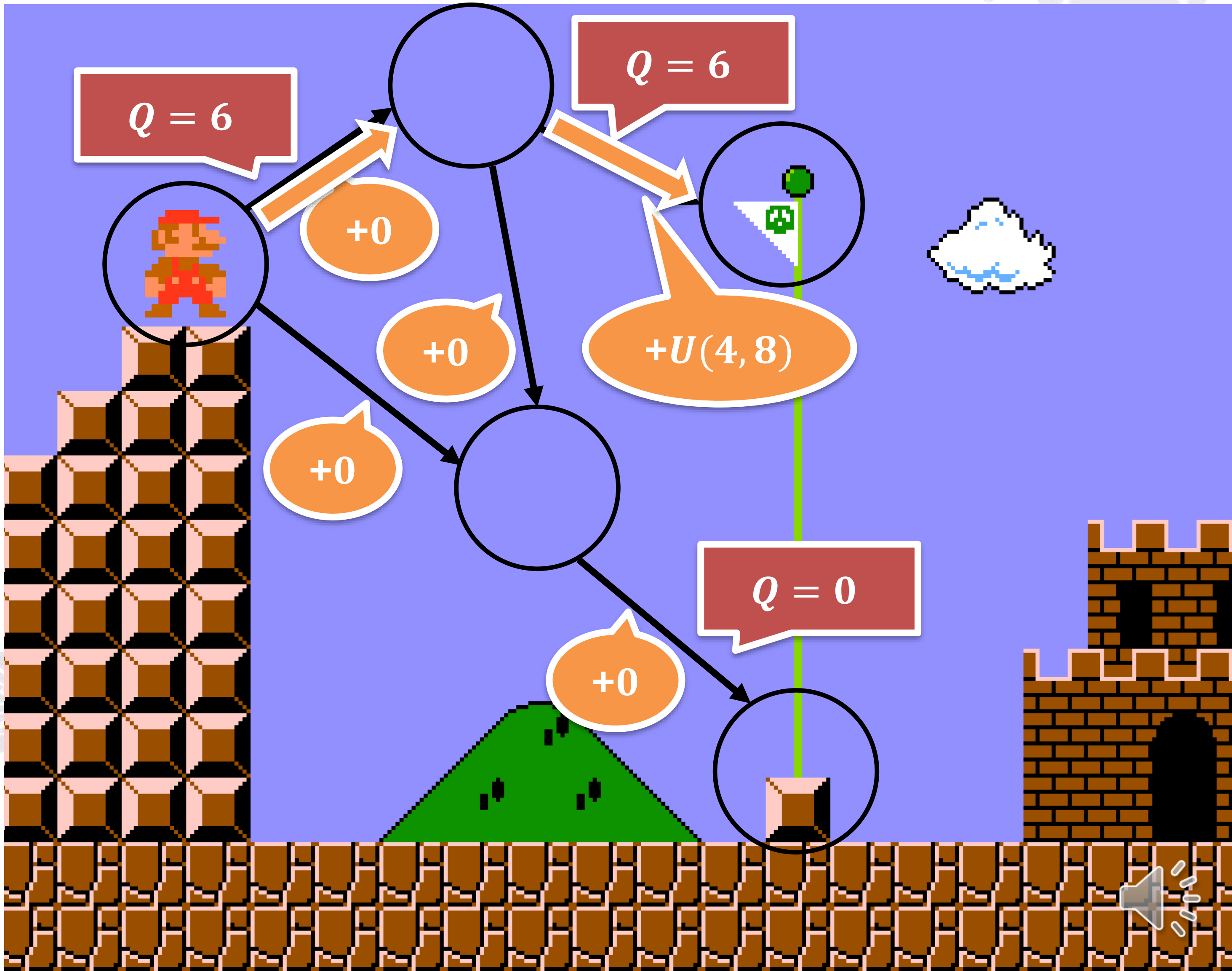
Reinforcement Learning (Focus on Q-Learning)

Single-Agent RL (SARL)

Prone to Overfitting

Optimal Policy

Methodology: Select the action with largest Q-Value for the current state



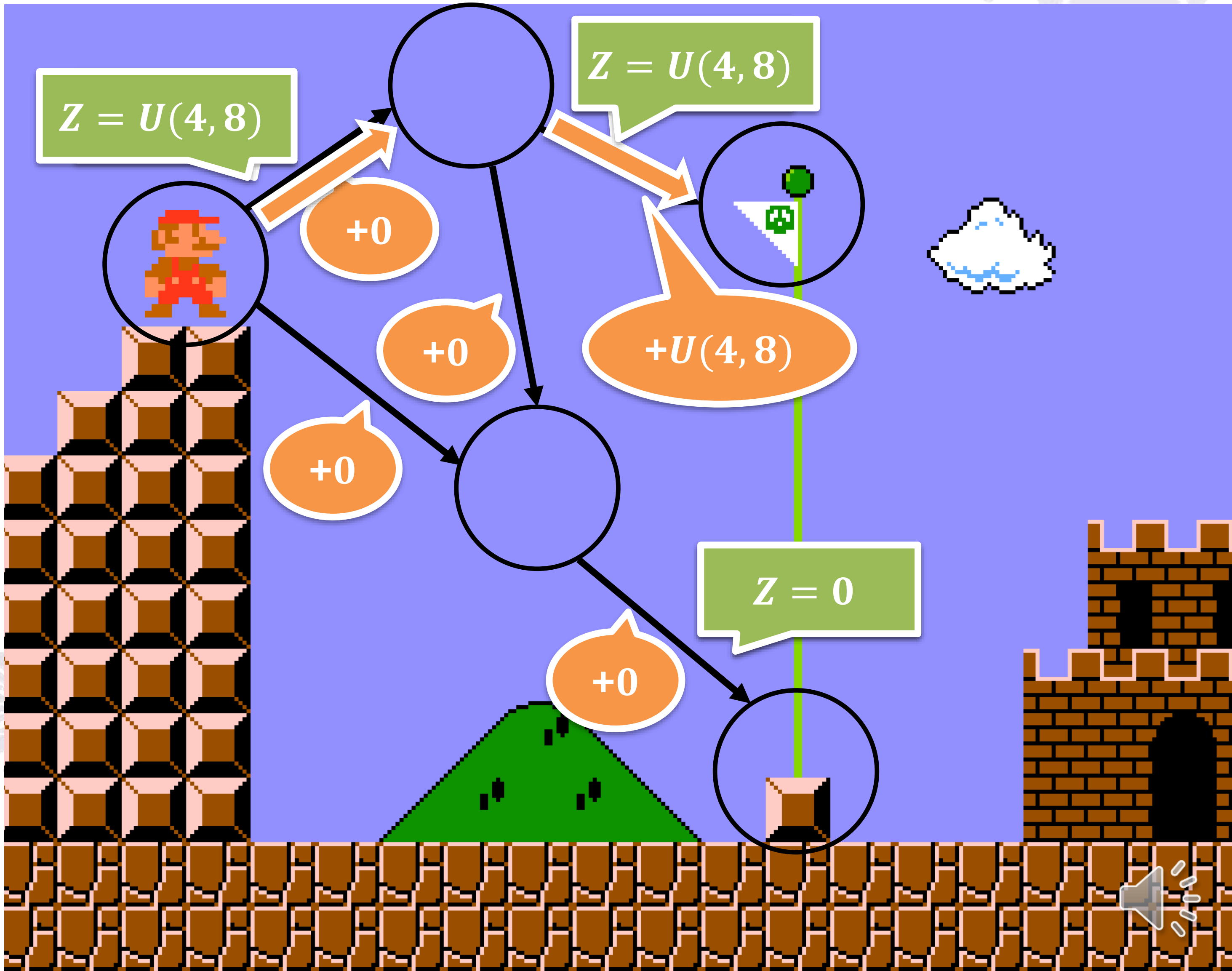
Reinforcement Learning
(Focus on Q-Learning)

Single-Agent RL
(SARL)

Prone to Overfitting

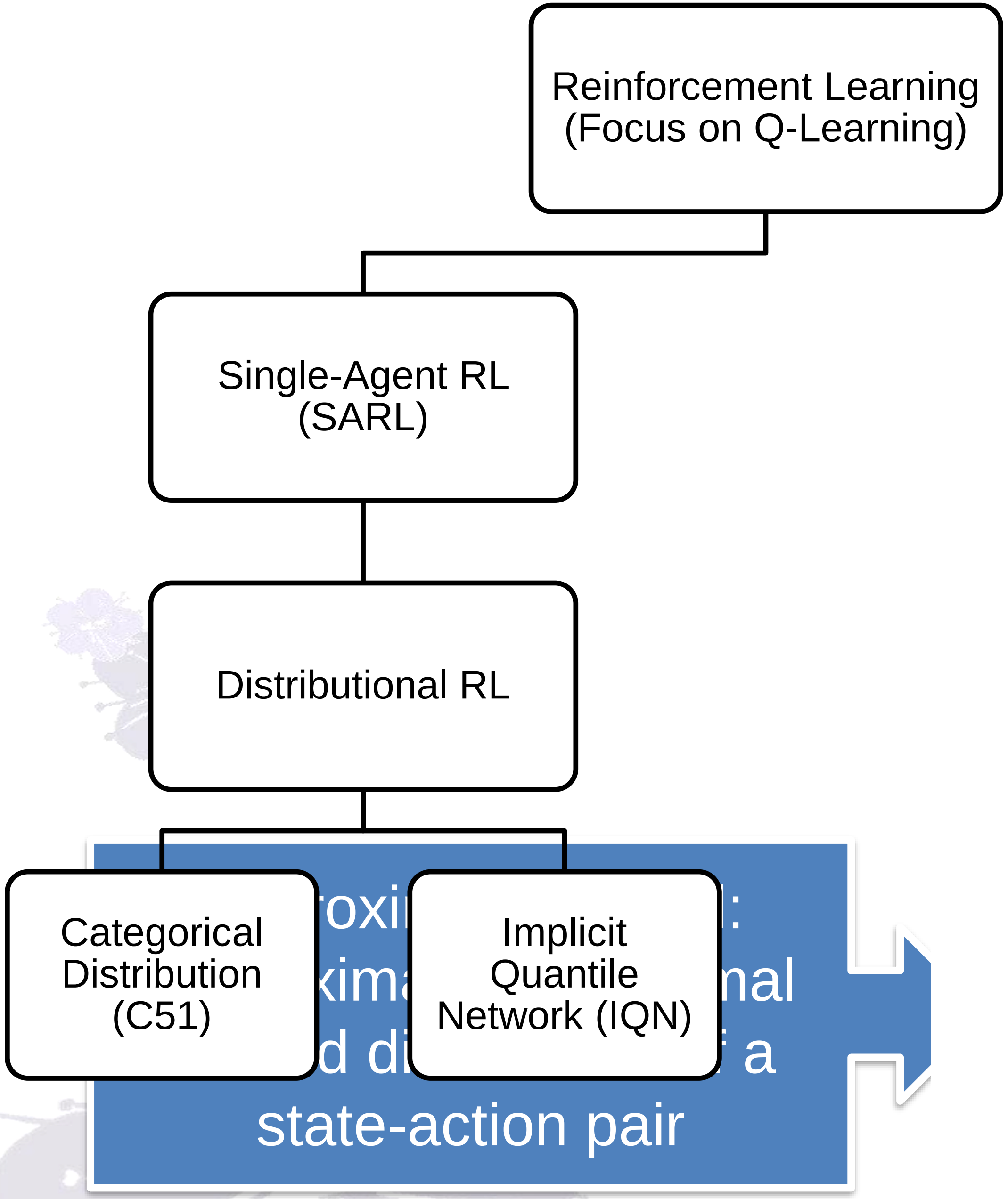
Distributional Q-Learning

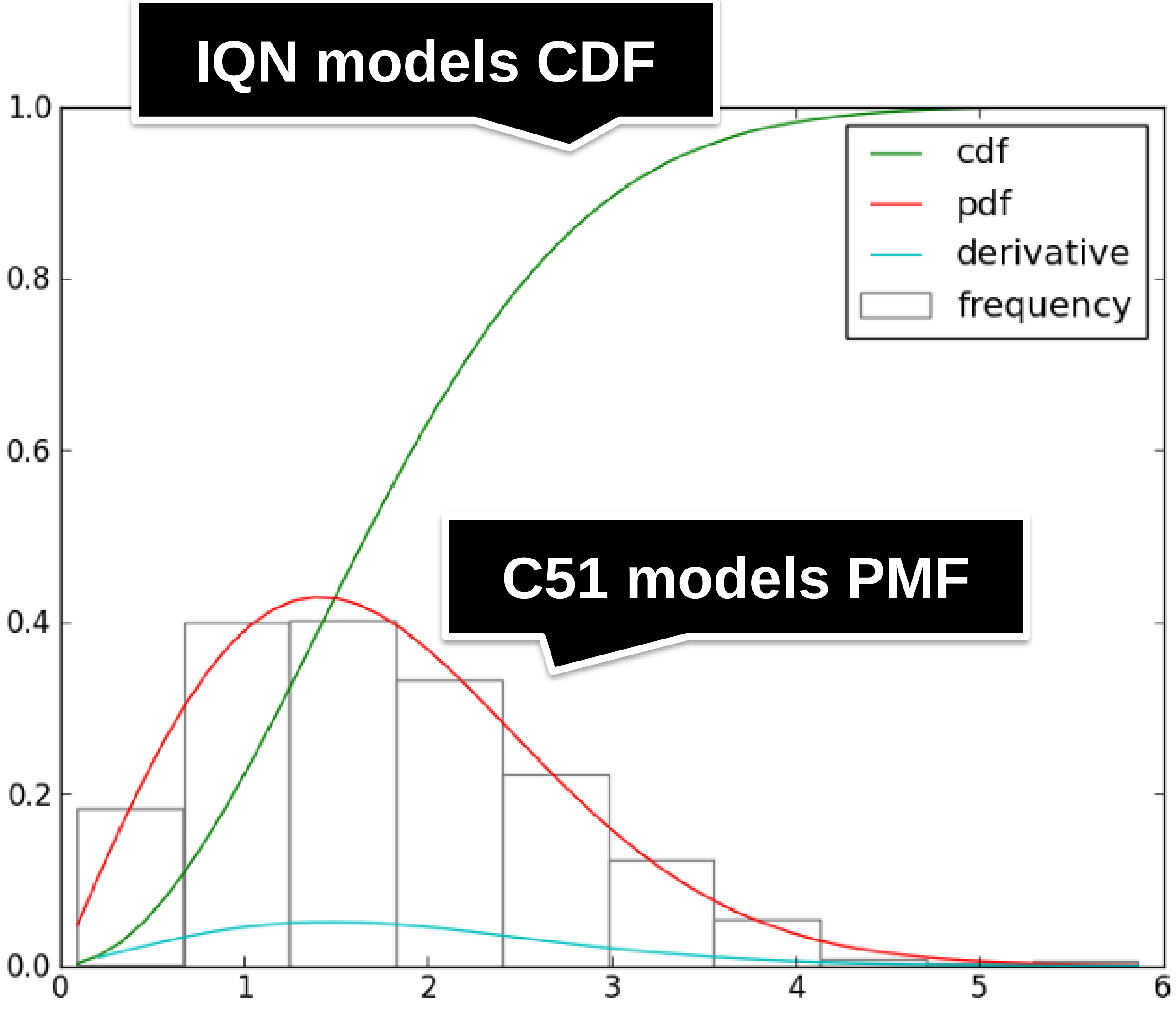
Approximation Goal: Approximate the optimal reward distribution of a state-action pair



Reduce Overfitting

Distributional Q-Learning





Fully Cooperative
Team Reward (not Individual Reward)
⇒ Have Issues in Q-Learning

Decentralized Partial Observable MDP

(Focus on Q-Learnin

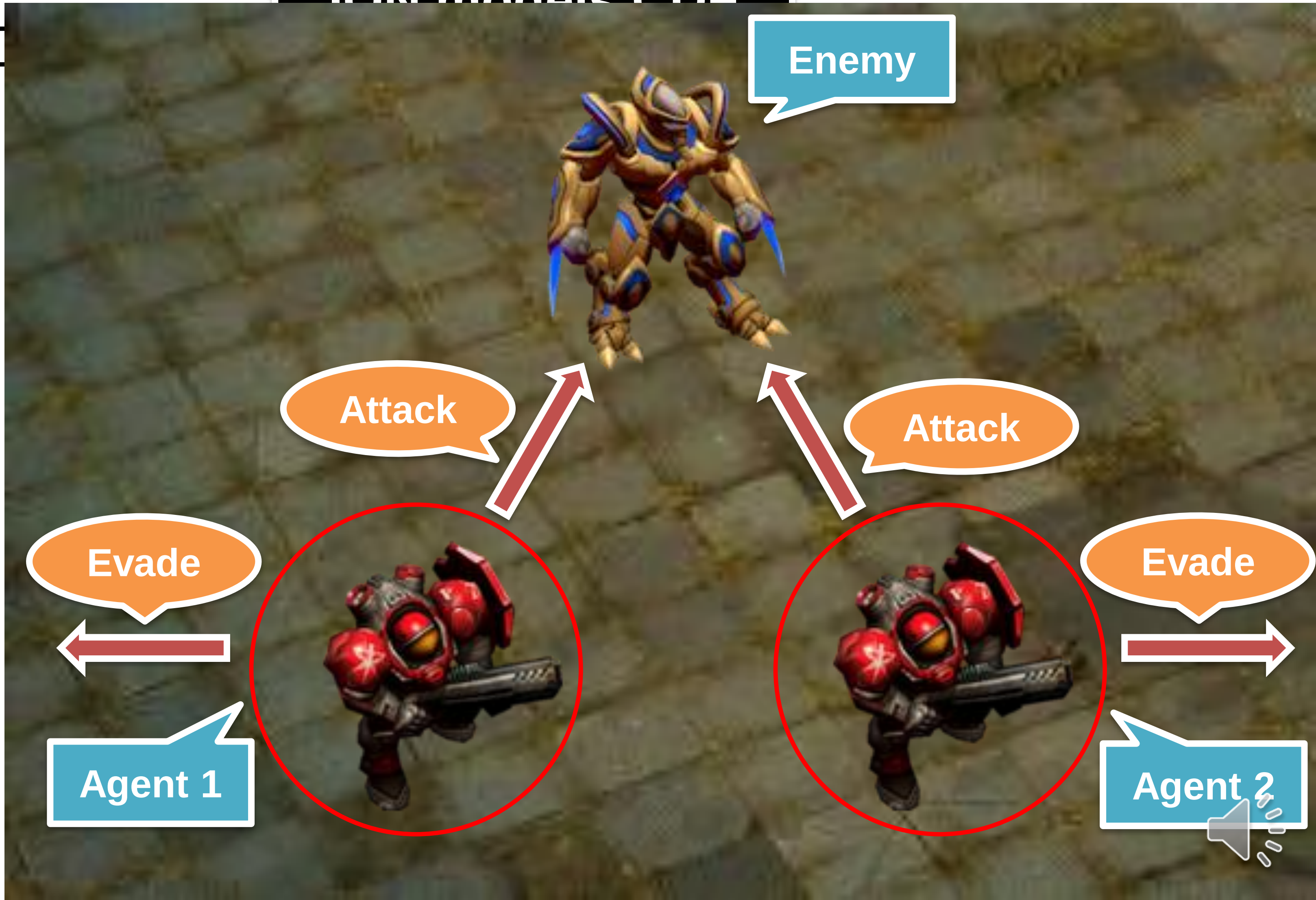
IQN models CDF

Single-Agent RL (SARL)

Distributional RL

Categorical Distribution (C51)

Implicit Quantile Network (IQN)

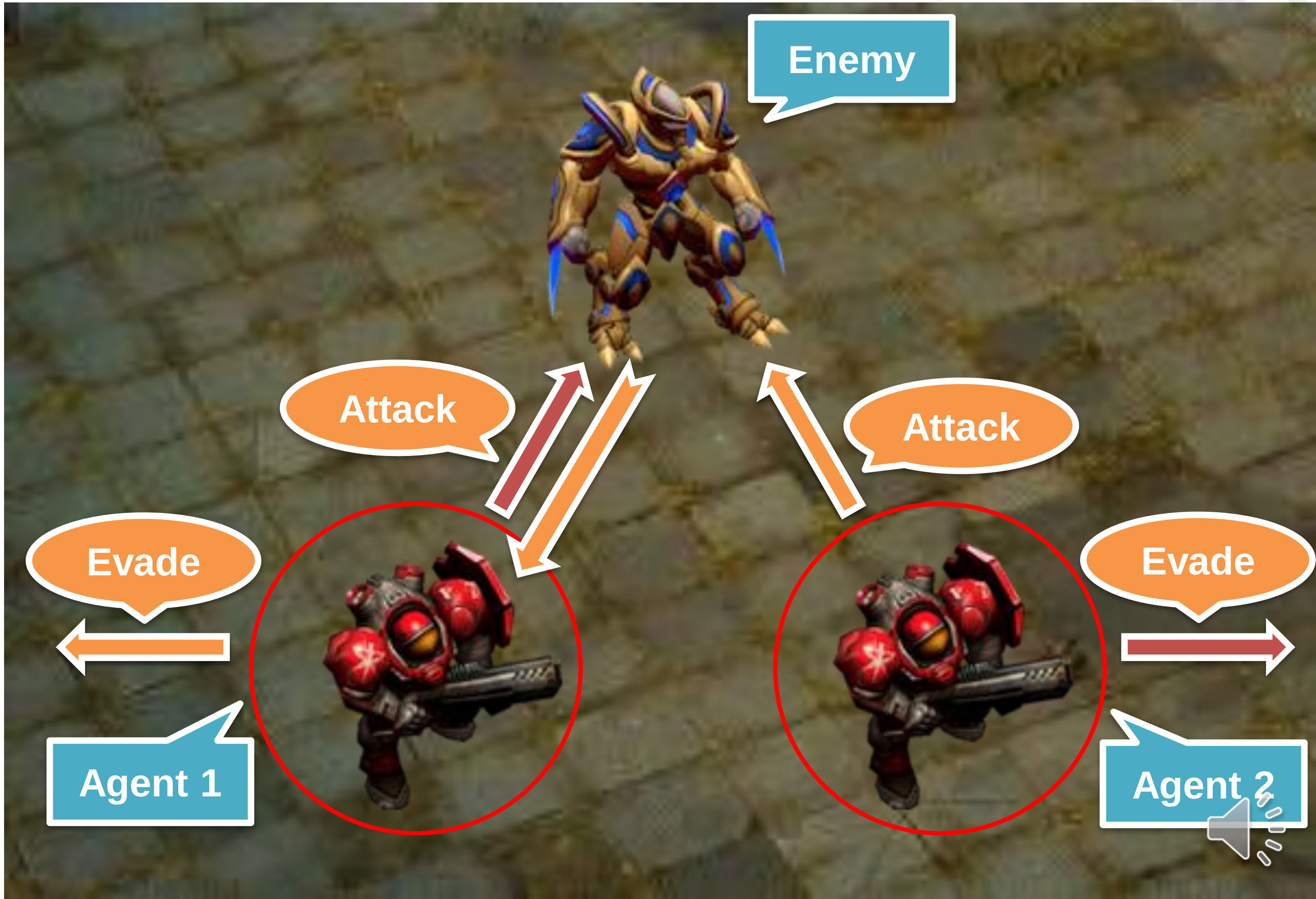


Enemy attacking Ag		
	Attack 2	
Attack 1	+5	+0
Evade 1	+10	+5

Fully Cooperative
Team Reward (not Individual Reward)
⇒ Have Issues in Q-Learning

Decentralized Partial Observable MDP

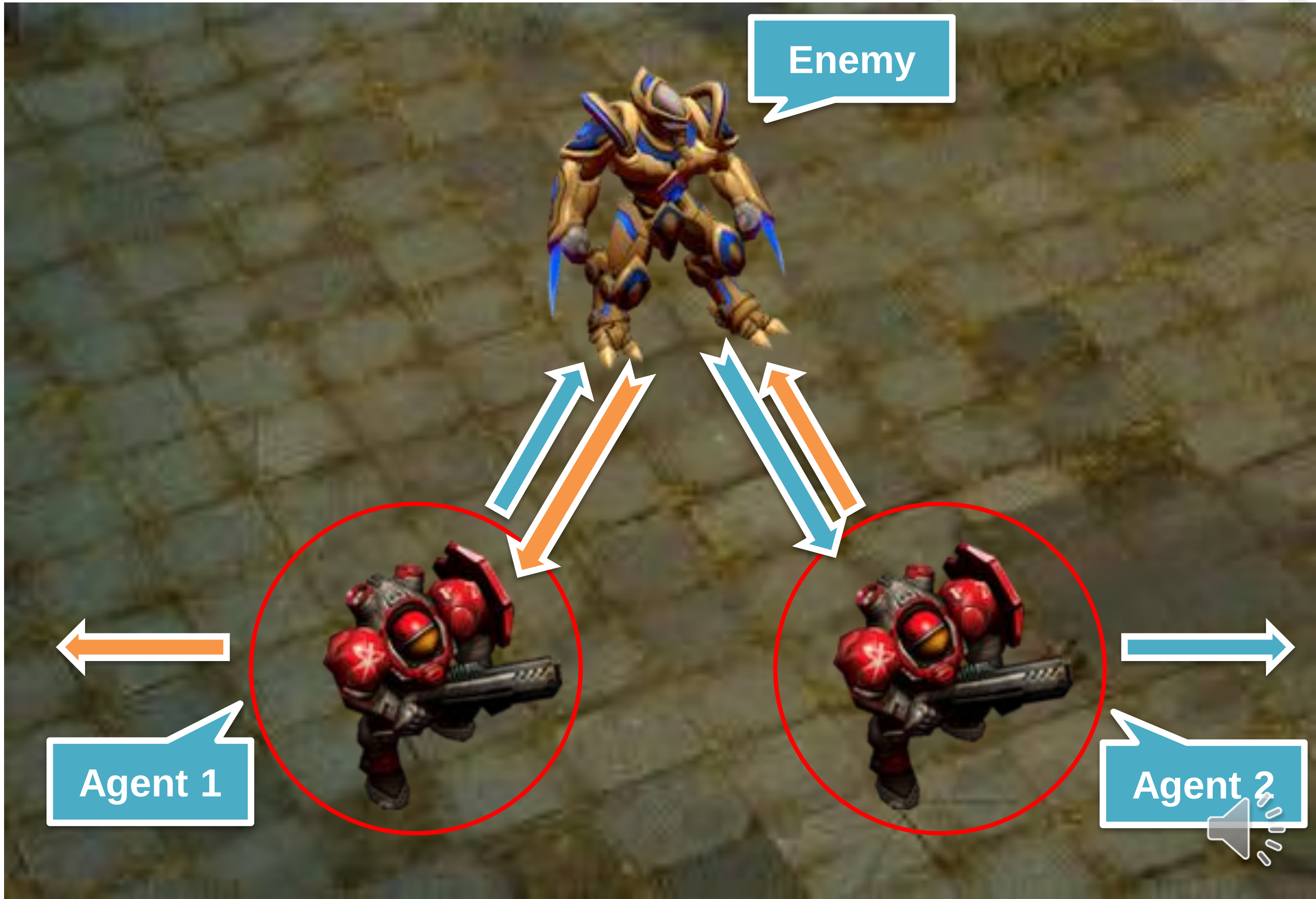
Fully Cooperative Multi-Agent RL (MARL)



Enemy attacking Agent 1 (Left)		
	Attack 2	Evade 2
Attack 1	+5	+0
Evade 1	+10	+5

Fully Cooperative Multi-Agent RL (MARL)

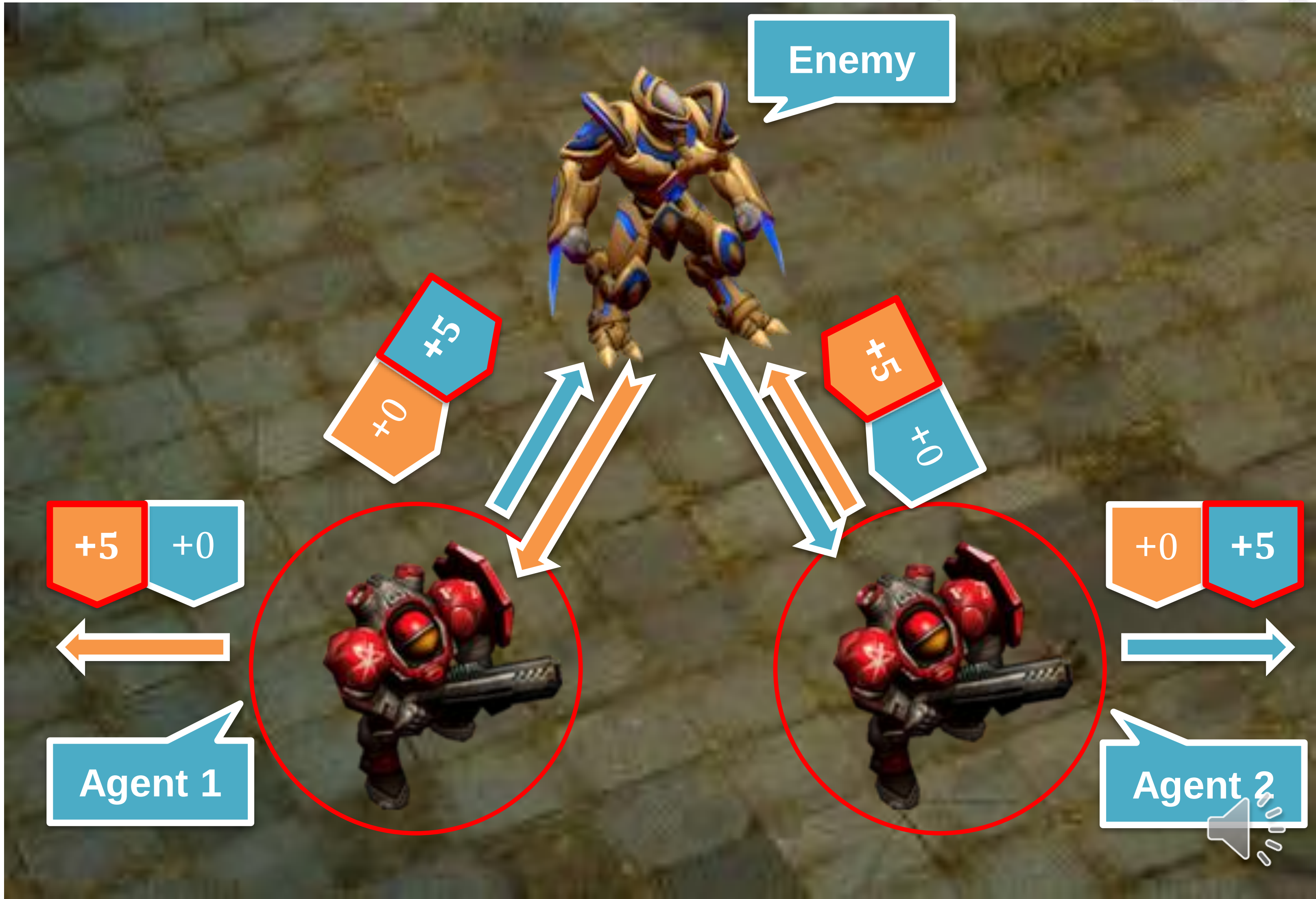
Enemy attacking Agent 2 (Right)		
	Attack 2	Evade 2
Attack 1	+5	+10
Evade 1	+0	+5



Enemy attacking Agent 1 (Left)		
	Attack 2	Evade 2
Attack 1	+5	+0
Evade 1	+10	+5
(Decomposed)		
	Attack 2	Evade 2
Attack 1	+0 / +5	+0 / +0
Evade 1	+5 / +5	+5 / +0

Enemy attacking Agent 2 (Right)		
	Attack 2	Evade 2
Attack 1	+5	+10
Evade 1	+0	+5
(Decomposed)		
	Attack 2	Evade 2
Attack 1	+5 / +0	+5 / +5
Evade 1	+0 / +0	+0 / +5

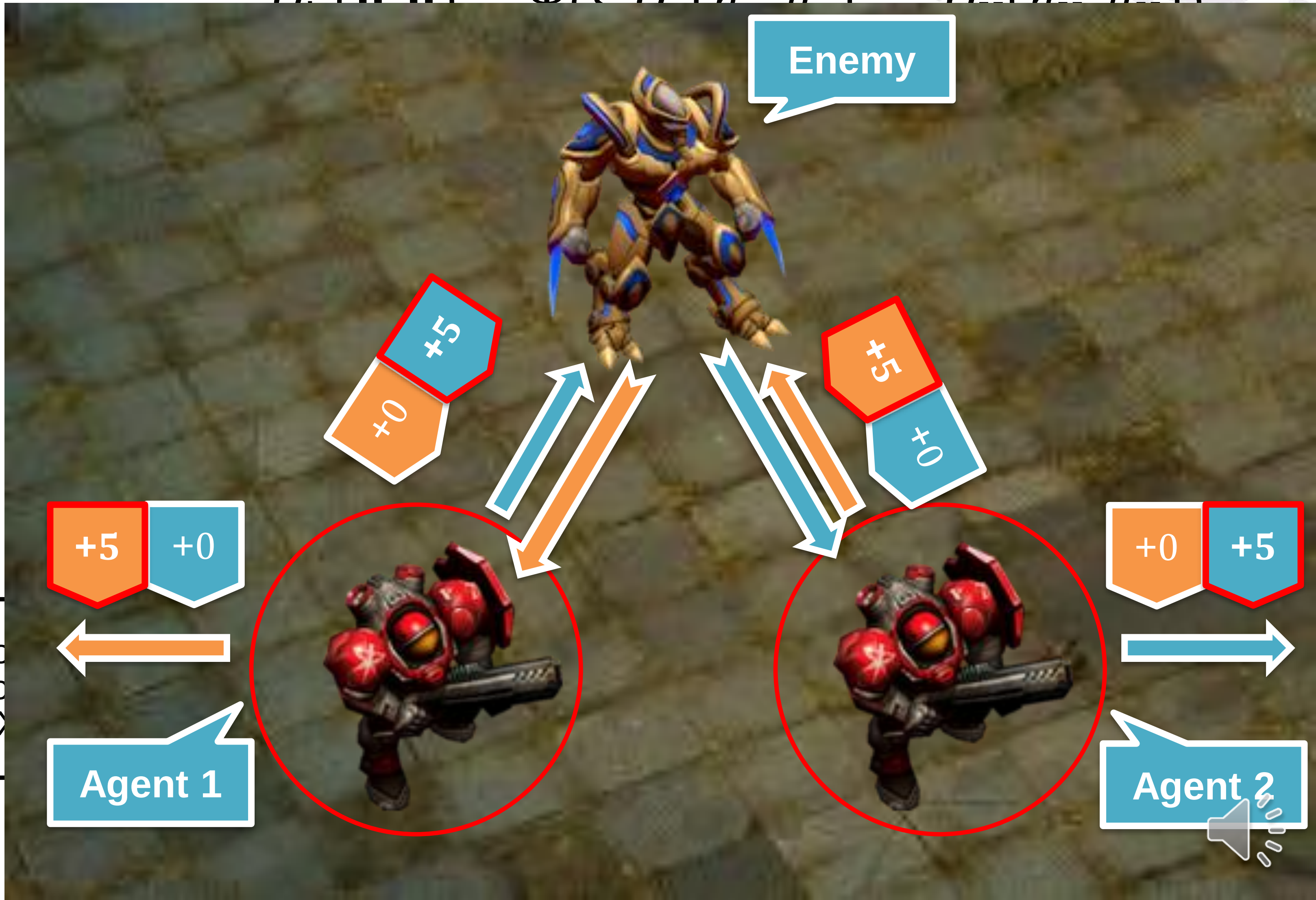
Decentralized Control → Value Factorization

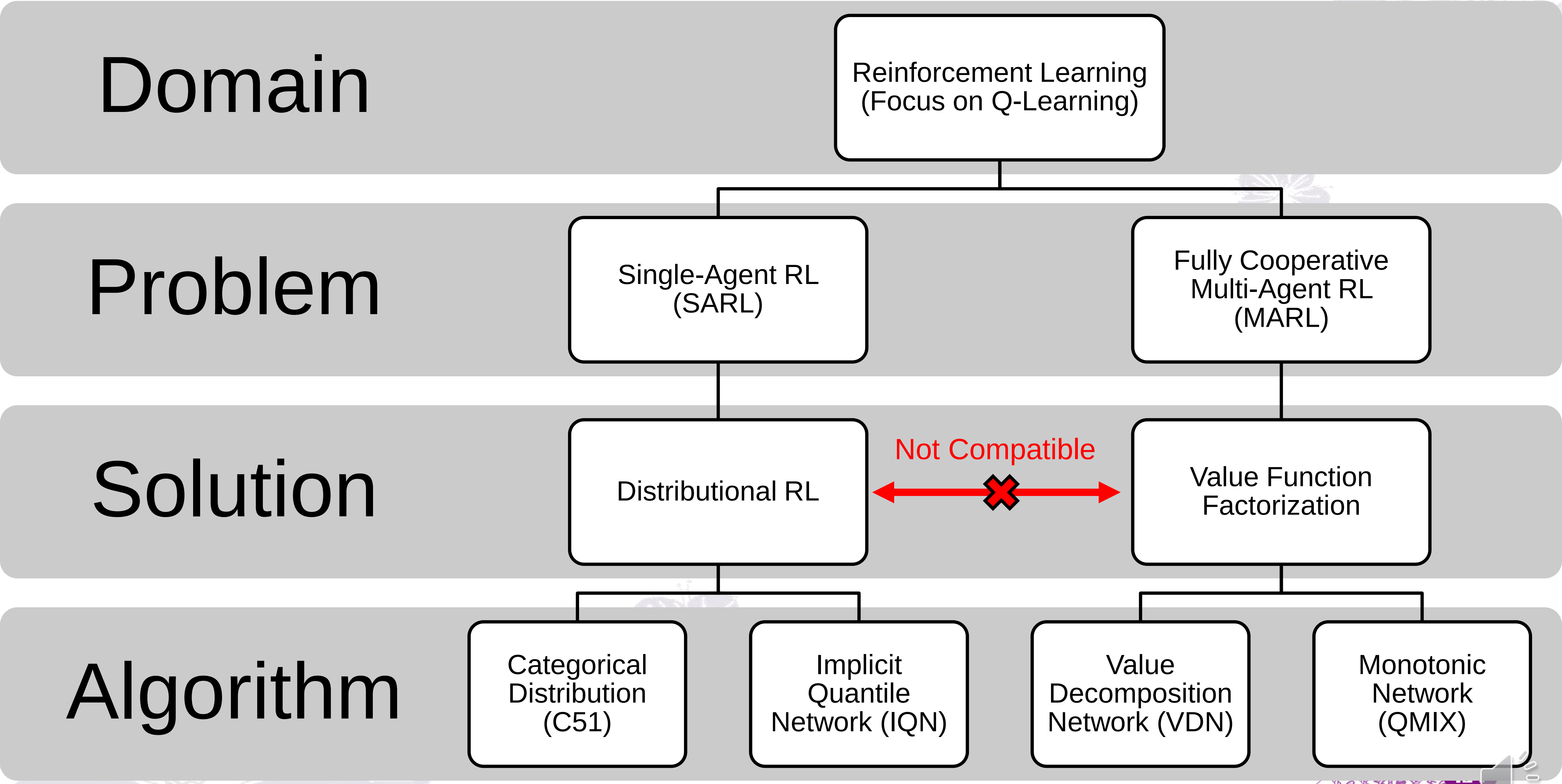


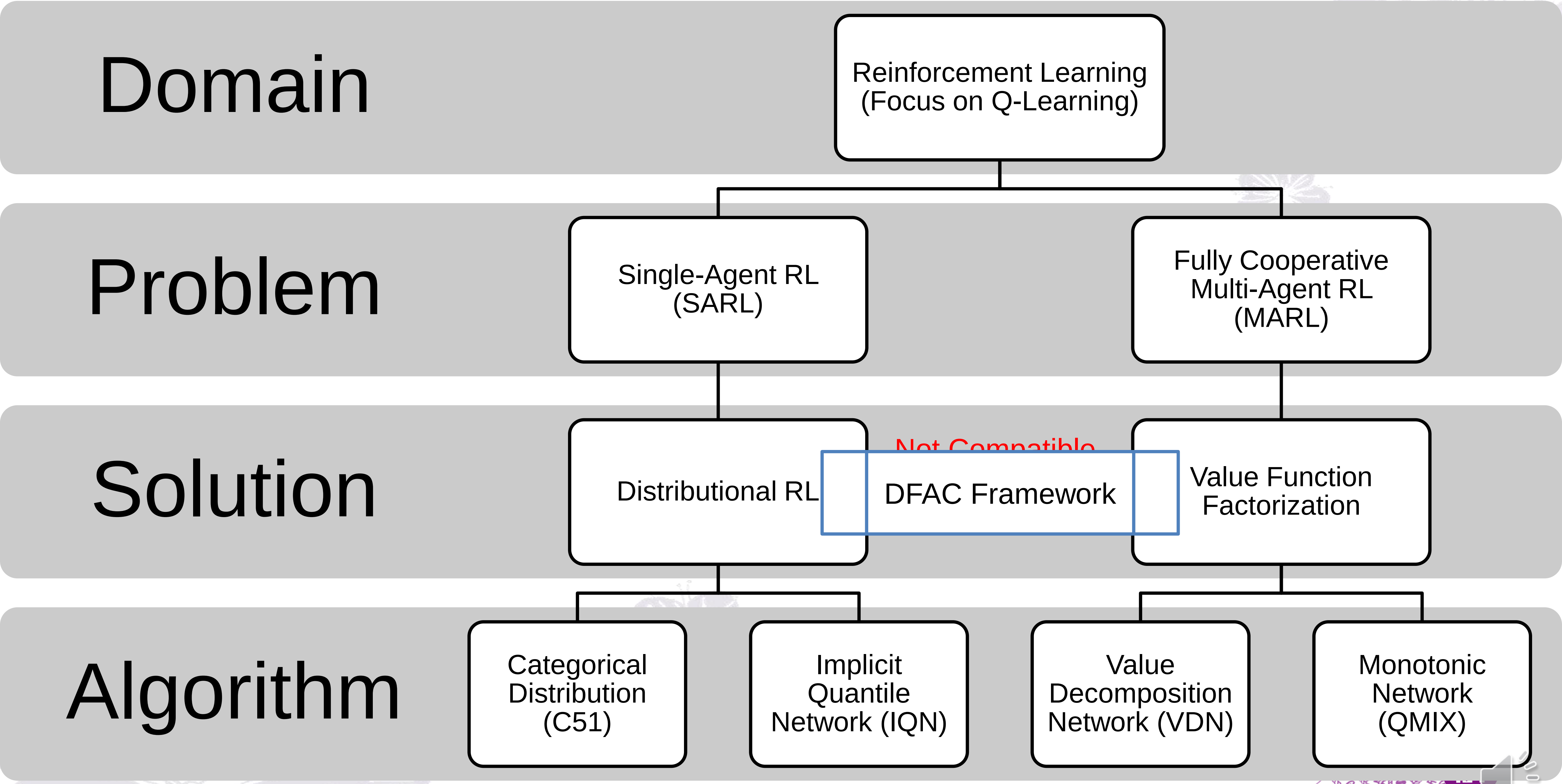
Decentralized Control → Value Factorization

Enemy attacking Agent 1 (Left)		
	Attack 2	Evade 2
Attack 1	+5	+0
Evade 1	**+10**	+5
(Decomposed)		
	Attack 2	Evade 2
Attack 1	+0 / +5	+0 / +0
Evade 1	**+5 / +5**	+5 / +0

Enemy attacking Agent 2 (Right)		
	Attack 2	Evade 2
Attack 1	+5	**+10**
Evade 1	+0	+5
(Decomposed)		
	Attack 2	Evade 2
Attack 1	+5 / +0	**+5 / +5**
Evade 1	+0 / +0	+0 / +5




Domain
Problem
Solution
Algorithm
Reinforcement Learning
(Focus on Q-Learning)
Single-Agent RL
(SARL)
Fully Cooperative
Multi-Agent RL
(MARL)
Distributional RL
Not Compatible
Value Function
Factorization
Categorical
Distribution
(C51)
Implicit
Quantile
Network (IQN)
Value
Decomposition
Network (VDN)
Monotonic
Network
(QMIX)


Domain
Problem
Solution
Algorithm
Reinforcement Learning
(Focus on Q-Learning)
Single-Agent RL
(SARL)
Fully Cooperative
Multi-Agent RL
(MARL)
Distributional RL
Not Compatible
DFAC Framework
Value Function
Factorization
Categorical
Distribution
(C51)
Implicit
Quantile
Network (IQN)
Value
Decomposition
Network (VDN)
Monotonic
Network
(QMIX)

Outline

- Background & Motivation
- **Proposed Method: DFAC**
- Experiment Results: Outperform all baselines

Theorem 1: Naïve generalization does not satisfy IGM

Given a factorization function Ψ that satisfies IGM in the following form:

$$Q_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) = \Psi(s, Q_1(h_1,u_1),\dots,Q_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}}))$$

The condition above is **not enough to guarantee** that

$$Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) = \Psi(s, Z_1(h_1,u_1),\dots,Z_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}}))$$

Satisfies IGM for random variables.

Theorem 2: Mean-Shape Decomposition satisfies IGM

Mean-Shape Decomposition:

$$Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) = \mathbb{E}\left[Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u})\right] + \left(Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u}) - \mathbb{E}\left[Z_{\mathrm{jt}}(\mathbf{h},\mathbf{u})\right]\right)$$
$$= Z_{\mathrm{mean}}(\mathbf{h},\mathbf{u}) + Z_{\mathrm{shape}}(\mathbf{h},\mathbf{u})$$

where

- $Z_{\mathrm{mean}}(\mathbf{h},\mathbf{u}) = \Psi\left(s, Q_1(h_1,u_1),\dots,Q_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}})\right)$
- $Z_{\mathrm{shape}}(\mathbf{h},\mathbf{u}) = \Phi\left(s, Z_1(h_1,u_1),\dots,Z_{\mathrm{K}}(h_{\mathrm{K}},u_{\mathrm{K}})\right)$
- Ψ satisfies IGM for $[Q_k]_{k=1}^{\mathrm{K}}$, $\mathrm{Var}(Z_{\mathrm{mean}}) = 0$, and $\mathbb{E}\left[Z_{\mathrm{shape}}\right] = 0$.

Mean-Shape decomposition is **guaranteed to satisfy IGM**.

Theorem 3: Quantile Mixture have the form of sum of random variables

Given a Quantile Mixture:

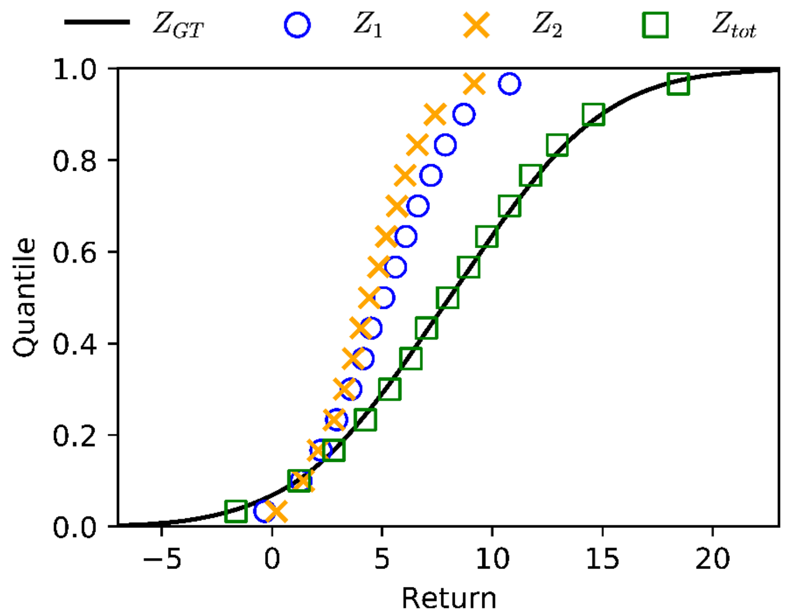
$$F_Z^{-1} = \sum_{k=1}^{K} \beta_k \cdot F_{Z_k}^{-1}$$

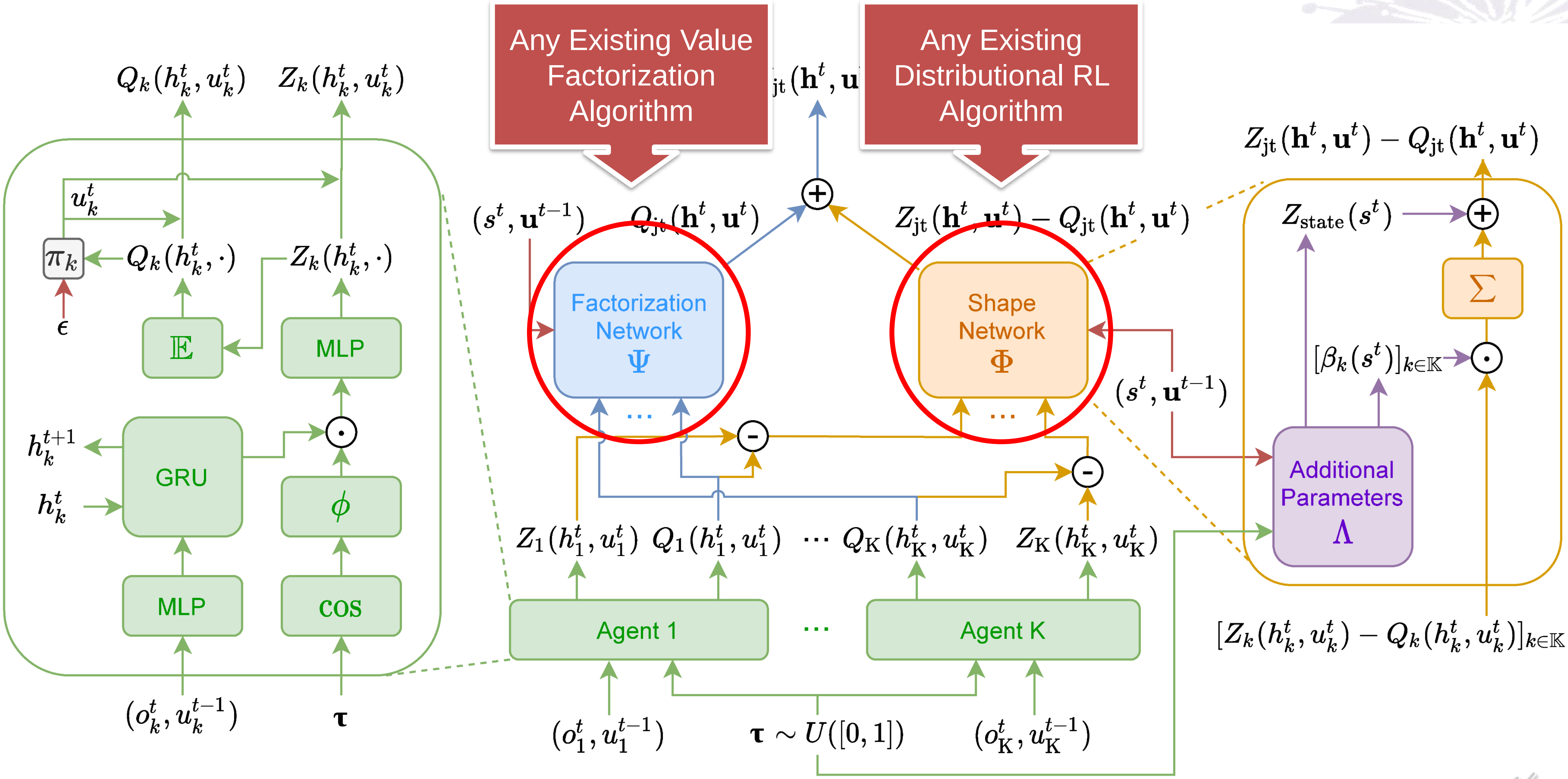
where $\beta_k \geq 0, \forall k$. There exist corresponding Z, $[Z_k]_{k=1}^{K}$ that satisfies:

$$\boldsymbol{Z = \sum_{k=1}^{K} \beta_k \cdot Z_k}$$

where the joint CDF of $[Z_k]_{k=1}^{K}$:

$$F_{\mathbf{Z}}(\mathbf{z}) = \min_k \left(F_{Z_k}^{-1}(z_k) \right)$$




Any Existing Value Factorization Algorithm
Any Existing Distributional RL Algorithm
$Q_k(h_k^t, u_k^t)$
$Z_k(h_k^t, u_k^t)$
u_k^t
π_k
$Q_k(h_k^t, \cdot)$
$Z_k(h_k^t, \cdot)$
ϵ
$\mathbb{E}$
MLP
h_k^{t+1}
h_k^t
GRU
ϕ
MLP
cos
(o_k^t, u_k^{t-1})
$\boldsymbol{\tau}$
$(s^t, \mathbf{u}^{t-1})$
$Q_{\text{jt}}(\mathbf{h}^t, \mathbf{u}^t)$
$Z_{\text{jt}}(\mathbf{h}^t, \mathbf{u}^t) - Q_{\text{jt}}(\mathbf{h}^t, \mathbf{u}^t)$
Factorization Network
Ψ
Shape Network
Φ
$(s^t, \mathbf{u}^{t-1})$
$Z_1(h_1^t, u_1^t)$ $Q_1(h_1^t, u_1^t)$ $\cdots$ $Q_K(h_K^t, u_K^t)$ $Z_K(h_K^t, u_K^t)$
Agent 1
Agent K
(o_1^t, u_1^{t-1})
$\boldsymbol{\tau} \sim U([0,1])$
(o_K^t, u_K^{t-1})
$Z_{\text{jt}}(\mathbf{h}^t, \mathbf{u}^t) - Q_{\text{jt}}(\mathbf{h}^t, \mathbf{u}^t)$
$Z_{\text{state}}(s^t)$
Σ
$[\beta_k(s^t)]_{k \in \mathbb{K}}$
Additional Parameters
Λ
$[Z_k(h_k^t, u_k^t) - Q_k(h_k^t, u_k^t)]_{k \in \mathbb{K}}$

DFAC Framework

For approximating $Z_{\mathrm{shape}} = \Phi$, we have the following choices:

(Assume we have K agents and N atoms/quantiles)

- C51 (models PMF)

  (convergence issue, not robust to hyperparameters, large network)

  - **Convolution** & Heuristic Projection ($\boldsymbol{O(KN^2)}$)
  - **FFT Convolution** + Heuristic Projection ($\boldsymbol{O(KN \log N)}$)

- IQN (models CDF)

  (convergence guarantee, robust to hyperparameters, light-weight)

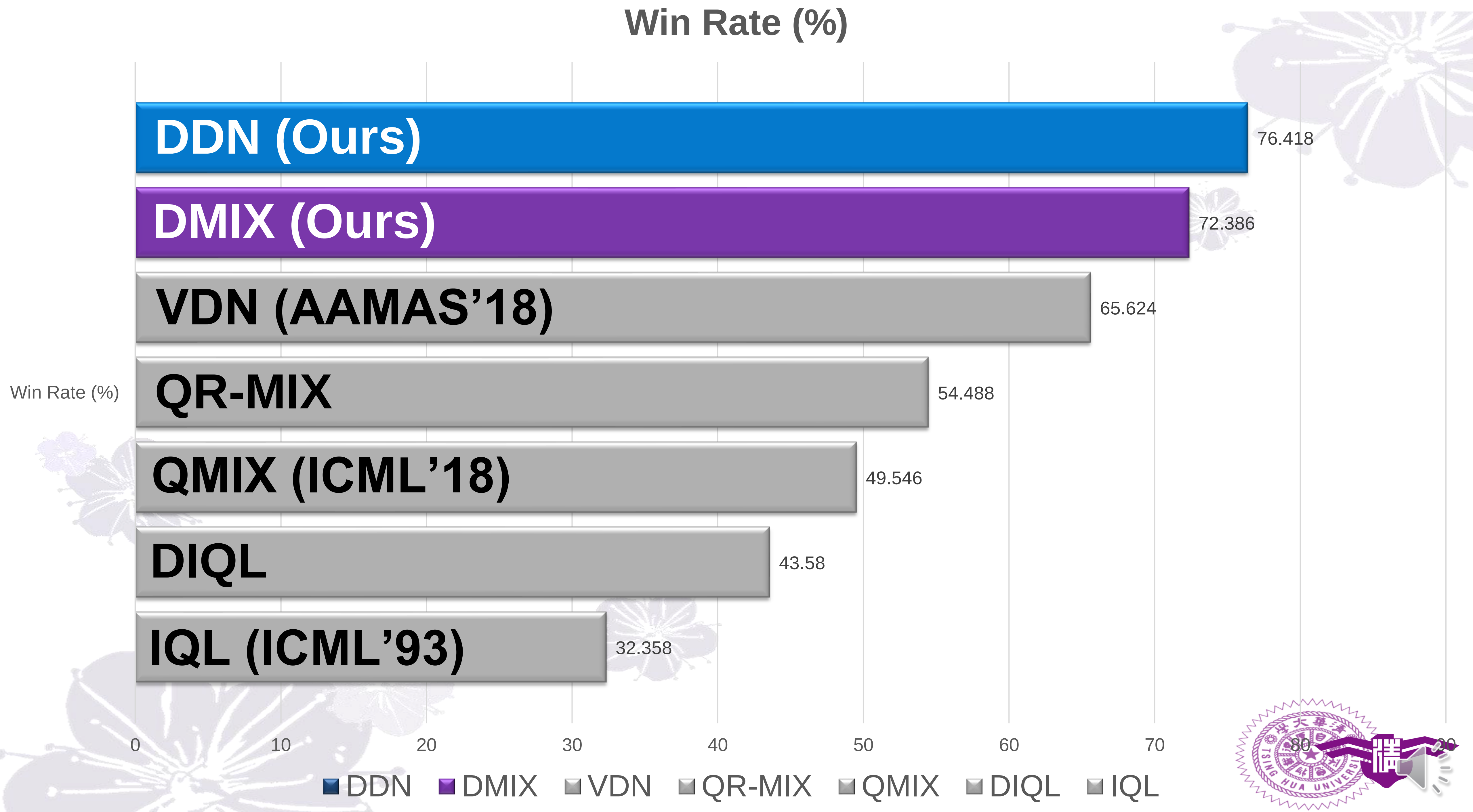
  - **Quantile Mixture** ($\boldsymbol{O(KN)}$)

Outline

- Background & Motivation
- Proposed Method: DFAC
- **Experiment Results: Outperform all baselines**

Experiment Results

	Map	IQL	VDN	QMIX	DIQL	**DDN**	**DMIX**
Win Rate (%)	6h_vs_8z	0.00%	0.00%	8.81%	0.00%	**83.52%**	68.75%
	3s5z_vs_3s6z	7.67%	90.91%	65.06%	29.83%	**94.60%**	90.62%
	MMM2	69.32%	87.78%	92.33%	83.52%	**97.44%**	95.17%
	27m_vs_30m	1.70%	64.20%	86.08%	12.50%	**94.60%**	86.08%
	corridor	83.10%	85.23%	4.26%	92.05%	**95.45%**	90.06%
Total Reward	6h_vs_8z	13.96	15.49	14.02	14.98	**19.32**	17.81
	3s5z_vs_3s6z	15.48	19.77	20.06	17.42	20.68	**20.78**
	MMM2	17.47	19.32	19.45	19.21	**21.06**	19.69
	27m_vs_30m	13.95	18.49	19.46	15.16	**19.72**	19.40
	corridor	19.30	19.38	13.44	19.57	**19.97**	19.61



Win Rate (%)
DDN (Ours)
76.418
DMIX (Ours)
72.386
VDN (AAMAS’18)
65.624
QR-MIX
54.488
QMIX (ICML’18)
49.546
DIQL
43.58
IQL (ICML’93)
32.358
Win Rate (%)
0
10
20
30
40
50
60
70
80
DDN
DMIX
VDN
QR-MIX
QMIX
DIQL
IQL

Thank you!

For more information, please refer to the QR code below: