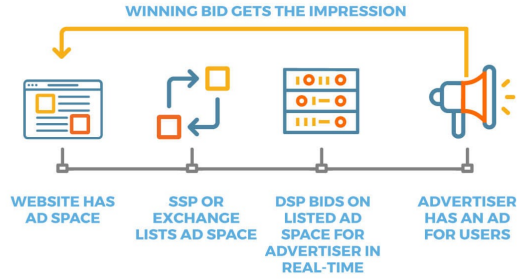
Hierarchical Multi-Agent Reinforcement Learning for Allocating Guaranteed Display Ads

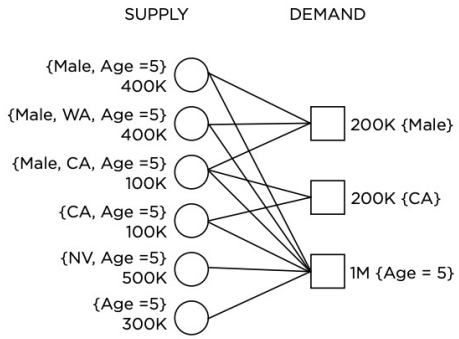
Lu Wang, Lei Han, Hao Chen, Xinru Chen, Chengchang Li, Junzhou Huang, Weinan Zhang, Wei Zhang, Xiaofeng He, Dijun Luo

Digital Advertising

- The total cost of digital advertising on world wide web is about $455.30 in 2021.
- Two major online ads
  - real-time bidding (RTB)
  - guaranteed display ads (GDAs)



- RTB
  - an advertiser submits a bid in auctions happening in real time.
- GDAs
  - an contract is signed to ensure a certain amount of ad impressions to be displayed to some targeted populations.



Allocating Guaranteed Display Ads

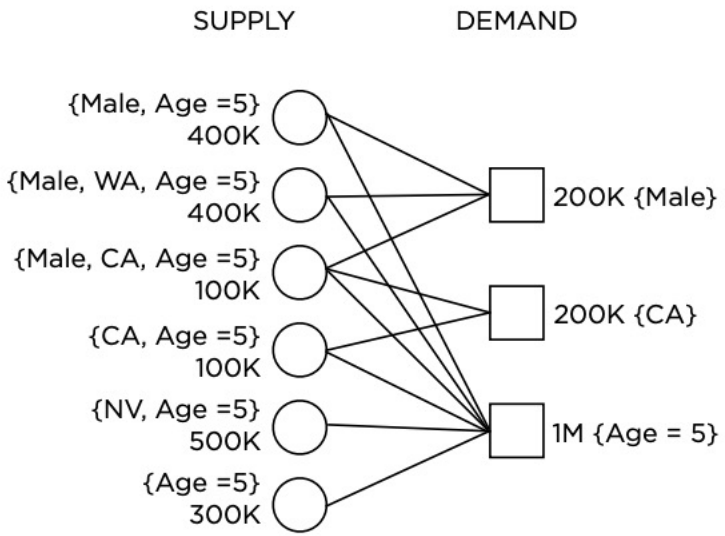
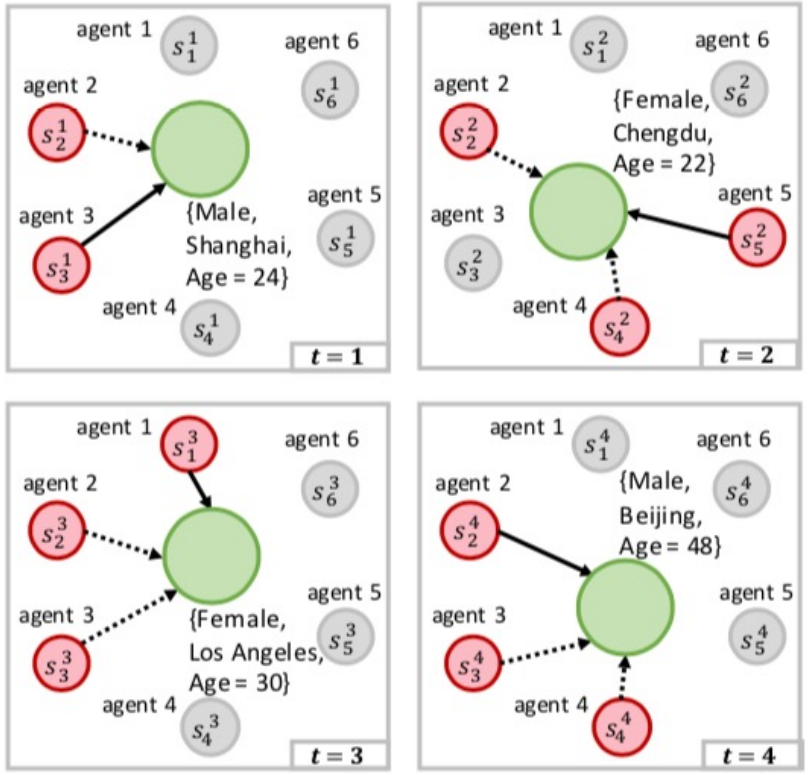


TABLE I: A case when advertisers compete.

	PID		Better decision	
	Ad1	Ad2	Ad1	Ad2
Impression 1	0.5	0.5	1	0
Impression 2	0	0.3	0	0.8

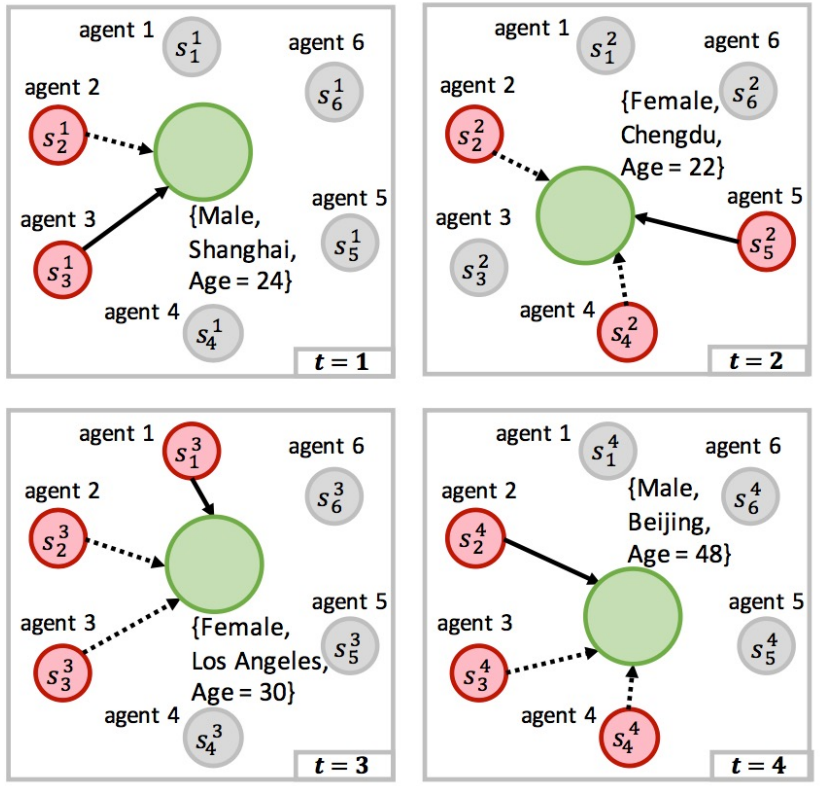


Static allocation optimization

PID: proportional-integral-derivative

Multi-agent Reinforcement Learning

Multi-agent Reinforcement Learning for GDAs

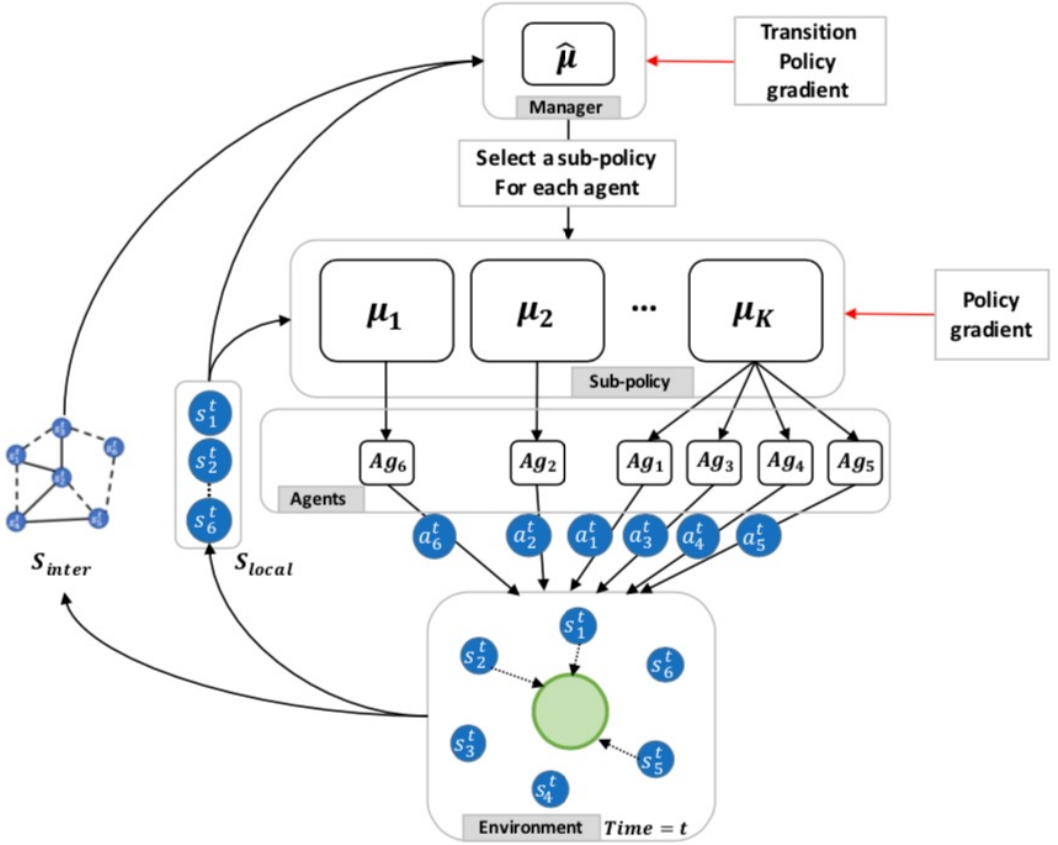


- **Agent:** each ad is with an allocation agent
- **State:** $s_t^n = [w^n, c_t^n, l_t^n, a_{t-1}^n]$
  - w^n: presents the degree of resource shortage of the contract $(w_n = d_n/S_n)$
  - c_t^n: indicates the distance between current time and the expiration time of the contract. $(c_t^n = t/T)$
  - l_t^n: presents the ratio of accumulated exposure divided by the demand until time step t $(l_t^n = \frac{\sum_{i=1}^{t} e_i^n)}{d_n}$
- **Action:** $a_t^n \in [0, 1]$ the ratio of displaying ad n.
- **Reward Function:**

$$r_t^n = \begin{cases} \lambda_z * \omega & \text{if } l_t^n \in (0.95 * \omega, 1.05 * \omega) \\ (1.05 - l_t^n) * \lambda_o * \omega & \text{if } l_t^n >= 1.05 * \omega \\ -\lambda_{u1} * \omega & \text{if } l_t^n < 0.5 * \omega \\ -\lambda_{u2} * \omega & \text{if } l_t^n <= 0.95 * \omega \end{cases}$$

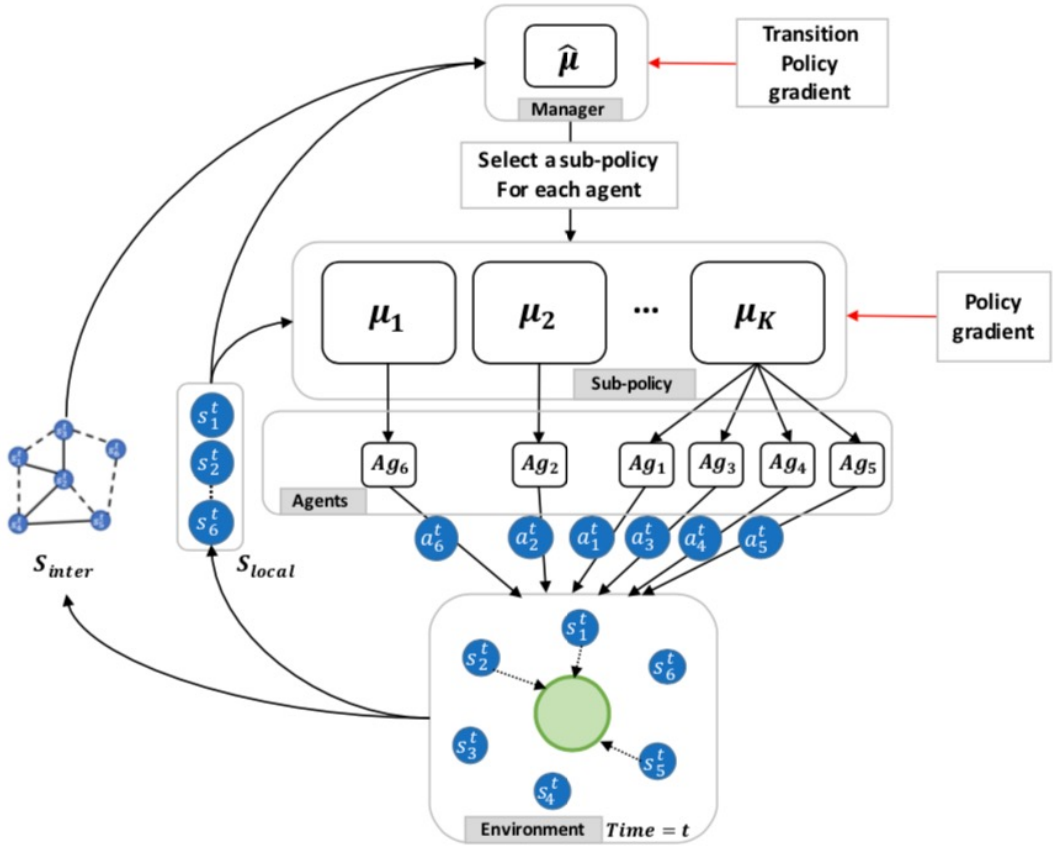
Hierarchical Multi-Agent Reinforcement Learning

- **Hierarchical MARL**
  - **All the agent share one policy**
    - Not Flexible
  - **Each agent learns one policy**
    - Parameter complexity is large
  - **Our method: N agents share K policies**
    - **K << N**



Hierarchical MARL for GDAs

- **Manager Policy**
  - **Chooses a sub-policy for each agent every V time steps.**
- **Sub-Policy**
  - **Each ad uses the selected sub-policy to take an action to determine whether display the ad.**



Hierarchical MARL for GDAs

- **Manager Policy**

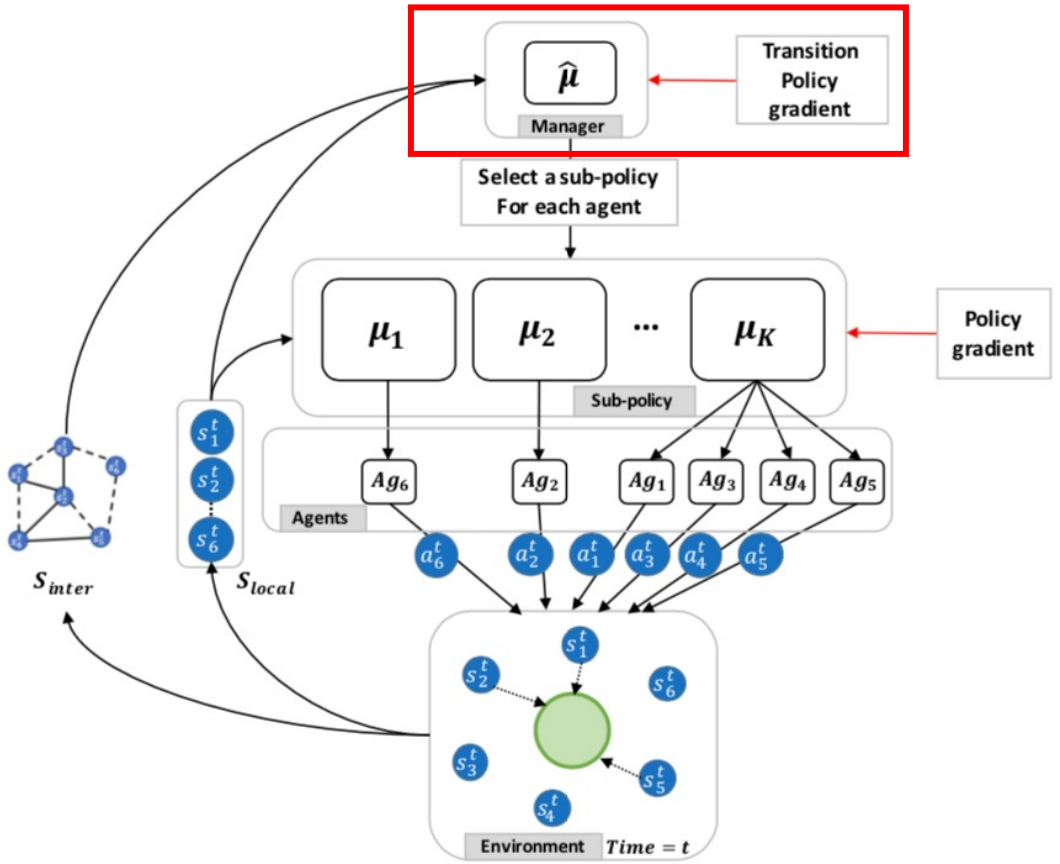
$$J(\hat{\theta}) = \sum_{n=1}^{N} \mathbb{E}_{\hat{s}_t^n, \hat{a}_t^n \sim \hat{\mu}_{\hat{\theta}}} [\sum_{t \in \Upsilon} \gamma^{\lfloor t/V \rfloor} \hat{r}_t^n(\hat{s}_t^n, \hat{a}_t^n)]$$

$$\Upsilon = \{1, 1+V, 1+2V, \dots, 1+T/V\}$$

  - **Choose a desired sub-policy for the agent based on interaction state and local state.**

$$p(\hat{s}_{t+V}^n \mid \hat{s}_t^n, \hat{a}_t^n)$$

  - **A multi-step transition probability function for the manager, which denotes the probability that action a causes the system to transform from state s_t^n to state s_{t+V}^n in V time steps.**

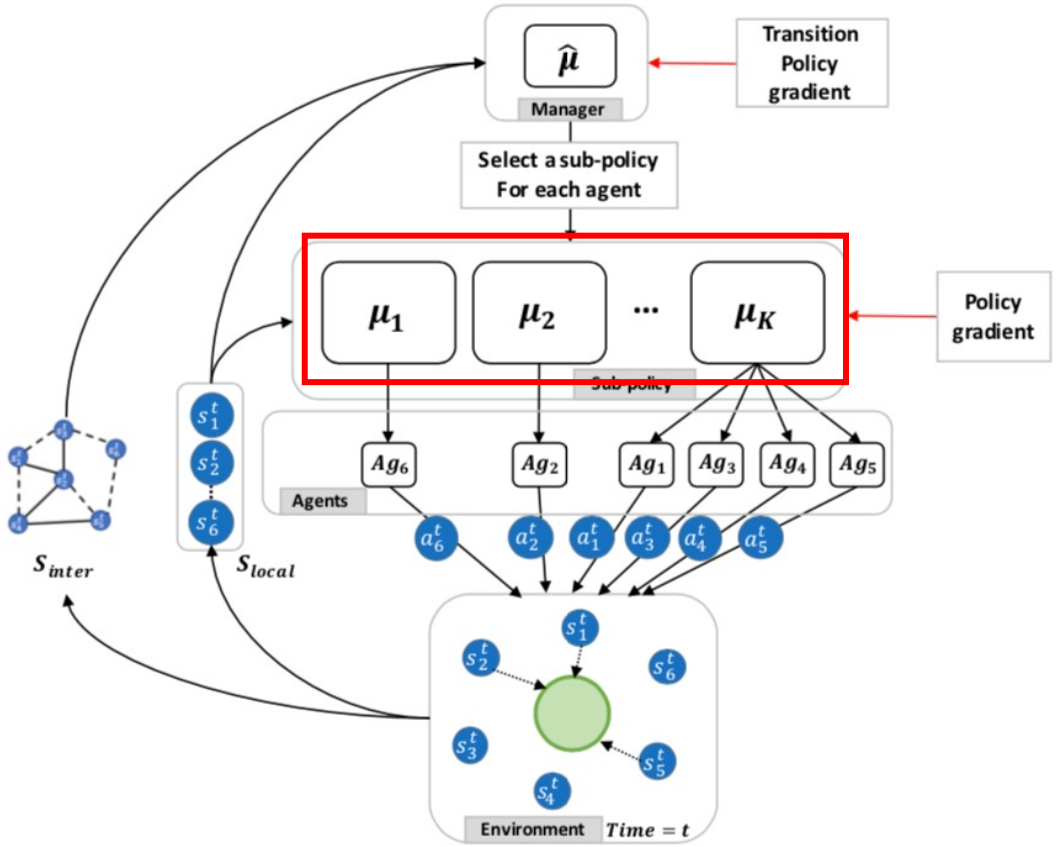


Hierarchical MARL for GDAs

- **Sub-Policy**

$$J(\theta_k) = \sum_{n=1}^{N} \mathbb{E}_{s_t^n, a_t^n \sim \mu_{\theta_k}(k=\arg\max \hat{a}_t^n)} [\sum_{t=1}^{T} \gamma^{t-1} r_t^n(s_t^n, a_t^n)] \quad (2)$$

  - **It's goal is to give a display ratio to obtain the impression for the ad.**



Experiment Results

- **Performance comparison on the three datasets with different periods for allocation of GDAs.**

Method	27/12/2018-03/01/2019		08/11/2018-11/11/2018		16/04/2018	
	UR	NR	UR	NR	UR	NR
Random	0.442	0.019	0.190	0.042	0.313	0.030
Static Action	0.728	0.090	0.913	0.086	0.813	0.171
DG	0.471	0.508	0.388	0.564	0.485	0.500
SG	0.388	0.601	0.526	0.435	0.724	0.261
PID	0.218	0.760	0.112	0.806	0.224	0.634
HWM	0.431	0.492	0.281	0.637	0.366	0.470
SHALE	0.502	0.472	0.221	0.702	0.244	0.606
LR	0.295	0.683	0.272	0.518	0.388	0.448
RFR	0.272	0.694	0.250	0.629	0.366	0.590
FeUdal	0.255	0.692	0.174	0.714	0.299	0.612
Flat_HMARL	0.237	0.716	0.142	0.761	0.225	0.647
HMARL	**0.138**	**0.835**	**0.058**	**0.875**	**0.089**	**0.813**

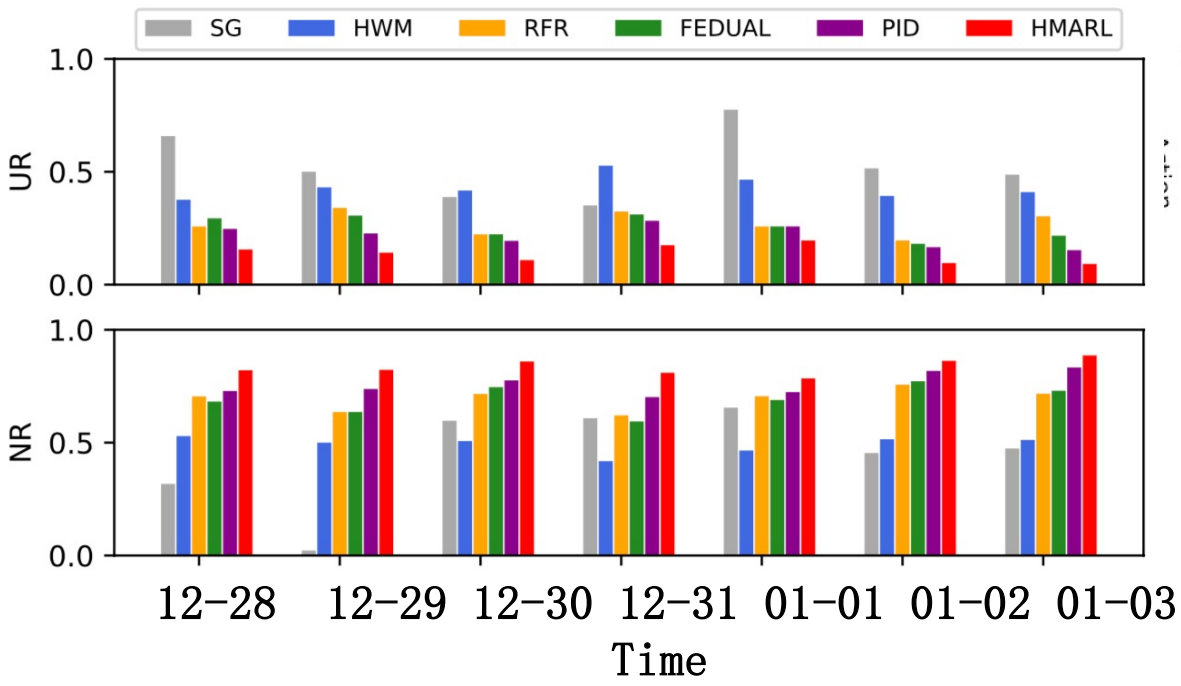
Tencent

State-of-the-art

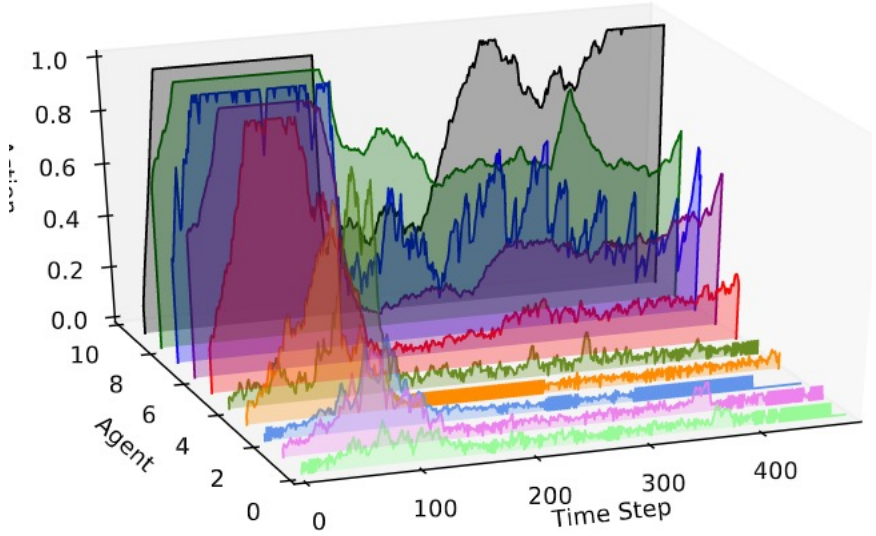
Our Method

Experiment Results

- **Visualization Results**



Performance on different days



Action distribution of selected agents

Thank You!