Fairness-Oracular MARL with Competitor-Aware Signals for Collaborative Inference

Hansong Zhou
Department of Computer Science
Florida State University
Tallahassee, FL 32304
hz21e@fsu.edu

Xiaonan Zhang
Department of Computer Science
Florida State University
Tallahassee, FL 32304
xzhang14@fsu.edu

Abstract

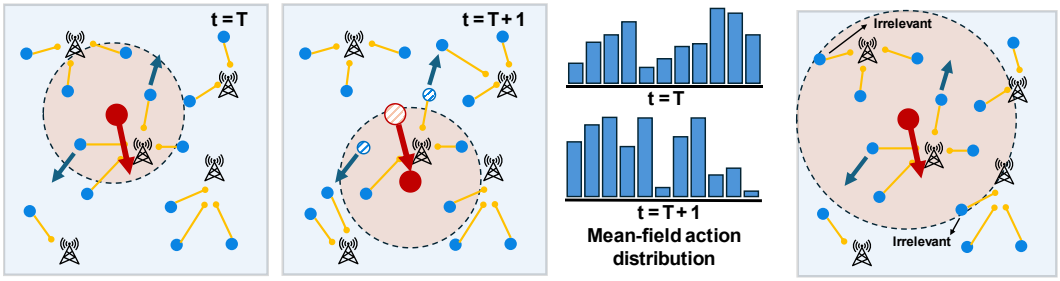
Collaborative inference (CI) in NextG networks enables battery-powered devices to collaborate with nearby edges on deep learning inference. The fairness issue in a multi-device, multi-edge (M2M) CI system remains underexplored. Mean-field multi-agent reinforcement learning (MFRL) is a promising solution due to its low complexity and adaptability to system dynamics. However, the mobile nature of M2M CI systems hinders their effectiveness, as it breaks the premise of stable mean-field statistics. We propose FOCI (Fairness-Oriented Collaborative Inference), an RL-based method with two components: (i) an oracle-shaping reward for approaching max-min fairness and (ii) a competitor-aware observation augmentation for stabilizing device behaviors. We provide a convergence guarantee with bounded estimation errors. According to the results from real-world device mobility traces, FOCI demonstrates the best performance across multiple metrics and tightens the tails. It reduces worst-case latency by up to 56% and worst-case energy by 46% compared to baselines, while halving the switch cost and preserving competitive QoS.

1 Introduction

Collaborative inference (CI) in mobile edge computing is attracting increasing attention in next-generation (NextG) wireless networks, where devices collaborate with nearby edge servers to perform deep learning inference. Such collaboration alleviates the computational burden on battery-powered devices, thereby extending their battery life and reducing service response latency Li et al. [2020]. In practice, such as city-wide AR assistants Khan et al. [2023], Wang et al. [2023], CI often manifests as multi-device–multi-edge (M2M) deployments, where devices autonomously request collaboration from edges and edges allocate resources accordingly. While most existing works optimize throughput or energy efficiency Li et al. [2023], Liu et al. [2023, 2024], few address fairness, i.e., improving the worst-off Quality-of-Service (QoS) across devices. Moreover, many solutions presume a central controller Tang et al. [2021] or rely on excessive peer information exchange Mohammed et al. [2020], both impractical at scale. To this end, we exploit multi-agent reinforcement learning (MARL) for its decentralized decision-making and adaptability to dynamic environments Gao et al. [2024], Li et al. [2025].

The effectiveness of some prior MARL-based CI methods requires per-device policies, which forces costly retraining once environments change and limits scalability Xiao et al. [2023, 2022]. Other approaches rely on detailed per-neighbor telemetry for coordination, which conflicts with privacy constraints Jiang et al. [2020]. In an M2M CI system, homogeneity of devices is natural, since those who subscribe to the same services generally share the same inference model and are only allowed to collaborate with nearby edges. This motivates us to adopt Mean-Field MARL (MFRL) Yang et al.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).



(a) High-variance proximity-based statistics (b) Irrelevant devices

Figure 1: Illustration of mobility challenges of mean-field statistics existing in an M2M CI system.

[2018], which conditions decisions on a mean-field statistic rather than explicit neighbor information, retaining scalability and privacy guarantee.

The effectiveness of MFRL relies on the premise that mean-field statistics are stable and informative, i.e., that the set of neighbors for each device remains consistent over time Sriram Ganapathi et al. [2022]. However, due to device mobility, each device's neighborhood and connections change frequently, as shown in Fig. 1a. This renders spatial-proximity-based statistics high-variance and noisy, undermining their effectiveness. Expanding the neighborhood helps reduce variance but dilutes the signal with non-influential devices that do not tend to compete for resources with the target device, as shown in Fig. 1b. Including these irrelevant devices in the observations during training may lead to slow convergence.

In this paper, we propose **Fairness-Oriented Collaborative Inference (FOCI)**, an efficient MARL method tailored for M2M CI that achieves fair QoS across devices with minimal coordination overhead. It combines two key components: (i) **Oracle-shaping fairness rewards (OFR):** each edge feeds back a max-min reward as an oracle fairness signal to guide device learning toward globally fair outcomes; and (ii) **Competitor-aware observation augmentation (COA):** each device augments its observations with the behavior distribution of potential resource competitors, yielding stable mean-field statistics. We prove that FOCI approximates equilibrium within a bounded error. Our experimental results validate its effectiveness with realistic mobility traces, showing substantial gains in worst-case latency, energy consumption, and QoS over strong baselines. Our contributions are summarized as follows:

- We propose an RL-based method, FOCI, and provide a convergence analysis of its bounded error.
- We apply an oracle-shaped reward to guide training towards a fairness target in the M2M CI system.
- We augment observation with potential competitor behaviors to address the mobility challenge.
- We validate it with real-world mobility traces to prove its effectiveness in the M2M CI system.

2 Preliminary

Mean-field approximation in MFRL. The interaction of agents in MARL is naturally modeled as a stochastic game Littman [1994] with joint policy $\boldsymbol{\pi} = \{\pi^1, \ldots, \pi^M\}$. A Nash equilibrium $\boldsymbol{\pi}^*$ satisfies, for every agent m, state s, and feasible unilateral deviation π^m, that the following inequality

$$v^m(s; \pi^m, \boldsymbol{\pi}_*^{-m}) \leq v^m(s; \pi_*^m, \boldsymbol{\pi}_*^{-m}), \quad \forall s \in \mathcal{S},\ \forall \pi^m \tag{1}$$

holds, where $\boldsymbol{\pi}_*^{-m}$ denotes the policy profile of all agents except agent m; $\mathcal{S}$ is the state space; and v^m represents the value of the current state s. Nash Q-learning Hu and Wellman [2003] solves this by alternately calculating the stage-game NE and updating Q-values at each stage. However, this approach yields unacceptable combinatorial complexity in large-scale M2M CI due to the need to evaluate all possible joint actions. In contrast, MFRL significantly simplifies the interactions by conditioning each agent's update on compact mean-field statistics rather than on per-neighbor telemetry exchange. The rationale behind this is that, under homogeneity and locality assumptions, average behavior patterns matter more than any individual agent's behavior. Specifically, the action-value for agent m, $Q^m(s, a^m, \mathbf{a}^{-m})$, is approximated by $Q^m(s, a^m, \mathbf{a}^{-m}) \approx Q^m\big(s, a^m, \bar{\mathbf{a}}^m\big)$, via first-order mean-field (Taylor) expansion, where $\bar{\mathbf{a}}^m \in \Delta^{|\mathcal{A}|-1}$ is the empirical action distribution of agents in a neighborhood $\mathcal{N}^m$ (e.g., $\bar{\mathbf{a}}^m := \mathbb{E}_{n \sim d(\mathcal{N}^m)}[\mathbf{e}(a^n)]$, with $\mathbf{e}(\cdot)$ one-hot over $|\mathcal{A}|$ actions).

This compacts the exponential dependence on the number of agents to a size-invariant representation, enabling efficient Q-updates.

3 M2M CI system model and problem formulation

System model. An M2M CI system is deployed in a densely populated area, as shown in Fig. 2. A cluster of edges $\mathcal{N}$ collaborates with a set of battery-powered devices $\mathcal{M}$ roaming in this area to provide inference services. The inference model on each device is abstracted as a K-layer structure. Devices autonomously make CI decisions $A^m = \{u^m, k^m\}$, including edge selection $u^m \in [0, N]$ and model partition $k^m \in [0, K]$, based on their local observations o^m, such as hardware status and channel conditions. Specifically, $\{u^m = 0, k^m = K\}$ indicates fully local inference. Each edge has L allocatable Streaming Multiprocessors (SMs). Once a connection change is detected, edge n updates its allocation pattern $\mathcal{L}_n = \{l_n^m\}_{\mathcal{M}_n}$ for the currently connected device set $\mathcal{M}_n$. Edges are assumed to be selfless and always fully utilize all available resources.

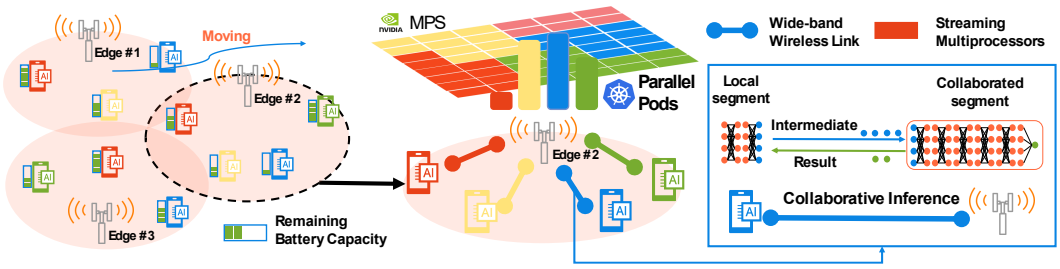


Figure 2: Diagram of M2M collaborative inference system.

Performance metrics. The common QoS criteria include latency D^m, energy consumption E^m, and connection switch cost C^m when a device changes its collaborating edge due to mobility. Formally, the QoS for each device is given by

$$R^m(A^m, L_n^m) = -\big(\alpha E^m(A^m) + \beta D^m(A^m, L_n^m) + \gamma C^m\big),$$

where α, β, and γ are weighting coefficients. The QoS is affected by the joint CI decisions across all devices. In particular, the partition decision k^m determines the model segments on the local and edge sides, thereby influencing the corresponding latency and on-device energy consumption. The connection decision u^m influences the transmission latency and energy. Note that the inference latency on the edge is also affected by u^m, since the connected set $\mathcal{M}_n$ competes for resources on edge n.

Learning objective. We seek to achieve QoS fairness among devices collaborating with edges. We formulate the following max–min optimization problem:

$$\textbf{(P-Orig)} \quad \max_{\mathcal{U},\mathcal{K},\mathcal{L}} \big(\min_m \{R^m(A^m, l_n^m)\}_{\mathcal{M}}\big) \quad \text{s.t} \quad D^m \le D^{st} \tag{2}$$

where $\{\mathcal{U}, \mathcal{K}\}$ are the joint actions of all devices, $\mathcal{L}$ is the joint resource allocation of all edges, and D^{st} is a latency constraint that ensures valid actions. Since edges make passive decisions based on existing device actions, the edge behavior can be defined as a mapping $\mathcal{F} : (\mathcal{U}, \mathcal{K}) \mapsto \mathcal{L}$. The original problem **(P-Orig)** is then reformulated as an MARL problem under device mobility and limited observability. Specifically, we aim to obtain the optimal device policies for max–min expected QoS:

$$\textbf{(P-MARL)} \quad \max_{\boldsymbol{\pi}} \min_{m \in \mathcal{M}} \mathbb{E}_{\tau \sim P^{\boldsymbol{\pi},\mathcal{F}}} \left[\sum_{t=0}^{\infty} \gamma^t R^m(A^m, \mathcal{F}) \right] \tag{3}$$

where $\gamma \in (0, 1)$ is the discount factor and $P^{\boldsymbol{\pi},\mathcal{F}}$ is the trajectory law induced by the device policies $\boldsymbol{\pi}$ and the environment with edge behavior $\mathcal{F}$. Our solution to **(P-MARL)** builds on MFRL, which offers scalability and privacy preservation by conditioning device policies on aggregated mean-field statistics rather than per-neighbor telemetry. However, in M2M CI systems, this naive mean-field assumption often fails: device mobility introduces high variance into local means, while relevance sparsity dilutes global means, both of which can mislead learning. Furthermore, MFRL does not explicitly align individual device updates with the system's global max–min QoS objective.

4 Method

We propose Fairness-Oriented Collaborative Inference (FOCI) to achieve fairness in MARL with performance guarantees. The fairness issue in **(P-MARL)** is first addressed by enabling an oracle-shaping fairness rewards to guide the training. We then handle the mobility challenge in MFRL through competition-aware observation augmentation. The convergence analysis of the bound gap equilibrium for FOCI is presented in the end.

4.1 Oracle-shaping fairness rewards

Oracle-shaping with a different runtime allocator. The difficulty of solving **P-MARL** arises from the coupling between proactive devices and passive edge allocators. To align learning with max–min fairness, we introduce a per-edge fairness oracle $\{\mathcal{F}_n^*\}$ used only for shaping the reward for each device. The $\mathcal{F}_n^*$ is defined as the solution to the following max-min reward problem

$$\textbf{(P-EDGE)} \quad R_n^* := \max_{\mathcal{L}_n} \min_{m \in \mathcal{M}_n} R^m(A^m, l_n^m) \tag{4}$$

where R_n^* is the oracle-shaping reward. Each device independently solves the above problem with joint decisions from connected devices $\{u^m, k^m\}_{m \in \mathcal{M}_\setminus}$. The oracular reward is fed back to the corresponding device set $\mathcal{M}_n$ for updating their policy in each step.

To obtain oracle rewards efficiently in highly dynamic environments, we propose a polynomial-time min-risk search algorithm. Specifically, the original problem is first relaxed to a minimax linear-fractional programming problem, and then solved by the Interval Partition Linearization-based Local Search (IPLS) algorithm Zhang et al. [2023a] to obtain the optimal continuous solution $\bar{\mathcal{L}}_n^*$. After that, we apply a min-risk backtracking step to round it to an integer solution, removing any over-allocation while minimizing the drop in the worst-device reward relative to $\bar{\mathcal{L}}_n^*$. This pipeline yields a discrete, fairness-aligned allocation suitable for training. See Appendix A for the algorithm details.

Complexity analysis for fairness-oracular rewards. In the worst case, obtaining $\mathcal{L}_n^*$ requires traversing all intervals with the dichotomy method, and the size of redundant resources in the proposed local search algorithm is equivalent to $|\mathcal{M}_n|$. Thus, it needs to update all $|\mathcal{M}_n|$ variables $\lceil log_2 \frac{2(ln(\bar{R}^u) - ln(\bar{R}^l))}{\varepsilon} \rceil$ times, where $\bar{R}^u$ and $\bar{R}^l$ are the upper and lower bounds of worst-off reward; ε is the error bound in IPLS algorithm. The min-risk search must calculate future returns and backtrack at most $|\mathcal{M}_n|$ times, indicating the complexity as $\mathcal{O}(|\mathcal{M}_n| log_2(\varepsilon^{-1}))$.

4.2 Competition-aware observation augmentation

Perception domain of potential competitors. To ensure autonomy and privacy, each device is prohibited from accessing the geographic locations or exact actions of others. As a result, devices cannot directly identify their potential competitors. To address this issue, we introduce the concept of a *perception domain*. A perception radius r_p^m is defined, which represents the maximum transmission radius that satisfies a specific latency requirement. Due to the symmetry of edges, devices located within a perception radius of the target device are considered potential competitors, while those beyond this range are excluded because they cannot meet the latency requirements.

Observation augmentation with anonymous mean-field statistics. When an edge allocates resources for connected devices, it will gather the exact decisions of each device and maintain updated action records $F_t^n \in \mathbb{R}^{K \times 1}$. The record is an aggregated action distribution of all connected devices. When device m makes a decision, it acquires all historical action records from nearby edges within the perception domain r_p^m as

$$F_{t-1}^m = \{\mathbb{I}(d^{m,n} \leq r_p^m) F_{t-1}^n\}_{\mathcal{N}} \tag{5}$$

where $F_{t-1}^m \in \mathbb{R}^{K \times N}$; $\mathbb{I}(\cdot)$ is an indicator function that determines the nearby edges. This allows each device to obtain all relevant information about potential competitors without exposing detailed status information or exact actions. After augmenting the local observation with the action distribution of potential competitors F_{t-1}^m, the value function with respect to the Q-value for device m becomes

$$v_t^m(O_{t+1}^m) = \sum\nolimits_{A_t^m} \pi_t^m(A_t^m | O_t^m, F_{t-1}^m) Q^m(O_t^m, A_t^m, F_{t-1}^m) \tag{6}$$

and the Q-function of device m in FOCI is updated as

$$Q_{t+1}^m = (1 - \eta) Q_t^m + \eta(R_t^m + \gamma v(O_{t+1}^m)) \tag{7}$$

This competitor-aware mean-field statistic achieves a better bias–variance trade-off by focusing on potential competitors, leading to steadier value updates. The rationale is that connection behavior is more stable than physical locality: devices may move frequently, but they tend to maintain similar edge preferences over time. Therefore, aggregating only the records from nearby edges within the perception domain yields a more stable mean-field statistic for training. In addition, compared with an expanded proximity-based statistic, the competitor-aware mean-field statistic filters out interactions from irrelevant devices, thereby preventing their dilution of the training signal.

Convergence Analysis. The procedure of augmenting with aggregated action records is equivalent to sampling from the global actions with MFRL. To demonstrate the convergence of the proposed augmentation method, we first prove the bounded error of Q-value between the sampled action and global actions (Theorem 1). Furthermore, we show that the Q-value remains at a bounded distance from the Nash Equilibrium Q-value (Theorem 2).

Assumption 1 (*L-continuous Q function w.r.t action distribution*) The Q-function in Eq. (7) is Lipschitz continuous w.r.t action distribution with a constant L, i.e.,

$$|Q(O_t, A_t, F^1_{t-1}) - Q(O_t, A_t, F^2_{t-1})| \leq L|F^1_{t-1} - F^2_{t-1}| \tag{8}$$

Theorem 1. *(Bounded gap to global distribution) Under the assumption that the Q-function is L-Lipschitz continuous with respect to the action distribution, the deviation in the Q-value between the sampling distribution F^m_{t-1} and the true global action distribution of all agents F_{t-1} is bounded as follows with probability $\sigma^{\phi-1}$:*

$$|Q(O_t, A_t, F^m_{t-1}) - Q(O_t, A_t, F_{t-1})| \leq L\frac{\phi - \lambda + 1}{2\lambda} \ln \frac{2}{\sigma} \tag{9}$$

where $\phi = (NK + 1) >> 1$ represents the dimension of entire action space; λ is the expected value of the non-zero dimension of F^m_{t-1}; and σ denotes the upper bound of on the deviation between the sampled action distribution and the global action distribution. See the proof in Appendix B.1.

Theorem 2. *(Bounded gap to Nash Equilibrium) Under the assumptions (Appendix B. 2.), if we employ an update for the Q-functions as elucidated in Eq. (7), the Q-value in FOCI satisfies*

$$|Q(O_t, a_t, F^m_{t-1}) - Q^*(O_t, \mathbf{a}_t)| \leq 2\delta \quad w.p.\sigma^{\phi-1}, \tag{10}$$

where δ is the deviation bound in value function as $|v(O_{t+1}|F^m_{t-1}) - v(O_{t+1}|F_{t-1})| = L\frac{\phi-\lambda+1}{2\lambda} \ln \frac{2}{\sigma} + \kappa\sqrt{2}$ and Q^* is the Nash equilibrium Q-value. See the proof in Appendix B.2.

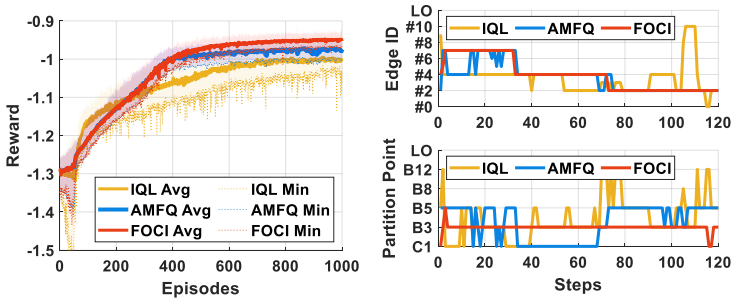
5 Experiments

Experiment settings. We simulate an M2M CI with 240 devices and 11 edges. To simulate device mobility, we utilize a real-world taxi trajectory dataset from Porto, Portugal Kaggle [2017], focusing on a $2.5\ km^2$ downtown area. We take the MobileNetV3-large Howard et al. [2019] as the backbone network for the inference model. The latency constraint is set to $1/24$s. The data for simulating hardware status and scaling functions of devices is collected from the run-time status of Nvidia Jetson TX2 NVIDIA [2024]. The number of allocatable SMs on each edge is set to 2×128. A simple 3-layer MLP architecture is applied on each device for RL training. More details about the experimental settings are given in Appendix. C.

Overall performance. We compare FOCI with : 1) ALP: all local processing ; 2) Neurosurgeon (NSG) Kang et al. [2017]: individual-optimal decision; 3) IQL Liang et al. [2019]: only local observation for RL training; and 4) AMFQ Zhang et al. [2023b]: MFRL-based resource allocation scheme. The results are shown in Table. 1.

Metrics/method	ALP	NSG	IQL	AMFQ	FOCI
Avg/Max latency (ms)	34.20/39.51	32.77/62.72	27.83/30.08	27.33/27.96	**26.82/27.20**
Avg/Max energy (mJ)	60.24/71.44	**22.16**/62.55	23.95/38.48	25.87/35.16	24.50/**33.50**
Avg switch cost	-	2.88	3.06	3.35	**1.52**
Avg/Max Reward	1.248/1.417	1.157/2.193	0.988/1.065	0.971/0.997	**0.953/0.973**

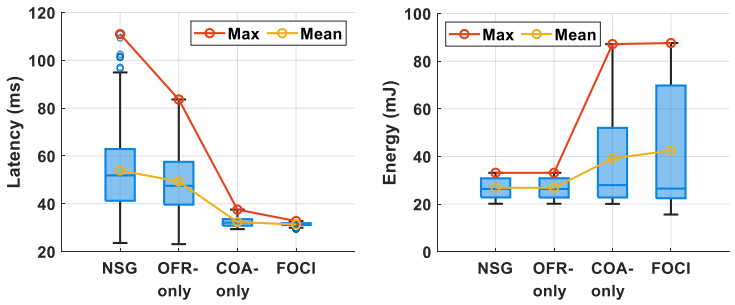
Table 1: Overall comparison of different CI methods.



(a) Training result (b) Switch behavior

Figure 3: Performance comparison in ablation study.

FOCI demonstrates the best tail performance while keeping strong averages, which matches its max–min design. On latency, both non-RL-based baselines fail on the worst-off QoS control, especially the local-optimal solution NSG. In comparison, FOCI attains the lowest average and a remarkably tight worst case in latency, e.g, 26.82 ms and 27.20 ms, and in QoS reward (lower is better), e.g., 0.953 and 0.973. Energy shows the same pattern that FOCI yields the lowest max energy at 33.50 J, cutting the worst case by 46% vs. $62.55J$ of NSG. As for the decision stability, we visualize the switch behavior of different RL-based methods in Fig. 3. FOCI achieves the most consistent decisions over time, leading to a 50% lower switch cost compared with other RL-based methods. Consequently, the reward is the best for FOCI in both average and max, reflecting that oracle-shaping fairness feedback suppresses outliers while the competitor-aware mean field augments the stationarity of the decisions.



(a) Latency per frame (b) Energy per frame

Figure 4: Performance comparison in ablation study.

Ablation study. We compare four variants: Neither (NSG), OFR-only, COA-only, and FOCI to quantify the contribution of the two components in FOCI. As shown in Fig. 4a, the max-latency drops monotonically from NSG, OFR-only, COA-only to FOCI, showing that the oracular reward alone already suppresses outliers efficiently, and adding COA yields the lowest tail and the smallest variance. Mean latency follows the same trend with incremental gains, indicating smoother partition choices. In Fig. 4b, energy shows a similar pattern: FOCI reduces both mean and max, while removing either component raises dispersion; COA-only lowers the average energy because its augmentation distribution is smoother and more predictive than that of the noisy proximity-based neighbor. In comparison, OFR-only lowers the tail with the explicitly equalized contention.

6 Conclusion

We modeled M2M CI as MARL and proposed FOCI, which aligns local learning with global fairness while avoiding peer telemetry. The method couples oracle-shaping rewards from the edge's max–min response with a competition-aware mean field-based observation augmentation, yielding a stable policy for achieving max-min QoS. We give a theoretical guarantee on its bounded convergence from the equilibrium. Empirically, FOCI shows fairer QoS with tighter latency and energy tails and more stable behavior patterns compared with multiple M2M CI baselines. This suggests a practical path to achieve scalable M2M CI in NextG wireless networks with MARL.

References

Zhang Bo, Gao Yuelin, Liu Xia, and Huang Xiaoli. Interval division and linearization algorithm for minimax linear fractional program. *Numerical Algorithms*, 95:839–858, 2023.

Sriram Ganapathi Subramanian, Matthew E. Taylor, Mark Crowley, and Pascal Poupart. Partially observable mean field reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, page 537–545, 2021.

Guanyu Gao, Yuqi Dong, Ran Wang, and Xin Zhou. Edgevision: Towards collaborative video analytics on distributed edges for performance maximization. *IEEE Transactions on Multimedia*, 26:9083–9094, 2024.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019. URL `https://arxiv.org/abs/1905.02244`.

Junling Hu and Michael P. Wellman. Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069, dec 2003. ISSN 1532-4435.

Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph convolutional reinforcement learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HkxdQkSYDB`.

Kaggle. Taxi trajectory data, 2017. ECML/PKDD 15: Taxi Trip Time Prediction (II) Competition, 2015.

Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '17, page 615–629, Xi'an, China, 2017. ACM.

Muhammad Asif Khan, Ridha Hamila, Aiman Erbad, and Moncef Gabbouj. Distributed inference in resource-constrained iot for real-time video surveillance. *IEEE Systems Journal*, 17(1):1512–1523, 2023.

En Li, Liekang Zeng, Zhi Zhou, and Xu Chen. Edge ai: On-demand accelerating deep neural network inference via edge computing. *IEEE Transactions on Wireless Communications*, 19(1):447–457, 2020.

Jing Li, Weifa Liang, Yuchen Li, Zichuan Xu, Xiaohua Jia, and Song Guo. Throughput maximization of delay-aware dnn inference in edge computing by exploring dnn model partitioning and inference parallelism. *IEEE Transactions on Mobile Computing*, 22(5):3017–3030, 2023.

Xulong Li, Wei Huangfu, Xinyi Xu, Jiahao Huo, and Keping Long. Attention-driven marl for aoi minimization in uav-assisted intelligent transport systems. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–15, 2025.

Le Liang, Hao Ye, and Geoffrey Ye Li. Spectrum sharing in vehicular networks based on multi-agent reinforcement learning. *IEEE Journal on Selected Areas in Communications*, 37(10):2282–2292, 2019.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 157–163. Morgan Kaufmann, San Francisco (CA), 1994.

Zhicheng Liu, Jinduo Song, Chao Qiu, Xiaofei Wang, Xu Chen, Qiang He, and Hao Sheng. Hastening stream offloading of inference via multi-exit dnns in mobile edge computing. *IEEE Transactions on Mobile Computing*, 23(1):535–548, 2024.

Zhiyan Liu, Qiao Lan, and Kaibin Huang. Resource allocation for multiuser edge inference with batching and early exiting. *IEEE Journal on Selected Areas in Communications*, 41(4):1186–1200, 2023.

Thaha Mohammed, Carlee Joe-Wong, Rohit Babbar, and Mario Di Francesco. Distributed inference acceleration with adaptive dnn partitioning and offloading. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pages 854–863, Virtual Conference, 2020. IEEE.

NVIDIA. Jetson tx2 developer kit. `https://developer.nvidia.com/embedded/jetson-tx2-developer-kit`, 2024.

Singh Satinder, Jaakkola Tommi, L. Littman Michael, and Szepesvári Csaba. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38:287–308, 2000.

R. J. Serfling. Probability Inequalities for the Sum in Sampling without Replacement. *The Annals of Statistics*, 2(1):39 – 48, 1974.

Subramanian Sriram Ganapathi, Taylor Matthew E., Crowley Mark, and Poupart Pascal. Decentralized mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 9439-9447. AAAI, 2022.

Xin Tang, Xu Chen, Liekang Zeng, Shuai Yu, and Lin Chen. Joint multiuser dnn partitioning and computational resource allocation for collaborative edge intelligence. *IEEE Internet of Things Journal*, 8(12):9511–9522, 2021.

Li Wang, Xin Wu, Yi Zhang, Xinyun Zhang, Lianming Xu, Zhihua Wu, and Aiguo Fei. Deepadainnet: Deep adaptive device-edge collaborative inference for augmented reality. *IEEE Journal of Selected Topics in Signal Processing*, 17(5):1052–1063, 2023.

Yilin Xiao, Kunpeng Wan, Liang Xiao, and Helin Yang. Energy-efficient collaborative inference in mec: A multi-agent reinforcement learning based approach. In *2022 8th International Conference on Big Data Computing and Communications (BigCom)*, pages 407–412, 2022.

Yilin Xiao, Liang Xiao, Kunpeng Wan, Helin Yang, Yi Zhang, Yi Wu, and Yanyong Zhang. Reinforcement learning based energy-efficient collaborative inference for mobile edge computing. *IEEE Transactions on Communications*, 71(2):864–876, 2023.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5567–5576, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Bo Zhang, Yuelin Gao, Xia Liu, and Xiaoli Huang. Interval division and linearization algorithm for minimax linear fractional program. *Numerical Algorithms*, 95:1–20, 07 2023a. doi: 10.1007/s11075-023-01591-0.

Hengxi Zhang, Chengyue Lu, Huaze Tang, Xiaoli Wei, Le Liang, Ling Cheng, Wenbo Ding, and Zhu Han. Mean-field-aided multiagent reinforcement learning for resource allocation in vehicular networks. *IEEE Internet of Things Journal*, 10(3):2667–2679, 2023b.

Appendix

A Min-risk searching algorithm for oracle-shaping fairness reward

We first rewrite the **(P-EDGE)** problem in the form as:

$$\textbf{P-EDGE-2:} \quad R_n^* := \max_{\mathcal{L}_n} \min_{m \in \mathcal{M}_n} (-R_c^m - \frac{R_f^m}{L_n^m}) \tag{11}$$

where $R_c^m = \alpha E^m + \beta(D^{m,lo} + D^{m,tr})$ refers to the constant part of the reward R^m, and $R_f^m = \beta L \sum_{k=k^m+1}^{K} D_k^{re} + \gamma C^m$ is the molecular part. **P-EDGE-2** is NP-hard due to the fractional form of variables l_n^m in each item. Thus, we first relax original integer programming to the polynomial solvable continuous problem. After obtaining the optimal continuous solutions via IPLS algorithm Zhang et al. [2023a], we backtrack to the integer results with a heuristic strategy. The detailis given in Algorithm. 1.

Algorithm 1 Min-risk searching algorithm for discrete resource allocation

Input: A set of $\{R_c^m, R_f^m\} > 0, \forall m \in \mathcal{M}_n$. The available SM L. Tolerance ε.

1: Relax the original problem to an Minimax Linear-Fractional Programming problem
2: Set the list of upper and lower bound candidates as $\{U^m\}_{\mathcal{M}_n} = \{R^m(1)\}_{\mathcal{M}_n}$, $\{L^m\}_{\mathcal{M}_n} = \{R^m(L - |\mathcal{M}_n|)\}_{\mathcal{M}_n}$,
3: Obtain upper bounds $\bar{R}^u = \max\{U^m\}_{\mathcal{M}_n}$ and lower bounds $\bar{R}^l = \max\{L^m\}_{\mathcal{M}_n}$,
4: Create steps $\varpi = \lfloor \log_{(1+\varepsilon)}(\bar{R}^u/\bar{R}^l) \rfloor$ and corresponding polynomial interval set as $\mathcal{P}^\varepsilon = \{\bar{R}^l, \bar{R}^l(1+\varepsilon), \ldots, \bar{R}^l(1+\varepsilon)^\varpi\} \triangleq \{p_1, p_2, \ldots, p_{\varpi+1}\}$,
5: Initialize iteration counter $k = 0$ and optimum $\bar{R}^*_{\mathcal{M}_n} = +\infty$.
6: **while** $P^\varepsilon \neq \emptyset$ **do**
7: Divider $D = |\mathcal{P}^\varepsilon|$, candidate optimal interval $\bar{p} = p_D$.
8: **if** $\sum_{\mathcal{M}_n} R_f^m/(\bar{p} - R_c^m) \leq L$ **then**
9: Update set $\mathcal{P}^\varepsilon = \{p_1, \ldots, p_D\}$,
10: Obtain the current optimal QoS $\bar{R}^*_{\mathcal{M}_n} = \min\{\bar{p}, \bar{R}^*_{\mathcal{M}_n}\}$, and solutions $\bar{\boldsymbol{l}}_n^* = \{\max\{R_f^m/(\bar{p} - R_c^m), 1\}\}_{\mathcal{M}_n}$.
11: **else**
12: Update set $P^\varepsilon = \{p_{D+1}, \ldots, p_{\varpi+1}\}$, reset $p_1 = p_{D+1}$
13: **end if**
14: **end while**
15: Given $\bar{\boldsymbol{l}}_n^*$, calculate $\{\bar{R}_n^{m*}\}_{\mathcal{M}_n}$, and $\bar{R}^*_{\mathcal{M}_n} - p^* \leq \varepsilon$ holds,
16: Initialize solution $\boldsymbol{l}_n^* = \lceil \bar{\boldsymbol{l}}_n^* \rceil$, calculate diff $\vartheta = \sum_{\mathcal{M}_n} l_n^{m*} - L$.
17: **while** $\vartheta \neq 0$ **do**
18: Exclude all m that satisfy $l_n^{m*} == 1$ from $\mathcal{M}_n$,
19: Calculate future value $\bar{R}_n^* = \{\bar{R}_n^{m*}(l_n^{m*} - 1)\}_{\mathcal{M}_n}$,
20: Locate the smallest index $\bar{m} = \text{argmin}\{\bar{R}_n^{m*} - R_n^{m*}\}_{\mathcal{M}_n}$,
21: Update $l_n^{\bar{m}*} = l_n^{\bar{m}*} - 1$, and $\vartheta = \vartheta - 1$.
22: **end while**

Return: Optimal SMs allocation $\boldsymbol{l}_n^*$

Details of IPLS algorithm. The main idea is that **P-EDGE-2** shares the same optimal global solution with the equivalent problem $\min_m\{\bar{R}_{\mathcal{M}_n} - \max_{\{\bar{l}_n^m\}}(R_c^m + R_f^m/\bar{l}_n^m)\}, \bar{R} \in [\bar{R}^l, \bar{R}^u]$, where $\bar{R}^u, \bar{R}^l$ is the upper and lower bound of worst-off QoS (Step 2 and 3). Given a tolerant error ε to the optimal global QoS, the solution space $[\bar{R}^l, \bar{R}^u]$ is divided into a set of polynomially countable small intervals $\mathcal{P}^\varepsilon$ (Step 4). The corresponding number of intervals is given as $\varpi = \lfloor \log_{(1+\varepsilon)}(\bar{R}^u/\bar{R}^l) \rfloor$. Compared to the set of constant intervals, the polynomial set dramatically reduces the searching space. We use the dichotomy to locate the interval where the optimal solution resides. According to the Theorem 3 in Bo et al. [2023], if the Constraint $\sum_{\mathcal{M}_n} R_f^m/(\bar{p} - R_c^m) \leq G l_n$ is satisfied by current candidate $\bar{p} \in \mathcal{P}^\varepsilon$ (Step 8), $\max_m\{R_c^m + R_f^m/l_n^m\}_{\mathcal{M}_n} \leq (1+\varepsilon)\max_m\{R_c^m + R_f^m/l_n^{m*}\}_{\mathcal{M}_n}$ holds, and the $\bar{R}^*_{\mathcal{M}_n}$ must be on the left side of the current interval. Subsequently, we update the polynomial set (Step 9) and the optimal variables (Step 10). By iteratively applying this process until the polynomial set is empty, we identify a $\bar{R}^*_{\mathcal{M}_n}$ that satisfies an ε-approximate global optimal solution.

B Proof of theorem related to competition-aware observation augmentation

B.1 Theorem 1: Bounded gap to global distribution

Proof. According to the assumption in Eq. (8), we first derive the bound of $|a_t^1 - a_t^2|$ where $a_t^1 = F_{t-1}^m$ and $a_t^2 = F_{t-1}$. Denote the perceptional action dimension of each agent m as ϕ_m, we have an expected dimensions over all agents as $\lambda = \mathbb{E}_{m\sim M}\phi_m$. We then denote the perception procedure as a set of random variables (RVs) $\{X_i \in [0,1], \forall i \in \lambda\}$, each of which indicates the sampling on one action and repeats λ times. We normalize F_{t-1}^m for each agent m as a probability distribution in the following analysis.

After that, the perception procedure of agent m is regarded as sampling from the global action F_{t-1} without replacement, so that X_i is no longer independent with each other. According to Serfling inequality (given in Corollary 1.1 in Serfling [1974]), we then provide an upper bound on the probability that the sum of bounded dependent RVs deviates from its expected value as

$$P(|S_\lambda - \lambda\mu| \geq \lambda p) \leq 2\exp(\frac{-2\lambda p^2}{(1-\frac{\lambda-1}{\phi})(b-a)}) \tag{12}$$

where $p > 0$ refers to the deviation; $S_\lambda = \sum_{i\in\lambda} X_i$ describes the sampling procedure; given a finite list of all values of RV X_i as x_i, $\mu = \sum_{i=1}^{\phi} x_i$ represents the global action distribution; and b and a are the upper and lower bound of variable X_i, respectively. If the perception range is large enough to cover the whole mobility area, the sum of distribution X_i approaches the true global action. Thus, let $S_\lambda/\lambda = F_{t-1}^m$, $\mu = F_{t-1}$, and $b = 1, a = 0$, we have

$$P(|F_{t-1}^m - F_{t-1}| \geq p) \leq 2\exp(\frac{-2\lambda\phi p^2}{\phi - \lambda + 1}) \tag{13}$$

Let the right hand of the above inequation as σ, we have $p = \sqrt{\frac{\phi-\lambda+1}{2\lambda\phi}\ln\frac{2}{\sigma}}$. We then obtain the difference between the mean action of COA and global mean action as

$$|F_{t-1}^m - F_{t-1}| \leq \sqrt{\frac{\phi-\lambda+1}{2\lambda\phi}\ln\frac{2}{\sigma}} \quad w.p.\sigma. \tag{14}$$

As the sum of individual components is 1, the last random variable becomes deterministic once we fix the first $\phi - 1$ RVs. Thus, with the above bound holds $w.p.\sigma$, we have the following bound holds $w.p.(\sigma)^{\phi-1}$ under Assumption 1 as

$$|Q(O_t, A_t, F_{t-1}^m) - Q(O_t, A_t, F_{t-1})| \leq L|F_{t-1}^m - F_{t-1}| \tag{15}$$

By replacing $|F_{t-1}^m - F_{t-1}|$ with $\phi(\sqrt{\frac{\phi-\lambda+1}{2\lambda\phi}\ln\frac{2}{\sigma}})^2$, we have proved the Theorem. 1. □

B.2 Theorem 2: Bounded gap to Nash Equilibrium

Given the conclusion in Theorem. 1, we will prove the convergence of FOCI under the following 4 common assumptions in MARL.

Assumption 2 (*Bounded reward*). Each action-value pair is visited infinitely with bounded reward.

Assumption 3 (*Rational agent*). Agent's policy is Greedy in the Limit with Infinite Exploration (GLIE).

Assumption 4 (*Equilibrium Optimality*). The Nash equilibrium is considered a global optimum or a saddle point in every stage game of the stochastic game.

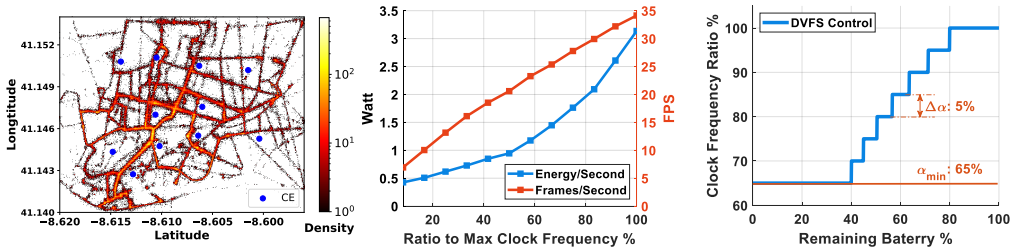
Assumption 5 (κ-*continuous Q function w.r.t actions*) The Q-function is Lipschitz continuous w.r.t the action with constant κ, which is $|Q(O_t, A_t^1, F_{t-1}^m) - Q(O_t, A_t^2, F_{t-1}^m)| \leq \kappa|A_t^1 - A_t^2|$

Proof. First, our learning model satisfies all above assumptions: the reward in FOCI is bounded; the annealing ϵ-greedy policy applied in our model is GLIE Satinder et al. [2000]; the optimality of Nash equilibrium and Lipschitz continuity are strong assumptions in most related work about MARL. After that, by replacing Eq. (21) in the proof in Subramanian et al. Ganapathi Subramanian et al. [2021] with the conclusion from Theorem. 1, we have proved the Theorem. 2. □

B.3 Detail experiment settings

The mobility of devices is randomly sampled from trajectories in the dataset Kaggle [2017]. We visualize the original trajectory set as a heatmap, as shown in Fig. 5a. The edge placement is also determined by the heatmap, following that higher-density regions host more densely deployed edges.

For the device prototype, we use the Jetson TX2 and measure its energy consumption and latency under different clock frequencies, as shown in Fig. 5b. Although the TX2 is not battery-powered, we simulate a mobile device by assigning an initial energy level between 3 Wh and 12 Wh, a typical range for a smartphone-class device (e.g., an iPhone). A DVFS controller, illustrated in Fig. 5c, is designed based on the TX2 power consumption model. Note that although we use the term "battery level" for ease of exposition, the actual factor influencing performance is the ratio of the current clock frequency to the maximum clock frequency. The minimum allowable clock-frequency ratio is set to $\alpha_{\min} = 65\%$, because lower frequencies would fail to meet the latency requirement. It is worth noting that our proposed method is adaptable to any device prototype, including variations in battery capacity and DVFS configurations, without significantly affecting the main experimental results.



(a) Edge placement scheme (b) Scaling function (c) DVFS controller on device

Figure 5: Detailed experiment settings

As for the training parameters, we apply a dueling DDQN with the same 3-layer MLP architectures for each device. The learning rate and discount factor are set as $\eta = 0.0001$ and $\gamma = 0.7$. The ϵ in greedy exploration exponentially decays from 1 by 0.99 every episode. For simplicity, we set all weights in the QoS reward to 1. Each episode contains 1 hour of the inference services.