Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Human-Centric Cooperative MARL for Reliable EV Charging Coordination in Smart Cities: A Controlled Multi-Seed Comparison of Centralised and Decentralised Training

Nour-Eddine Moumni

noureddine_moumni@um5.ac.ma

Laboratory of Research in Computing and Telecommunications (LRIT) https://orcid.org/0009-0008-5803-0604

Rachid Alaoui

Laboratory of Research in Computer Science and Telecommunications (LRIT), Mohammed V University, Rabat, Morocco Higher School of Technology of Salé (EST Salé), Mohammed V University, Salé, Morocco https://orcid.org/0000-0001-9948-329X

Driss Kiouach

L2MASI Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco https://orcid.org/0000-0002-1497-3028

Research Article

Keywords: Electric Vehicle Charging, Multi-Agent Reinforcement Learning, Deep Reinforcement Learning, QMIX, CTDE, DTDE, SUMO

Posted Date: October 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10139360/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Human-Centric Cooperative MARL for Reliable EV Charging Coordination in Smart Cities: A Controlled Multi-Seed Comparison of Centralised and Decentralised Training

Nour-Eddine Moumni[1*], Rachid Alaoui[1,2] and Driss Kiouach[3]

[1*]Laboratory of Research in Computing and Telecommunications (LRIT), Mohammed V University in Rabat, Rabat, B.P.1014, Morocco.
[2]Higher School of Technology of Salé (EST Salé), Mohammed V University, Salé, Morocco.
[3]L2MASI Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco.

*Corresponding author(s). E-mail(s): noureddine_moumni@um5.ac.ma;
Contributing authors: rachid.alaoui@est.um5.ac.ma; kiouach@gmail.com;

Abstract

Efficient allocation of electric vehicles (EVs) to charging stations in dense urban environments requires decision-making under partial observability and time-varying traffic demand. We present a controlled empirical comparison of two multi-agent reinforcement learning paradigms for cooperative EV charging assignment in a SUMO-based urban network: Independent Q-Learning (IQL; DTDE) and QMIX (CTDE). Across 7 random seeds and 200 training episodes per seed (2,800 episodes in total), QMIX attains a mean training acceptance rate of $\mathbf{81.5 \pm 17.1\%}$, versus $\mathbf{66.0 \pm 16.8\%}$ for IQL. Under pure-greedy exploitation ($\boldsymbol{\varepsilon}$**=0**, frozen weights, 32 episodes per seed), QMIX reaches $\mathbf{83.8 \pm 20.5\%}$ acceptance with **95.4%** success-per-accept, versus $\mathbf{64.4 \pm 15.5\%}$ and **93.6%** for IQL (Wilcoxon: $\boldsymbol{T = 0}$, $\boldsymbol{p \leq 0.016}$; paired Cohen's $\boldsymbol{d_z = 8.6}$). Non-RL baselines (NEAREST, WEIGHTED) accept all requests (**100%**) but incur higher timeout rates (**9.3%** and **4.1%**), highlighting the value of learned selective acceptance. An ablation analysis suggests that, within this tabular-IQL vs. deep-QMIX comparison, CTDE/deep mixing—rather than reward shaping alone—better explains the gap.

Keywords: Electric Vehicle Charging, Multi-Agent Reinforcement Learning, QMIX, CTDE, DTDE, SUMO

1 Introduction

The accelerating adoption of electric vehicles (EVs) is reshaping urban mobility and placing increasing pressure on charging infrastructure [1]. In dense Smart City environments, uncoordinated charging behaviour leads to station congestion, prolonged waiting times, range anxiety in EV drivers [2], and imbalanced load across the network [3]. Centralised coordination can be effective, but becomes difficult to scale and maintain in large urban deployments [4]. Critically, these failures are experienced directly by EV drivers as longer waiting times, aborted charging sessions, and heightened range anxiety, so coordination quality is ultimately a human-centric service-quality question rather than a purely technical one.

Decentralised multi-agent reinforcement learning (MARL) provides a practical middle ground: agents can learn cooperative behaviours during training while retaining local decision-making at deployment time [5, 6]. Two paradigms dominate the literature. *Decentralised Training and Execution* (DTDE), exemplified by Independent Q-Learning (IQL) [7], treats each agent independently and incurs no coordination overhead but ignores inter-agent dynamics during learning. *Centralised Training with Decentralised Execution* (CTDE), exemplified by QMIX [8], trains a joint value function subject to the Individual-Global-Max (IGM) constraint, enabling globally consistent decentralised policies.

Despite an extensive MARL literature, controlled head-to-head comparisons on realistic EV charging scenarios remain scarce. Most prior work either evaluates a single algorithm in related charging or energy-management settings [9, 10] or relies on synthetic grid topologies without realistic traffic dynamics. To the best of our knowledge, this paper provides a controlled multi-seed comparison of IQL and QMIX for EV charging assignment on a real urban road network imported into SUMO [11].

Contributions:

1. A seven-seed, 200-episode empirical comparison of IQL and QMIX on a Dec-POMDP formulation of the EV charging assignment problem.
2. Statistical validation using Wilcoxon signed-rank tests and paired Cohen's d_z, showing higher QMIX acceptance than IQL both at episode 200 and under pure-greedy exploitation.
3. An ablation analysis (A0, A1, A3, A4; with A2 reserved for a future deep-IQL control) suggesting that the observed gap is more consistent with CTDE/deep mixing than with reward shaping alone.
4. A five-flux curriculum training protocol enabling systematic evaluation under varying traffic density.

2 Related Work

2.1 EV Charging Optimisation

Early approaches relied on centralised optimisation, formulating the problem as a mixed-integer programme [12] or a stochastic scheduling task [13]. These achieve optimal solutions in small-scale settings but do not scale to real-time, city-wide coordination. Online control and model-predictive scheduling methods [14] trade optimality for responsiveness but lack adaptation. Reinforcement learning methods [15] have been applied to single-agent charging control [9] and grid-level demand-response [10], achieving strong results in simulation. Crucially, in a related residential energy-flexibility setting, Charbonnier et al. [16] show that the standard Independent Q-Learning (IQL) approach exhibits coordination performance degradation under partial observability at scale, motivating the move to cooperative MARL paradigms. Our work extends this line to a controlled multi-agent cooperative setting with realistic urban traffic dynamics.

2.2 Multi-Agent Reinforcement Learning for Transport

MARL has been widely studied in transport-adjacent coordination settings, including autonomous driving and shared-mobility coordination [17]. In the cooperative setting, QMIX [8] and its variants QPLEX [18] and MAPPO [19] have set competitive baselines in game-playing tasks.

The application of CTDE to physical infrastructure management is growing on several fronts. For smart-grid charging, Park and Moon [20] schedule EV charging under CTDE, while Yan et al. [21] combine individual and collective policy models for cooperative control. Beyond scheduling, CTDE has been extended to V2G integration with local execution and no centralised entity at deployment [22], to cross-site EV redirection on real charging data [23], and to PV/ESS-enabled stations [24]; real-time station recommendation under traffic uncertainty has likewise been tackled with MADDPG [25, 26].

Closest to our setting, two strands address the IQL–QMIX question only in part. Wang et al. [27] apply QMIX to EV-facilitated peak shaving under dynamic traffic, but without an IQL baseline; conversely, Fu et al. [28] contrast IQL with a non-cooperative game-based MARL in large-scale urban scheduling, confirming IQL's weakness in the absence of explicit coordination. Other work is adjacent rather than comparative: federated MARL for energy management does not evaluate charging assignment under traffic uncertainty [29], and a CTDE-DDPG vs. decentralised comparison [30] (preprint) examines a residential charging setting. To the best of our knowledge, no existing work provides a controlled, multi-seed IQL vs. QMIX ablation study on a realistic urban road network under varying traffic conditions — a gap this paper fills.

2.3 Simulation Platforms

SUMO [11] with TraCI [31] provides microscopic, vehicle-level traffic simulation with real-network import. It has been used for large-scale mobility and vehicular-network simulation studies [32] and traffic management [17] but rarely in conjunction with

cooperative MARL training loops for EV charging assignment, which our framework addresses.

3 Problem Formulation

3.1 Dec-POMDP Definition

We model the EV charging assignment problem as a Decentralised Partially Observable Markov Decision Process (Dec-POMDP) [33–35]:

$$\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}, \mathcal{T}, \{r_i\}, \{\mathcal{O}_i\}, \mathcal{O}, \gamma \rangle \tag{1}$$

where $\mathcal{N} = \{1, \ldots, n\}$ is the set of station-side decision agents, $\mathcal{A}_i = \{\texttt{accept}, \texttt{reject}\}$ is the binary action space of agent i, $\mathcal{T}$ is the stochastic transition function, $\mathcal{O}_i$ is agent i's local observation, and $\gamma = 0.99$ for QMIX ($\gamma = 0.95$ for IQL). In the implemented decision layer, the MARL agents are station-side decision agents: each charging station evaluates whether to accept or reject an incoming EV charging request under its local queue, load, distance, and reservation context. EVs therefore appear as service requests/candidates rather than as the learning agents themselves.

3.2 State and Observation Space

The global state $\mathbf{s}$ encodes the joint charging network status passed to the QMIX mixing network. For QMIX, the station-side local observation fed to each shared Q-network consists of eight normalised base features plus a station-identity one-hot vector:

$$o_i = [\, \rho_i^{\text{occ}},\ q_i,\ \rho_i^{\text{res}},\ \text{soc}_i,\ d_i,\ \sin(t),\ \cos(t),\ a_{i,\text{prev}},\ \mathbf{e}_i^{\text{station}} \,], \tag{2}$$

where ρ_i^{occ} is station occupancy, q_i is queue pressure, ρ_i^{res} is reservation pressure, soc_i and d_i describe the candidate EV request, and $\mathbf{e}_i^{\text{station}}$ identifies the station under weight sharing. With $n = 6$ stations, this gives $D_{\text{local}} = 8 + n = 14$ for QMIX. Separately, the discrete state space used for IQL Q-table indexing uses binned SoC, distance, load, queue, time, and previous-action factors, with size:

$$|\mathcal{S}| = 4_{\text{SoC}} \times 3_d \times 4_{\text{load}} \times 3_{\text{queue}} \times 4_t \times 2_a = 1{,}152 \tag{3}$$

3.3 Reward Function

On the `accept` branch, the individual reward combines the outcome signal, a load-balancing cooperative bonus, a state-of-charge (SoC) urgency bonus, and three normalised shaping penalties (queue, distance, saturation):

$$r_i = R_{\text{outcome}} + \lambda_c\, r_{\text{coop}} + \lambda_s\, b_i^{\text{soc}} - \lambda_q\, p_i^{\text{queue}} - \lambda_d\, p_i^{\text{dist}} - \lambda_{\text{sat}}\, \mathbf{1}[\text{saturated}], \tag{4}$$

while the `reject` branch yields a fixed R_{rej} (no shaping). The cooperative term rewards load balancing, $r_{\text{coop}} = (\bar{l} - l_i)/c$, where $\bar{l}$ is the mean station load, l_i the target station's load, and c its capacity; the penalties are normalised to $[0, 1]$ as $p_i^{\text{queue}} =$

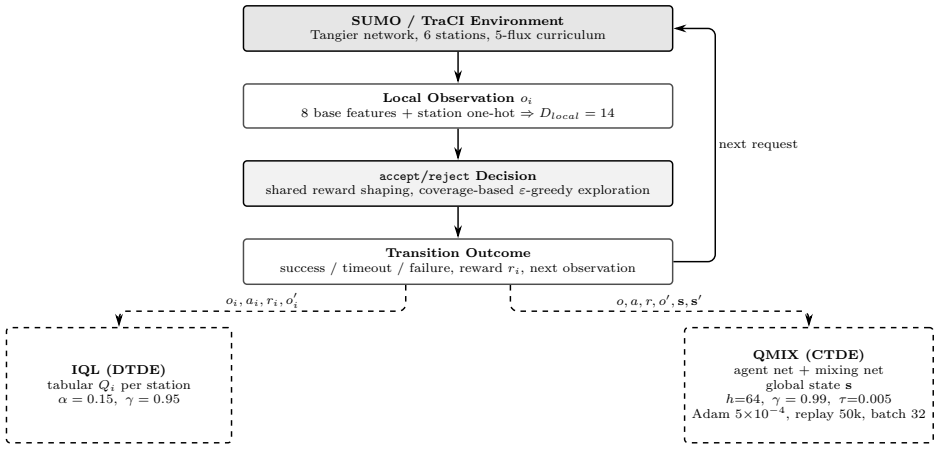


Fig. 1 System architecture for EV charging coordination. The temporal execution flow (solid lines) utilizes a shared local `accept/reject` interface. The algorithmic update flow (dashed lines) diverges depending on the training paradigm: independent tabular learning in IQL versus global-state-conditioned value mixing in QMIX.

$\min(n_{\text{wait}}, q_{\text{cap}})/q_{\text{cap}}$ and $p_i^{\text{dist}} = \min(d_i, d_{\text{cap}})/d_{\text{cap}}$, and $\mathbf{1}[\text{saturated}]$ fires when the chosen station is at capacity. Parameter values: R_{succ}=1.0, R_{fail}=−0.8, R_{timeout}=−0.6, R_{rej}=−0.1; λ_c=0.2, λ_s=0.1, λ_q=0.2, λ_d=0.1, λ_{sat}=0.3; q_{cap}=3, d_{cap}=5,000 m. The reward hierarchy $R_{\text{succ}} > R_{\text{timeout}} > R_{\text{rej}} > R_{\text{fail}}$ and all shaping weights are *identical* for IQL and QMIX, ensuring a fair comparison.

4 Algorithms

4.1 System Overview

Figure 1 summarises the shared execution loop and the training-time divergence between the two learning paradigms.

At decision time, the SUMO/TraCI environment streams EV charging requests; each charging station observes a local state o_i and emits an `accept/reject` decision. The two paradigms diverge only at training time: IQL updates each station's Q-table independently, whereas QMIX mixes per-station utilities through a monotonic network conditioned on the global state $\mathbf{s}$.

4.2 IQL (DTDE Baseline)

IQL [7] treats each agent i as an independent tabular Q-learner:

$$Q_i(o,a) \leftarrow Q_i(o,a) + \alpha\big[r + \gamma \max_{a'} Q_i(o',a') - Q_i(o,a)\big] \tag{5}$$

with α=0.15 and γ=0.95. Exploration follows a *coverage-based* adaptive ε-greedy schedule: ε starts at 0.30 and decays exponentially towards a 0.08 floor as state-space coverage grows (after a 5-episode warmup at ε=0.30). Crucially, this identical schedule

Algorithm 1 IQL training (DTDE), one episode

Initialise per-station Q-tables Q_i; set coverage-based ε
for each decision step t in the episode **do**
  for all stations i evaluating a candidate request **do**
    Observe o_i; choose $a_i \leftarrow \varepsilon$-greedy$\big(Q_i(o_i, \cdot)\big)$
  end for
  Advance SUMO; observe outcomes, rewards r_i, next obs o_i'
  for all stations i **do**
    $Q_i(o_i, a_i) \mathrel{+}= \alpha\big[r_i + \gamma \max_{a'} Q_i(o_i', a') - Q_i(o_i, a_i)\big]$
  end for
  Update ε from state-space coverage
end for

Algorithm 2 QMIX training (CTDE), one episode

Initialise shared agent net Q_θ, mixing net ϕ, target copies, replay $\mathcal{D}$
for each decision step t in the episode **do**
  for all stations i **do**
    $a_i \leftarrow \varepsilon$-greedy$\big(Q_\theta(o_i, \cdot)\big)$
  end for
  Advance SUMO; store $(\mathbf{o}, \mathbf{a}, \mathbf{r}, \mathbf{o}', \mathbf{s}, \mathbf{s}')$ in $\mathcal{D}$
  Sample a minibatch; compute $Q_{\text{tot}} = \text{Mix}_\phi\big(Q_1, \ldots, Q_6; \mathbf{s}\big)$ with $\partial Q_{\text{tot}}/\partial Q_i \geq 0$
  $y \leftarrow r_{\text{tot}} + \gamma\, Q^-_{\text{tot}}(\mathbf{o}', \mathbf{s}')$; minimise $(y - Q_{\text{tot}})^2$ via Adam
  Soft-update targets (τ=0.005); update ε from coverage
end for

is applied to both IQL and QMIX, so that performance differences are attributable to the training paradigm rather than to the exploration policy. Algorithm 1 summarises one training episode.

4.3 QMIX (CTDE)

QMIX [8] factorises the joint action-value function via a monotonic mixing network, enforcing the IGM property. Each agent's local Q-network builds on the deep Q-network architecture [36] with experience replay and a target network.

$$\frac{\partial Q_{\text{tot}}}{\partial Q_i} \geq 0 \quad \forall\, i \tag{6}$$

This guarantees that local greedy policies are globally consistent. Network configuration: hidden size 64, Adam lr $= 5{\times}10^{-4}$, replay buffer 50k, batch 32, soft target update τ=0.005. Algorithm 2 summarises one training episode.

4.4 Non-RL Baselines

Two fixed heuristics with no learning state serve as lower-bound references, using the same spatial and station signals as the MARL agents for a fair comparison. **NEAREST** assigns each request to the closest non-saturated station ($i^\star = \arg\min_{i \in \mathcal{S}_{\text{avail}}} \text{dist}(v, i)$). **WEIGHTED** assigns to the station minimising a composite cost

$$\text{score}_i = 0.40\,\hat{d}_i + 0.35\,\hat{l}_i + 0.20\,\hat{q}_i - 0.05\,u_i, \tag{7}$$

combining normalised distance $\hat{d}_i$, load $\hat{l}_i$, queue $\hat{q}_i$, and an SoC-urgency bonus u_i. Both reject a request only when *all* stations are saturated; in practice this is rare, so both effectively accept every request (hence their 100% acceptance and the absence of a learned selective-rejection mechanism).

5 Experimental Setup

5.1 Simulation Environment

Experiments run on the Aïn Dalia/Boukhalef district road network of Tangier, Morocco (OpenStreetMap export, 2026-03-08: 7.97 km^2, 11,204 edges, 2,754 junctions, imported into SUMO [11]). Six heterogeneous charging stations (2 fast, 2 medium, 2 slow) serve up to 5 concurrent EV agents over 10-hour simulated episodes (3,600 simulation steps each).

5.2 Five-Flux Curriculum

Training follows a sequential curriculum (episodes 1–140) then random flux sampling (episodes 141–200):

Table 1 Five-flux curriculum protocol.

Flux label	Demand level	Vehicles/episode	Episodes
Low	Low traffic	100	1–28
Weekend	Off-peak urban	120	29–56
Normal	Moderate	140	57–84
High	High load	200	85–112
Extreme	Saturated	250	113–140
Random mix	All above	—	141–200

5.3 Evaluation Protocol

Seeds: 7 (42, 99, 777, 1234, 2024, 5555, 9999). **Training:** 200 episodes per seed per algorithm (1,400 per algorithm; 2,800 episodes in total across IQL and QMIX). **Exploitation:** 32 episodes per seed at ε=0 with frozen weights, yielding 224 evaluation

episodes per algorithm (896 total across IQL, QMIX, NEAREST, WEIGHTED). **Metrics:** acceptance rate (%), mean reward, success-per-accept (%), timeout-per-accept (%), failure-per-accept (%), and success-per-request (%, throughput; see Table 7). **Statistics:** Wilcoxon signed-rank test on per-seed means; paired effect size Cohen's d_z (mean of per-seed QMIX−IQL differences divided by their standard deviation), consistent with the paired design. Baselines NEAREST and WEIGHTED serve as non-RL lower bounds.

6 Results

6.1 Final Performance at Episode 200

Table 2 reports per-seed results at episode 200. QMIX achieves higher acceptance and reward than IQL on all 7 seed pairs.

Table 2 Per-seed performance at episode 200.

Algo	Seed	AccRate%	Reward	Succ%	TO%
IQL	42	69.6	0.533	97.9	0.0
IQL	99	65.3	0.437	93.5	0.0
IQL	777	90.2	0.720	97.3	0.0
IQL	1234	78.7	0.637	98.7	0.0
IQL	2024	69.2	0.477	94.0	0.0
IQL	5555	43.4	0.191	89.8	2.9
IQL	9999	45.8	0.269	95.7	0.0
QMIX	42	92.6	0.676	97.7	0.0
QMIX	99	89.5	0.673	98.7	0.0
QMIX	777	93.4	0.799	100.0	0.0
QMIX	1234	92.4	0.718	98.6	0.0
QMIX	2024	88.4	0.673	97.6	0.0
QMIX	5555	51.2	0.276	94.7	0.0
QMIX	9999	62.9	0.377	95.4	0.0

Table 3 summarises aggregate statistics with Wilcoxon test. The statistic $T = 0$ (all signed ranks favour QMIX) at $n = 7$ yields $p \leq 0.016$. The paired effect size $d_z = 1.98$ on acceptance rate indicates a *very large* effect.[1]

6.2 Learning Dynamics

Figure 2 shows smoothed learning curves (w=10) with ±1 std bands across 7 seeds. Both algorithms improve through the curriculum phase (ep. 1–140). After the curriculum switch (ep. 141), QMIX converges more consistently: the IQL band widens under high traffic loads, reflecting greater variance in tabular exploration efficiency.

[1]The paired effect size is $d_z = \bar{d}/s_d$, where s_d is the sample standard deviation (ddof=1) of the per-seed differences; equivalently $d_z = t/\sqrt{n}$ for the paired t statistic. All standard deviations reported in this paper use ddof=1.

Table 3 Aggregate statistics (episode 200). Mean and standard deviation computed across the 7 per-seed means.

	AccRate (%)		Reward	
	mean	std	mean	std
IQL	66.0	16.8	0.466	0.189
QMIX	**81.5**	17.1	**0.599**	0.194
Δ QMIX−IQL	+15.5	—	+0.133	—
Wilcoxon T	0		0	
p-value	≤ 0.016		≤ 0.016	
Cohen's d_z (paired)	1.98		2.13	

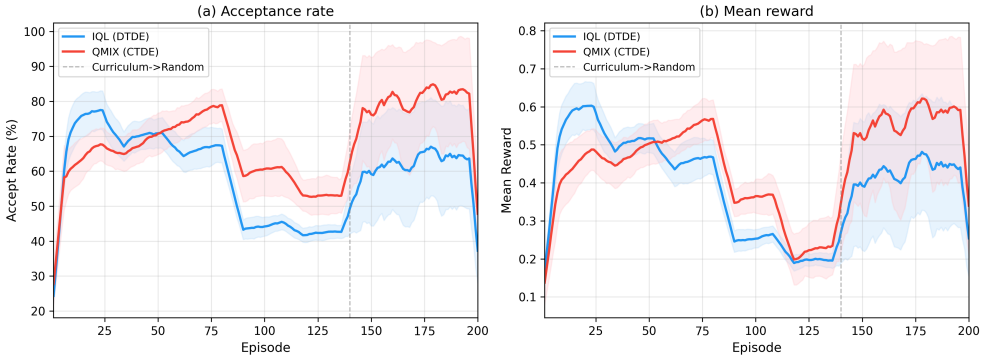


Fig. 2 Learning curves (mean ± std, 7 seeds, smoothing w=10). Dashed line: curriculum transition at episode 140.

6.3 Quality Metrics

Figure 5 shows success and timeout rates. Both algorithms achieve high success-per-accept (≥95% at convergence). Timeout converges to near zero for QMIX across all seeds; IQL seed 5555 exhibits a residual 2.9% timeout rate at episode 200, the sole anomaly across the experiments, consistent with a harder traffic initialisation regime.

6.4 Exploitation Phase (epsilon = 0, Frozen Policies)

Table 4 reports per-seed acceptance rate under pure-greedy exploitation. QMIX attains higher acceptance than IQL on all 7 seed pairs (Wilcoxon $T = 0$, $p \leq 0.016$, paired $d_z = 8.6$). Notably, IQL's best seed (70.7%, seed 777) remains below QMIX's worst seed (81.3%, seed 2024): the two acceptance-rate distributions are *fully separated*, a stronger guarantee than the effect size alone. Timeout rates remain low and not significantly different ($p = 1.0$), confirming both policies are reliable.

Table 5 compares all four algorithms. Non-RL baselines accept all requests but achieve only 67.0% (NEAREST) and 79.6% (WEIGHTED) success-per-accept, with timeout rates of 9.3% and 4.1% respectively. QMIX combines high acceptance

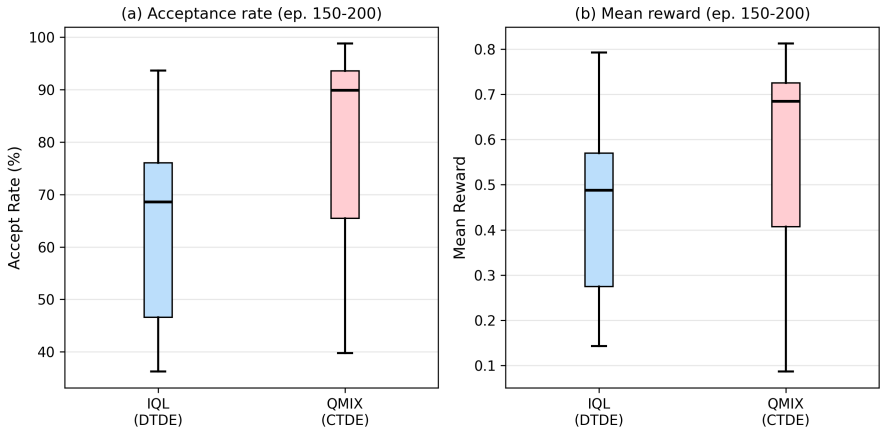


Fig. 3 Box plots of acceptance rate and reward (ep. 150–200, convergence phase, pooled across 7 seeds).

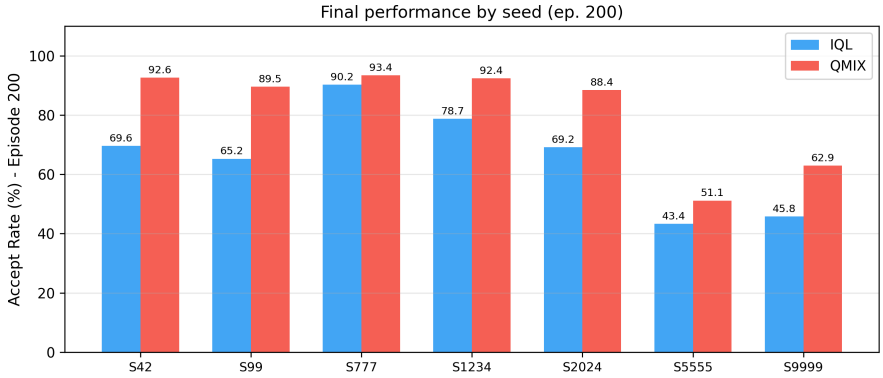


Fig. 4 Per-seed acceptance rate at episode 200. QMIX > IQL on all 7 seeds.

(83.8%) with the best reward (0.618), indicating that learned selective acceptance can outperform these heuristic strategies on the throughput-reliability trade-off.

7 Ablation Study

To isolate the source of QMIX's advantage, we run three ablation configurations (7 seeds, 50 episodes each) and report the mean acceptance rate over the final 10 episodes (convergence window). A1 removes the cooperative bonus from QMIX (λ_c=0); A3 removes experience sharing from IQL (ρ=0, cooperative reward on); A4 is a pure IQL baseline (ρ=0, no cooperation). We reserve the label A2 for a deep independent learner (deep-IQL without the mixing network), the control that would isolate the CTDE structure from the deep function-approximation class; this run is deferred to future work (Section 8). These are compared against A0, the full QMIX system; for reference we cite its frozen-exploitation acceptance (83.8%), noting that A0 is measured under

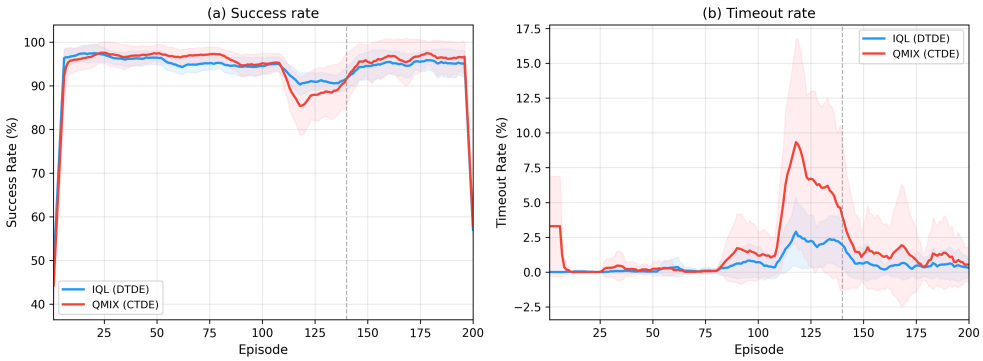


Fig. 5 Success and timeout rates over 200 episodes (mean ± std, 7 seeds).

Table 4 Per-seed exploitation results (epsilon = 0).

	IQL (DTDE)			QMIX (CTDE)		
Seed	Acc%	Succ%	TO%	Acc%	Succ%	TO%
42	63.8	93.2	3.1	82.9	92.2	5.5
99	63.8	91.7	4.0	86.9	94.7	2.6
777	70.7	92.4	3.9	86.9	92.1	5.8
1234	63.7	93.7	2.0	84.5	95.3	2.0
2024	63.3	95.6	1.2	81.3	97.7	0.2
5555	63.9	94.6	1.3	82.0	98.5	0.0
9999	61.9	94.1	1.8	82.4	97.4	0.6

pure greedy (ε=0) while A1–A4 are convergence-window estimates, so the Δ values are indicative rather than exact. Table 6 reveals two indicative patterns. First, A1 (QMIX without cooperative bonus) achieves $83.1 \pm 5.5\%$ — numerically close to A0 under this non-matched comparison — suggesting that cooperative reward shaping alone is unlikely to explain most of QMIX's advantage. Second, A3 and A4 ($72.3 \pm 2.6\%$ and $73.9 \pm 7.2\%$) both fall ∼10 pp below A0, in the IQL performance range; their mutual difference (1.6 pp) remains small relative to seed variability, so the cooperative reward does not appear to change performance materially within the IQL family either. Within this tabular-IQL vs. deep-QMIX comparison, the ablation is therefore more consistent with a CTDE/deep mixing contribution than with a reward-shaping-only explanation (see Section 8 for the function-class caveat).

8 Discussion

8.1 Why CTDE Helps

IQL's tabular Q-function cannot represent correlations between agents' simultaneous decisions. When two agents target the same station concurrently, neither detects the collision until the outcome is observed. QMIX's mixing network internalises these

Table 5 Aggregate exploitation results (ε=0). Reported mean $\pm$ std are pooled over all 7×32 episodes (per-episode dispersion); the Wilcoxon test and paired d_z are computed on the 7 per-seed means (cf. Table 4), which is why the per-seed distributions are fully separated despite the wider pooled bands.

Algorithm	AccRate%	Succ/Acc%	Timeout%	Reward
IQL (DTDE)	64.4 ± 15.5	93.6 ± 5.5	2.5 ± 4.7	0.443 ± 0.190
QMIX (CTDE)	$\mathbf{83.8 \pm 20.5}$	$\mathbf{95.4 \pm 7.0}$	2.4 ± 6.4	$\mathbf{0.618 \pm 0.254}$
NEAREST	100.0 ± 0.0	67.0 ± 25.9	9.3 ± 8.3	0.272 ± 0.448
WEIGHTED	100.0 ± 0.0	79.6 ± 21.8	4.1 ± 7.5	0.493 ± 0.403
Wilcoxon T (IQL vs QMIX)	$T = 0$, $p \leq 0.016$		$p = 1.0$ (n.s.)	
Cohen's d_z (paired)	8.6 (AccRate)		—	

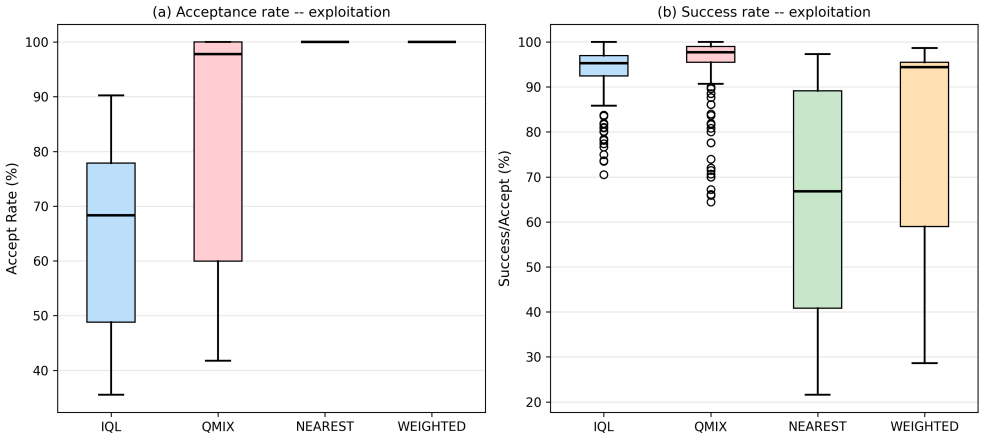


Fig. 6 Exploitation box plots: acceptance rate and success rate (ε=0, 7 seeds $\times$ 32 episodes per algorithm).

joint dynamics during training, producing policies that implicitly coordinate station selection even with only local observations at execution time.

8.2 Harder Seeds

Seeds 5555 and 9999 produce lower absolute performance for both algorithms, consistent with harder traffic patterns generated by these random seeds in SUMO's vehicle spawning. QMIX still maintains higher acceptance on these seeds, suggesting that the advantage is robust across traffic difficulty levels.

8.3 Baseline Comparison

Non-RL baselines (NEAREST, WEIGHTED) achieve 100% acceptance by never rejecting a request. However, this naïve strategy saturates stations under high flux, producing 9.3% timeout rates for NEAREST and 4.1% for WEIGHTED. QMIX's

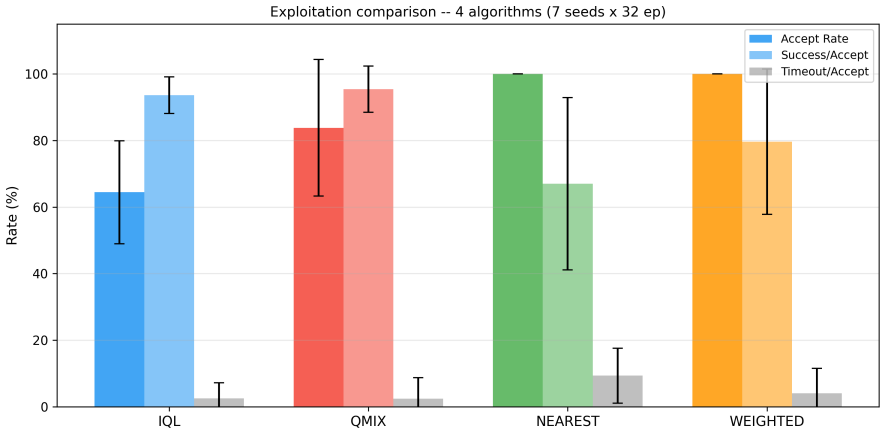


Fig. 7 Aggregate exploitation comparison: acceptance, success, and timeout rates across all four algorithms (mean ± std).

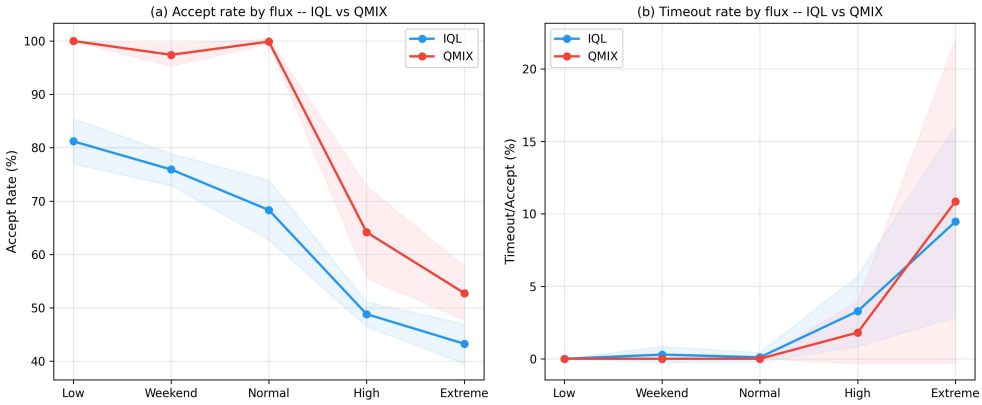


Fig. 8 Exploitation performance by traffic flux: IQL vs QMIX. QMIX consistently outperforms IQL across all flux levels.

learned policy accepts 83.8% of requests while maintaining 2.4% timeouts — selective acceptance emerges from training without explicit programming. IQL shows the same selectivity (64.4% acceptance, 2.5% timeout) but rejects too aggressively, sacrificing throughput. The gap between QMIX and WEIGHTED (83.8% vs 100% acceptance at comparable timeout) is consistent with a better throughput-reliability balance under the learned objective.

Because acceptance is a *policy action* rather than a service-quality outcome, the fair comparison rests on outcome metrics. Table 7 reports, for all four algorithms, the quality of accepted requests (success, timeout, failure per accepted) and two denominator-fair throughput measures: success per *request* (counting a learner's rejections against it) and the absolute number of successful charges per episode. The picture is unambiguous:

Table 6 Ablation study.

ID	Description	AccRate%	Δ vs A0
A0	QMIX full (exploit reference)	83.8 ± 20.5	—
A1	QMIX, no cooperative bonus (λ_c=0)	83.1 ± 5.5	−0.7
A3	IQL, no experience sharing (ρ=0, coop on)	72.3 ± 2.6	−11.5
A4	IQL pure baseline (ρ=0, coop off)	73.9 ± 7.2	−9.9

NEAREST's 100% acceptance is hollow — 23.6% of its accepted requests fail and 9.3% time out, so only 67.0% of all requests are actually served. QMIX rejects 16% of requests yet still delivers the highest useful throughput (80.9% success per request) and the most successful charges per episode (89.4 vs 75.7 for NEAREST). Even charging its own rejections as failures, QMIX serves more vehicles than the always-accept baseline. IQL, by contrast, has the lowest throughput (60.9%): it inherits selective acceptance but rejects too conservatively. Selective rejection therefore *increases* realised service rather than reducing it.

Table 7 Detailed exploitation outcomes (epsilon = 0).

Outcome metric	IQL	QMIX	NEAREST	WEIGHTED
Success / request (%, throughput)	60.9 ± 16.8	$\mathbf{80.9 \pm 22.5}$	67.0 ± 25.9	79.6 ± 21.8
Successful charges / episode	86.1 ± 19.1	$\mathbf{89.4 \pm 20.3}$	75.7 ± 14.5	84.2 ± 17.9
Success / accepted (%)	93.6 ± 5.5	$\mathbf{95.4 \pm 7.0}$	67.0 ± 25.9	79.6 ± 21.8
Timeout / accepted (%)	2.5 ± 4.7	$\mathbf{2.4 \pm 6.4}$	9.3 ± 8.3	4.1 ± 7.5
Failure / accepted (%)	3.9 ± 1.9	$\mathbf{2.2 \pm 1.7}$	23.6 ± 21.8	16.3 ± 16.3
Mean reward / decision	0.443 ± 0.190	$\mathbf{0.618 \pm 0.254}$	0.272 ± 0.448	0.493 ± 0.403

8.4 Human-Centric Perspective

The coordination objective is human-centric in both its ends and its means. On the ends, the metrics that separate the algorithms—success per request and timeout rate—are precisely the quantities a driver experiences as reliable, low-anxiety charging. On the means, decentralised execution lets each station act on local observations without a central controller, keeping individual mobility and charging data local and avoiding a single point of control—a privacy- and autonomy-preserving property relevant to citizen acceptance. Making these policies explainable to human operators, to build trust at deployment, is a natural next step (Section 9).

8.5 Limitations and Future Work

Two design choices temper the causal interpretation of our ablation. First, the two paradigms also differ in function-approximation class: IQL is a tabular Q-learner while QMIX uses deep Q-networks with a mixing network. The performance gap is therefore

attributable to the combination of centralised training *and* deep approximation, not to CTDE in isolation; a deep independent learner (deep-IQL, no mixing) would be the clean control to disentangle the two, and we leave this targeted ablation to future work. We note, however, that this confound is attenuated by the problem scale: the discretised state space is small ($|\mathcal{S}| = 1{,}152$) and likely to be substantially covered by the tabular Q-table during training, which reduces, though does not eliminate, the concern that function approximation alone explains the gap. This makes an inter-agent coordination explanation plausible, but not definitively isolated here. Second, the discount factor differs by design (γ=0.95 for tabular IQL, γ=0.99 for QMIX), reflecting the longer-horizon credit assignment the centralised critic supports; while deliberately tuned per algorithm, this remains an uncontrolled variable in the strict sense.

Beyond these, the study uses a single road network (Tangier, 6 stations, 5 agents). Extensions to larger networks, V2G integration, MAPPO/QPLEX comparison, and explainability mechanisms for human-operator trust constitute the further directions for future work.

9 Conclusion

This paper presented a controlled multi-seed comparison of IQL (DTDE) and QMIX (CTDE) for EV charging assignment in a realistic urban Dec-POMDP, with non-RL heuristics as lower-bound baselines. Across 7 seeds and 200 training episodes, QMIX achieved higher acceptance than IQL during both training (81.5% vs 66.0%) and pure-greedy exploitation (83.8% vs 64.4%), while maintaining comparable timeout rates and higher useful throughput. The ablation analysis further suggests that the observed gap is more consistent with CTDE/deep mixing than with reward shaping alone, although a deep-IQL control remains necessary for a cleaner causal attribution. These results motivate considering CTDE for urban EV charging coordination and support future extensions to larger networks, V2G integration, and explainable MARL for human-in-the-loop operation.

Acknowledgements. This work was conducted at LRIT Laboratory, Mohammed V University, Rabat, Morocco.

Declarations.

- **Funding:** This research received no external funding.
- **Competing interests:** The authors declare no competing interests.
- **Ethics approval:** Not applicable.
- **Consent to participate / for publication:** Not applicable.
- **Data and code availability:** The simulation code, configuration files, and experimental data supporting the findings of this study are available from the corresponding author upon reasonable request.
- **Author contributions:** N.-E.M. designed the experimental protocol, implemented the simulation framework, ran all experiments, and wrote the manuscript. R.A. provided scientific supervision and critical revision of the manuscript. D.K. contributed to the theoretical framework and reviewed the manuscript.

References

[1] IEA: Global EV Outlook 2024. International Energy Agency, Paris (2024). https://www.iea.org/reports/global-ev-outlook-2024

[2] Rainieri, G., Buizza, C., Ghilardi, A.: The psychological, human factors and socio-technical contribution: A systematic review towards range anxiety of battery electric vehicles' drivers. Transp. Res. Part F **99**, 52–70 (2023) https://doi.org/10.1016/j.trf.2023.10.001

[3] Ghosh, A.: Possibilities and challenges for the inclusion of the electric vehicle (ev) to reduce the carbon footprint in the transport sector: A review. Energies **13**(10), 2602 (2020) https://doi.org/10.3390/en13102602

[4] Ding, L., Cui, Y., Yan, G., Huang, Y., Fan, Z.: Distributed energy management of multi-area integrated energy system based on multi-agent deep reinforcement learning. Int. J. Electr. Power Energy Syst. **157**, 109867 (2024) https://doi.org/10.1016/j.ijepes.2024.109867

[5] Gronauer, S., Diepold, K.: Multi-agent deep reinforcement learning: A survey. Artif. Intell. Rev. **55**, 277–356 (2021) https://doi.org/10.1007/s10462-021-09996-w

[6] Amato, C.: An Introduction to Centralized Training for Decentralized Execution in Cooperative Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2409.03052 (2024). https://doi.org/10.48550/arXiv.2409.03052

[7] Tan, M.: Multi-agent reinforcement learning: Independent vs. cooperative agents. In: Machine Learning Proceedings 1993, pp. 330–337. Elsevier, San Francisco, CA (1993). https://doi.org/10.1016/B978-1-55860-307-3.50049-6

[8] Rashid, T., Samvelyan, M., de Witt, C.S., Farquhar, G., Foerster, J., Whiteson, S.: Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In: Proc. ICML, pp. 4295–4304 (2018). https://doi.org/10.48550/arXiv.1803.11485

[9] Wan, Z., Li, H., He, H., Prokhorov, D.: Model-free real-time ev charging scheduling based on deep reinforcement learning. IEEE Trans. Smart Grid **10**(5), 5246–5257 (2019) https://doi.org/10.1109/TSG.2018.2879572

[10] Ahrarinouri, M., Rastegar, M., Seifi, A.R.: Multiagent reinforcement learning for energy management in residential buildings. IEEE Trans. Ind. Inf. **17**(1), 659–666 (2021) https://doi.org/10.1109/TII.2020.2977104

[11] Lopez, P.A., *et al.*: Microscopic traffic simulation using sumo. In: Proc. ITSC, pp. 2575–2582 (2018). https://doi.org/10.1109/ITSC.2018.8569938

[12] He, Y., Venkatesh, B., Guan, L.: Optimal scheduling for charging and discharging

of electric vehicles. IEEE Trans. Smart Grid **3**(3), 1095–1105 (2012) https://doi.org/10.1109/TSG.2011.2173507

[13] Gan, L., Topcu, U., Low, S.H.: Optimal decentralized protocol for electric vehicle charging. IEEE Trans. Power Syst. **28**(2), 940–951 (2013) https://doi.org/10.1109/TPWRS.2012.2210288

[14] Tang, W., Zhang, Y.J.: A model predictive control approach for low-complexity electric vehicle charging scheduling: Optimality and scalability. IEEE Trans. Power Syst. **32**(2), 1050–1063 (2017) https://doi.org/10.1109/TPWRS.2016.2585202

[15] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, Second edition edn. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts (2018)

[16] Charbonnier, F., Morstyn, T., McCulloch, M.D.: Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility. Applied Energy **314**, 118825 (2022) https://doi.org/10.1016/j.apenergy.2022.118825

[17] Donatus, R., Ter, K., Ajayi, O.-O., Udekwe, D.: Multi-Agent Reinforcement Learning in Intelligent Transportation Systems: A Comprehensive Survey. arXiv preprint arXiv:2508.20315 (2025). https://doi.org/10.48550/arXiv.2508.20315

[18] Wang, J., Ren, Z., Liu, T., Yu, Y., Zhang, C.: Qplex: Duplex dueling multi-agent q-learning. In: Proc. ICLR (2021). https://doi.org/10.48550/arXiv.2008.01062

[19] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., Wu, Y.: The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games. arXiv preprint arXiv:2103.01955 (2022). https://doi.org/10.48550/arXiv.2103.01955

[20] Park, K., Moon, I.: Multi-agent deep reinforcement learning approach for ev charging scheduling in a smart grid. Applied Energy **328**, 120111 (2022) https://doi.org/10.1016/j.apenergy.2022.120111

[21] Yan, L., Chen, X., Chen, Y., Wen, J.: A cooperative charging control strategy for electric vehicles based on multiagent deep reinforcement learning. IEEE Trans. Ind. Inf. **18**(12), 8765–8775 (2022) https://doi.org/10.1109/TII.2022.3152218

[22] Dong, J., Yassine, A., Armitage, A., Hossain, M.S.: Multi-agent reinforcement learning for intelligent v2g integration in future transportation systems. IEEE Trans. Intell. Transport. Syst. **24**(12), 15974–15983 (2023) https://doi.org/10.1109/TITS.2023.3284756

[23] Yang, Y., Tang, R.: Cooperative multi-agent reinforcement learning for grid-aware ev charging management with cross-site redirection. Energy and AI **23**, 100664 (2026) https://doi.org/10.1016/j.egyai.2025.100664

[24] Shin, M., Choi, D.-H., Kim, J.: Cooperative management for pv/ess-enabled electric vehicle charging stations: A multiagent deep reinforcement learning approach. IEEE Trans. Ind. Inf. **16**(5), 3493–3503 (2020) https://doi.org/10.1109/TII.2019.2944183

[25] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proc. NeurIPS, pp. 6379–6390 (2017). https://doi.org/10.48550/arXiv.1706.02275

[26] Bachiri, K., Yahyaouy, A., Gualous, H., Malek, M., Bennani, Y., Makany, P., Rogovschi, N.: Multi-agent ddpg based electric vehicles charging station recommendation. Energies **16**(16), 6067 (2023) https://doi.org/10.3390/en16166067

[27] Wang, L., Liu, S., Wang, P., Xu, L., Hou, L., Fei, A.: Qmix-based multi-agent reinforcement learning for electric vehicle-facilitated peak shaving. In: Proc. IEEE GLOBECOM, pp. 1693–1698. IEEE, Kuala Lumpur, Malaysia (2023). https://doi.org/10.1109/GLOBECOM54140.2023.10437665

[28] Fu, L., Wang, T., Song, M., Zhou, Y., Gao, S.: Electric vehicle charging scheduling control strategy for the large-scale scenario with non-cooperative game-based multi-agent reinforcement learning. International Journal of Electrical Power & Energy Systems **153**, 109348 (2023) https://doi.org/10.1016/j.ijepes.2023.109348

[29] Qiu, D., Xue, J., Zhang, T., Wang, J., Sun, M.: Federated reinforcement learning for smart building joint peer-to-peer energy and carbon allowance trading. Applied Energy **333**, 120526 (2023) https://doi.org/10.1016/j.apenergy.2022.120526

[30] Shojaeighadikolaei, A., Talata, Z., Hashemi, M.: Centralized vs.Decentralized Multi-Agent Reinforcement Learning for Enhanced Control of Electric Vehicle Charging Networks. ArXiv preprint arXiv:2404.12520 (2024). https://doi.org/10.48550/arXiv.2404.12520

[31] Wegener, A., *et al.*: Traci: An interface for coupling road traffic and network simulators. In: Proc. MSWiM, pp. 155–163 (2008). https://doi.org/10.1145/1400713.1400740

[32] Codecà, L., Frank, R., Engel, T.: Luxembourg SUMO traffic (LuST) scenario: 24 hours of mobility for vehicular networking research. In: 2015 IEEE Vehicular Networking Conference (VNC), pp. 1–8 (2015). https://doi.org/10.1109/VNC.2015.7385539

[33] Bernstein, D.S., Givan, R., Immerman, N., Zilberstein, S.: The complexity of decentralized control of markov decision processes. Math. Oper. Res. **27**(4), 819–840 (2002) https://doi.org/10.1287/moor.27.4.819.297

[34] Oliehoek, F.A., Amato, C.: A Concise Introduction to Decentralized POMDPs. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28929-8

[35] Kraemer, L., Banerjee, B.: Multi-agent reinforcement learning as a rehearsal for decentralized planning. Neurocomputing **190**, 82–94 (2016) https://doi.org/10.1016/j.neucom.2016.01.031

[36] Mnih, V., et al.: Playing Atari with Deep Reinforcement Learning. arXiv preprint arXiv:1312.5602 (2013). https://doi.org/10.48550/arXiv.1312.5602