

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Khan, M. A., Burghout, W., Jenelius, E., Cats, O., & Cebecauer, M. (2026). A Reinforcement Learning-based framework for optimizing mobile charging pod operations. In *2026 IEEE Forum for Innovative Sustainable Transportation Systems, FISTS 2026* (pp. 75-80). (2026 IEEE Forum for Innovative Sustainable Transportation Systems, FISTS 2026). IEEE. <https://doi.org/10.1109/FISTS67319.2026.11421762>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

A Reinforcement Learning-based framework for optimizing mobile charging pod operations

1st Mohd Aiman Khan
Division of Transport Planning
KTH Royal Institute of Technology
Stockholm, Sweden
makhan6@kth.se

2nd Wilco Burghout
Division of Transport Planning
KTH Royal Institute of Technology
Stockholm, Sweden
wilco.burghout@abe.kth.se

3rd Erik Jenelius
Division of Transport Planning
KTH Royal Institute of Technology
Stockholm, Sweden
erik.jenelius@abe.kth.se

4th Oded Cats
Department of Transport & Planning and Division of Transport Planning
Delft University of Technology and KTH Royal Institute of Technology
Delft, the Netherlands and Stockholm, Sweden
o.cats@tudelft.nl

5th Matej Cebecauer
Division of Transport Planning
KTH Royal Institute of Technology
Stockholm, Sweden
matej.cebecauer@abe.kth.se

Abstract—The rise of autonomous electric vehicles (AEVs) presents new challenges and opportunities for an efficient and flexible charging infrastructure. This study proposes a reinforcement learning (RL) based framework for optimizing the control and operation of mobile autonomous charging pods (MAPs) for maintaining the operation of AEVs through dynamic charging. We formulate a time and energy aware Markov Decision Process (MDP) to maximize the energy delivered, and the number of AEVs serviced, while also minimizing energy consumed and increasing efficiency. We integrate this framework with SUMO to enable realistic MAP-AEV interactions. A Proximal Policy Optimization (PPO) algorithm was used to train this MDP and identify the optimal control strategies for initiating, terminating, and balancing the network. The results show that the PPO agent can service around 175 AEVs, with an efficiency of 91.5%, representing a 25% improvement over baseline greedy heuristics. Moreover, the battery capacities of AEVs can also be reduced by up to 26%, without compromising the performance. The simulation results show the potential of the proposed method in providing a flexible, and scalable charging for future transport.

Index Terms—Autonomous vehicles, Dynamic Charging, SUMO, Vehicle-to-Vehicle charging, Mobile charging stations

I. INTRODUCTION

Electric vehicles are central to the transition toward sustainable and efficient transportation systems. Advances in vehicle connectivity and automation, like autonomous electric vehicles (AEVs), can further enhance the operational performance by improving traffic efficiency, reducing operating costs, and strengthening system-level safety [1]. However, conventional charging strategies, which often involve detours, queuing, and dwell time, introduce substantial delays that undermine these benefits [2].

Dynamic charging technologies, including vehicle-to-vehicle (V2V) approaches, have emerged as promising alternatives to static infrastructure [3]. Mobile autonomous charging pods (MAPs), autonomous battery-powered units capable of

providing in-motion V2V charging, offer a mechanism to reduce downtime, enhance route efficiency, and lower AEV battery size requirements. Prior studies have demonstrated the feasibility of V2V-based charging and its potential advantages over traditional dynamic charging, including improved travel times and more efficient fleet-level energy management [4].

Past research on MAP deployment shows consistent operational benefits. In stylized networks, MAPs can reduce travel time by up to 7%, and AEV battery capacity by up to 14% [5]. Case studies on urban bus networks report cost savings of 37% when MAPs complement depot-based charging [6], [7]. However, these studies rely predominantly on simple, greedy heuristics for initiating and terminating charging and for allocating MAPs spatially within the network. Such approaches neglect long-term system dynamics and are therefore limited in optimizing overall performance [8].

To overcome these limitations, this study introduces a reinforcement learning (RL) framework that optimizes MAP deployment and operation within dynamic AEV networks simulated in SUMO. Using Proximal Policy Optimization (PPO), the framework learns policies that minimizes energy consumption while maximizing efficiency and the number of supported AEVs. Formulating the problem as a Markov Decision Process (MDP) enables explicit modeling of these temporal dependencies, while RL provides a means to derive effective policies in the highly coupled, spatio-temporal nature of MAP-AEV charging operations, which interact in a nonlinear and stochastic manner.

The key contributions of the study are:

- A PPO-based strategy for MAP assignment, termination, and redistribution that enhances charging performance in dynamic traffic environments.
- A time- and energy-aware MDP capturing the stochastic and continuous characteristics of MAP-AEV interactions.
- A comprehensive evaluation on a grid network, benchmarked against a baseline heuristic to assess efficiency,

This research was funded by the Swedish Transport Administration grant nr 2022/8287, Project SMART3.

energy use, and scalability.

The remainder of this article is organized as follows: Section II reviews related work; Section III presents the methodological framework and MDP formulation; Section IV describes the experimental setup; Section V reports and discusses the results; and Section VI concludes with future research directions.

II. RELATED WORKS

Many studies have used RL-based methods in the domain of fleet management. These studies have considered different problems, such as vehicle dispatching, routing, matching, and repositioning around the network [9]. Moreover, recent studies have also considered the potential of RL for the planning and operation of charging systems, particularly those involving mobile charging vehicles (MCVs) [10].

A deep reinforcement learning (DRL) approach for the joint planning and operation of both fixed and mobile chargers was presented in [11], demonstrating how these methods can help balance and improve infrastructure utilization. Other studies have incorporated energy and time-aware decision frameworks, and developed an MDP model with the objective of minimizing travel distance and time for MCVs, using PPO algorithm to solve this MDP [12]. In a complementary study, [13] proposed a model-free DRL method that enables MCVs to learn optimal routing and scheduling policies directly through interaction with the environment, without requiring explicit system modeling. Meanwhile, [14] introduced a multi-agent DRL (MADRL) approach that allows both MCVs and electric vehicles (EVs) to make autonomous scheduling decisions, achieving a fully distributed coordination strategy.

Compared to MCVs proposed in previous works, MAPs are considerably smaller and designed to work cooperatively to electrify a network. This allows them to be deployed closer to demand, which greatly simplifies routing; this, however, also introduces new challenges. Specifically, when to initiate or terminate charging so as to balance the network and ensure availability of pods for future demand, how to prioritize vehicles based on state of charge (SOC) or distance, and how to redistribute MAPs dynamically to maintain service balance across the network.

To the best of our knowledge, this study is pioneering in addressing these operational challenges using a reinforcement learning framework, specifically focusing on the joint optimization of MAP assignment, termination, and redistribution in a dynamic AEV environment.

III. METHODOLOGY

To model the MAP-AEV interaction, including charging, lane changing, and car following, this study uses a microscopic simulation approach implemented in the open-source traffic simulator SUMO. The autonomous behavior of AEVs and MAPs is governed by the Cooperative Adaptive Cruise Control (CACC) car-following model available in SUMO [15]. Charging interactions between MAPs and AEVs are assumed to occur through conductive robotic arms as proposed by [16].

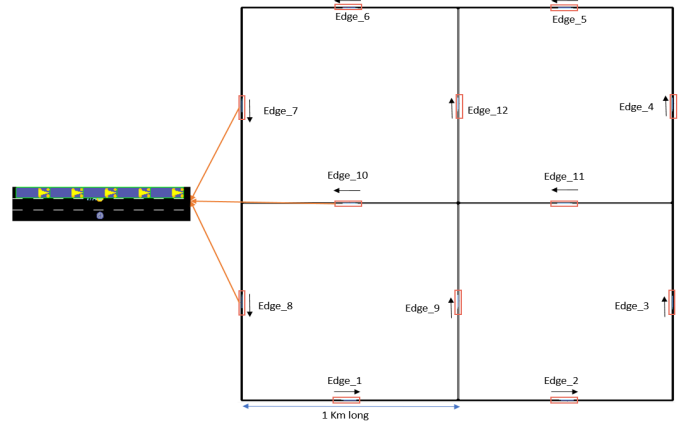


Fig. 1. Simulation network of 12 edges of 1 Km long each, with pods stationed in the middle.

The proposed framework integrates the microscopic traffic simulator, network inputs, and the MAP-AEV charging logic proposed in previous works with performance evaluation metrics to analyze the operational feasibility, energy efficiency, and potential reductions in battery capacity requirements. We consider that each MAP can charge only one AEV at a time, and AEV-AEV charge sharing is not allowed.

A. Network setup

To evaluate the feasibility and performance of the proposed MAP-based dynamic charging concept, a 2x2 grid network is developed. This network captures essential MAP-AEV interactions while maintaining computational tractability. The setup, illustrated in Fig. 1, consists of twelve directed road segments, each 1 km long, resulting in a total length of 12 km. Each edge contains three lanes, with the leftmost lane reserved for MAP operations to ensure safe parking, while also allowing it to reach the targeted AEV with little interference. The parking areas are embedded at the midpoints of these reserved lanes at each edge and also act as charging stations for the MAPs.

The network enforces unidirectional flow. AEVs enter from Edge 1 and circulate through three nested loops—first traversing edges 1,2,3,4,5,6,7,8, then 1,2,3,11,10,8, and finally 1,9,12,6,7,8, ensuring that all edges are visited. Overlapping edges produce varying demand densities, reflecting real-world heterogeneity in traffic and charging needs. AEVs operate under steady-state conditions, continuously circulating without exiting the network. All MAPs are initialized with full charge and evenly distributed across the parking sites.

B. Algorithms for AEVs and MAPs

Each AEV follows a simple behavioral logic, where after entering the simulation, it cruises at nominal speed and decelerates and changes lane when its SOC falls below a threshold, signaling the need for charging. Once charged by a MAP, it resumes cruising speed [5], [17].

The RL agent manages all MAPs collectively, making sequential decisions on assignment, charging, and redistribution.

The objective is to maximize charging service effectiveness, measured in terms of successful charge completions and network energy efficiency, while minimizing operational costs related to movement and energy use.

C. Markov decision process formulation

The operational control of MAPs is formulated as an episodic MDP, defined by the tuple

$$(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, T),$$

where $\mathcal{S}$ denotes the set of possible system states, $\mathcal{A}$ the discrete action space, $\mathcal{P}$ the transition dynamics describing the environment evolution, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ the stage-wise reward function, γ the discount factor that regulates the importance of future rewards, and T the simulation time.

At each timestep t , the agent observes the system state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$ according to its policy $\pi(a_t | s_t)$, and receives a scalar reward r_t . The environment, implemented as the coupled traffic–energy simulation framework, then transitions deterministically to the next state s_{t+1} according to the simulator dynamics $\mathcal{P}(s_{t+1} | s_t, a_t)$. This interaction continues until episode termination at $t = T$. The objective of the learning agent is to discover an optimal policy π^* that maximizes the expected discounted return over each episode.

D. State representation

At each timestep t , the environment provides a fixed-length observable state vector:

$$s_t = [g_t, \text{occ}_t(1:P), e_t(1:N_e), p_t(1:N_p)],$$

where g_t contains global network-level scalars, occ_t are normalized parking-area occupancies, P the total number of parking areas, e_t are feature vectors describing individual AEVs, and p_t are feature vectors representing MAPs.

The global feature vector is expressed as:

$$g_t = \left[\frac{n_t^{EV}}{N_e}, \frac{n_t^{pod}}{N_p} \right],$$

where n_t^{EV} and n_t^{pod} denote the numbers of active electric vehicles and charging pods at time t , normalized by their respective maxima N_e and N_p . This normalization maintains a consistent feature scale and a fixed input dimension for the policy network.

Each electric vehicle j (for $j \leq N_e$) is represented by the feature vector:

$$e_t(j) = [\hat{c}_t^j, \hat{\ell}_t^j, \hat{d}_t^j, a_t^j],$$

where $\hat{c}_t^j$ is the normalized SOC, $\hat{\ell}_t^j$ the scaled lane position, $\hat{d}_t^j$ the normalized distance to the nearest pod, and $a_t^j \in \{0, 1\}$ indicates whether the vehicle is currently assigned to a charging pod.

Similarly, each charging pod k (for $k \leq N_p$) is represented by:

$$p_t(k) = [\hat{c}_t^k, \hat{\ell}_t^k, f_t^k, a_t^k],$$

where $\hat{c}_t^k$ is the normalized pod SOC, $\hat{\ell}_t^k$ the scaled lane position, $f_t^k \in \{0, 1\}$ a binary indicator for parking or idle state, and $a_t^k \in \{0, 1\}$ denotes whether the pod is currently assigned to a vehicle.

The overall dimensionality of the observable state vector is given by:

$$\dim(s_t) = k_{\text{global}} + N_e k_{ev} + N_p k_{pod},$$

where k_{global} , k_{ev} and k_{pod} denote the dimensionalities of the global feature vector, the per-AEV feature vector, and the per-MAP feature vector, respectively, ensuring a fixed-size input for the policy network even as the number of active entities fluctuates during simulation. This structured representation enables the reinforcement learning agent to effectively perceive global network states, local charging dynamics, and the spatio-temporal interactions between MAPs and AEVs.

E. Action space

The action space is discrete. Each action encodes a high-level control intent that the environment interprets with respect to a selected vehicle. The vehicles are chosen based on a deterministic priority policy based on SOC levels and proximity to MAP locations. For example, a vehicle with a low SOC (around 20%) is prioritized for charging, while a vehicle with slightly higher SOC may also be selected if in proximity to MAPs. Typical high-level actions are:

- Assignment: select and assign a suitable charging pod to the priority vehicle based on SOC of MAPs, location, and availability. Priority is given to nearby pods with sufficient SOC, in order to reduce distance and energy consumed by these pods. After the assignment, the pod travels to the vehicle and continues following and charging in a platoon.
- Termination: terminate the active charging process for the selected vehicle when predefined stopping criteria are met (e.g., SOC range, proximity to parking area). After termination, the pods park at the nearest parking area.
- Redistribution: trigger a rebalancing heuristic that repositions pods between parking areas to improve spatial coverage and anticipate demand. This action also accounts for potential charging opportunities along the route, thereby avoiding empty trips by MAPs. Moreover, MAPs with higher SOC are prioritized for redistribution.

Invalid or inapplicable actions are handled gracefully by the environment and result in a small negative stage reward rather than hard failure.

F. Stage-based reward design

The reward at each simulation step is decomposed into three conceptual components: an *action-specific* reward r_t^{action} , a *per-step* shaping term r_t^{step} , and an additional terminal reward r_T^{term} computed at episode end. This stage decomposition supports learning of both immediate, tactical behaviors and long-term, global objectives.

1) *Action-specific reward*: When the agent issues an action, the environment computes a reward that reflects the immediate quality of that decision in the local context. For example:

- For an assignment action, the reward is a composite function of assignment urgency (a decreasing function of the target vehicle's SOC), assignment proximity (closer pods yield higher reward), and pod capability (higher pod SOC yields higher reward). The composite form is a weighted sum of normalized subcomponents.

$$r^{\text{assign}} = w_u \cdot \frac{\max(0, U - c_i)}{U} + w_d \cdot \left(1 - \frac{d_{i,p}}{d_{\max}}\right) + w_s \cdot \frac{c_p}{100},$$

where c_i and c_p are EV and pod SOC (%), U a constant controlling urgency scaling (e.g., 50%), $d_{i,p}$ is the 2D distance between EV i and pod p , $d_{\max}$ is the search distance, and w_u, w_d, w_s are tunable weights.

- For a termination action, the reward depends on the SOC of the receiving vehicle at stop: stopping within an operational SOC band yields a positive reward; while early or late yields a negative reward.

$$r^{\text{stop}}(c) = \begin{cases} -\kappa_{\text{early}} (50 - c), & c < 50\% \\ \frac{R_{\text{peak}}}{25} (c - 50), & 50\% \leq c < 75\% \\ R_{\text{peak}}, & 75\% \leq c \leq 80\% \\ -\kappa_{\text{over}} (c - 80\%), & c > 80\% \end{cases}$$

where κ_{early} and κ_{over} are the penalty coefficients and R_{peak} is the maximum reward value for charging completion. This function is piecewise and asymmetric to penalize unsafe early stopping and inefficient overcharging.

- For a redistribution action, the reward reflects improvement in spatial balance (for instance, a reduction in variance across parking-area occupancies) and the opportunity for opportunistic charging en route. Empty trips are penalized.

Each action reward is scaled so that its magnitude is commensurate with per-step shaping and terminal rewards.

2) *Per-step shaping reward*: A continuous shaping term is applied at each timestep to encode operational objectives that should persist throughout the episode. The per-step reward r_t^{step} combines several additive components:

$$r_t^{\text{step}} = r_t^{\text{charge}} - w_z |\mathcal{Z}_t| - w_e E_t^{\text{pods}} - w_m \Delta d_t, \quad (1)$$

where:

- r_t^{charge} is a positive reward proportional to the instantaneous energy successfully transferred from a MAP to an EV. It encourages sustained charging contact and efficient alignment between pods and vehicles.
- $|\mathcal{Z}_t|$ denotes the number of EVs whose SOC has reached zero at time t , associated with a strong safety penalty weighted by w_z .
- E_t^{pods} represents the electrical energy consumed by all pods during movement or charging, penalized by a weight w_e to promote energy-efficient operations.

- Δd_t denotes the total travel distance of pods during the timestep, penalized by a weight w_m to discourage unnecessary movement and excessive repositioning.

The weights w_z , w_e , and w_m are chosen such that the continuous shaping terms influence local behavior without dominating the long-term learning signal. The charging reward r_t^{charge} provides a continuous positive gradient toward maintaining active, energy-efficient charging sessions between pods and EVs, while the penalties ensure safety and operational efficiency are not compromised.

Shaping magnitudes are chosen to guide exploration without overwhelming the long-horizon objective; in practice, they are normalized relative to the typical scale of action and terminal rewards.

3) *Terminal reward*: At episode termination, the environment computes global performance metrics and applies a terminal reward that reflects long-term system success. The terminal reward is a weighted combination of components:

$$r_T^{\text{term}} = -\alpha N^{\text{zero}} + \beta \eta - \delta E^{\text{pods}} + B \cdot \mathbf{1}\{N^{\text{zero}} = 0\},$$

where:

- N^{zero} is the count of unique AEVs that reached zero SOC during the episode (primary minimization objective);
- η = (energy charged by the MAPs/ Energy provided to MAPs), is a charging efficiency ratio (maximize);
- E^{pods} is cumulative pod electrical energy use (minimize);
- B is a safety bonus for no zero-range events (optional).
- α, β , and δ are scalar weights.

All terminal terms are normalized before weighting to ensure consistent trade-offs across experimental conditions.

G. Cumulative objective

The learning objective is to maximize the expected discounted return over episodes:

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \gamma^t (r_t^{\text{action}} + r_t^{\text{step}}) + \gamma^T r_T^{\text{term}} \right],$$

where r_T^{term} the terminal reward, and π_θ the policy parameterized by θ . The discount factor γ governs the trade-off between short and long-term returns.

IV. EXPERIMENTAL DESIGN

A series of experiments are conducted to evaluate the effectiveness and operational viability of the proposed charging coordination approach. The Python implementation of the experimental framework, as well as SUMO files, are publicly available at [18]. The key simulation parameters and their respective values are presented in Table I.

A. Training configuration

The MDP was optimized using the Proximal Policy Optimization (PPO) algorithm implemented in Stable-Baselines3 [19]. PPO is well-suited to this problem due to its stable performance in discrete action spaces and its robustness to

TABLE I
PARAMETER VALUES USED IN THE SIMULATION (BEFORE SCALING)

Parameter	Value
Maximum Speed of AEVs (s_c)	36 Km/h
Maximum Speed of MAPs (s_m)	108 Km/h
Maximum battery capacity of AEVs	64 KWh
Maximum battery capacity of MAPs	200 KWh
Maximum weight of AEVs	1830 Kg
Maximum weight of MAPs	500 Kg
Charging Time (t_c)	20 minutes

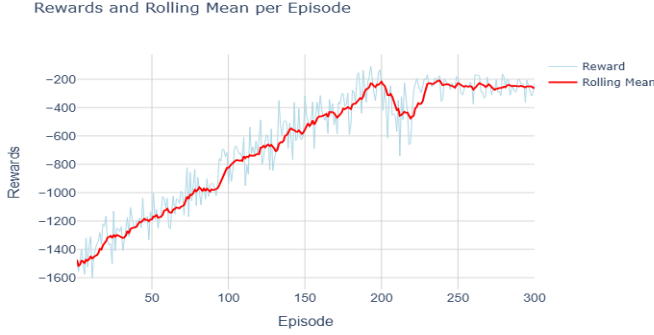


Fig. 2. Episodic reward value for the agent per episode

the stochastic and noisy feedback characteristic of microscopic traffic simulations.

Training was conducted over 300 episodes, each episode comprising 2000 timesteps, with a total wall-clock training time of approximately 48 hours. The learning rate was set to 0.003, with a batch size of 64 and a discount factor γ of 0.99. Reward weighting coefficients were tuned through short iterative runs to balance safety, charging efficiency, and pod-movement penalties. To account for variations in network scale, the most influential weights were normalized with respect to the maximum number of AEVs under each scenario and subjected to a lightweight sensitivity analysis in which key weights were varied by $\pm 50\%$. Performance trends and qualitative behaviors remained stable across these tests, indicating robustness to moderate parameter perturbations.

Convergence was assessed based on the stabilization of two performance indicators: (i) the number of discharged AEVs and (ii) MAP charging efficiency. Training consistently exhibited a two-phase learning pattern: the agent first prioritized preventing discharge events and subsequently refined its movements and charging decisions to improve overall efficiency. Figure 2 illustrates this progression.

All simulations were executed on a workstation equipped with an Intel i7-13650 CPU (1.8 GHz) and 32 GB of RAM running Python 3.11.

B. Experimental evaluation

To assess system scalability, the first experiment evaluates the maximum number of AEVs that can be sustainably supported by a fixed number of MAPs. The number of AEVs in

the network is gradually increased while maintaining a constant number of MAPs, and system performance is monitored until the number of discharged AEVs begins to rise.

Following this analysis, a second experiment investigates the minimum required battery capacity of the AEVs while ensuring that all vehicles receive sufficient charge. This experiment identifies the extent to which AEV battery sizes can be reduced without compromising system reliability.

Finally, the performance of the PPO-based strategy is benchmarked against a baseline greedy heuristic introduced in previous work [5]. The simulation scaling procedure and parameter adjustments follow the methodology established in [5], [17] to ensure comparability and consistency across studies.

V. RESULTS

The results of the SUMO-based simulations are illustrated in Fig. 3. The proposed RL approach demonstrates a significant improvement over the baseline greedy heuristic. When 60 MAPs are deployed in the network, the PPO agent successfully supports approximately 175 AEVs, representing an increase of about 52% compared to the baseline, which can only support 115 AEVs.

In terms of system efficiency, the PPO agent achieves an average efficiency of 91.5%, an increase of about 25% compared to 72.9% when using the baseline heuristic. These results indicate that the PPO-based coordination strategy not only supports more vehicles but also utilizes the available MAPs and energy more effectively, minimizing both idle operation and unnecessary pod movement.

To evaluate the potential for reducing vehicle battery capacities, an additional experiment was conducted under the scenario with 175 AEVs. The battery capacities were progressively reduced while monitoring the system's ability to maintain full operability. As shown in Fig. 4, the network remains stable and fully functional with the AEV battery capacity reduced to approximately 470 kWh, a 26% reduction from the original configuration, without compromising overall performance.

VI. CONCLUSIONS AND FUTURE WORK

This study developed a reinforcement learning-based framework for optimizing the assignment, charging termination, and redistribution of Mobile Autonomous Pods (MAPs) to enhance the performance of electric vehicle charging networks. A time- and energy-aware MDP was formulated and integrated with the SUMO microscopic simulator to capture realistic MAP-AEV interactions and operational constraints. Using PPO, the agent successfully learned adaptive charging and coordination strategies that respond to dynamic demand.

The results demonstrate that the proposed RL policy substantially outperforms the baseline heuristic. The agent can sustain up to 175 AEVs, a 52% improvement, while also achieving an average charging efficiency of 91.5%, representing a 25% increase over the baseline. The policy also enables a reduction of battery capacity requirements by up to 26%.

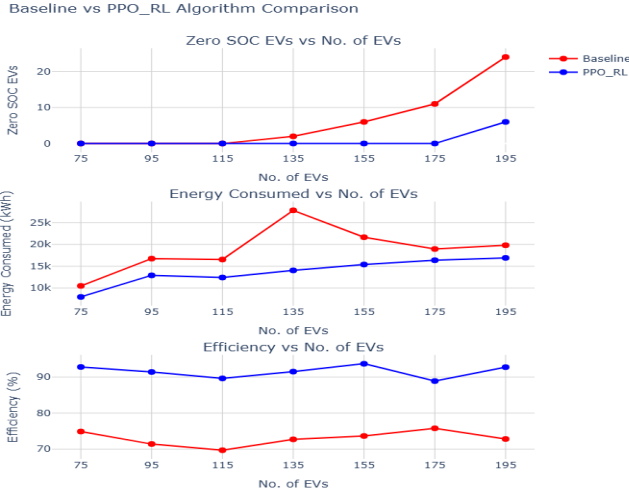


Fig. 3. Performance comparison of PPO agent with baseline methods across the number of EVs supported, energy consumption, and efficiency (η)

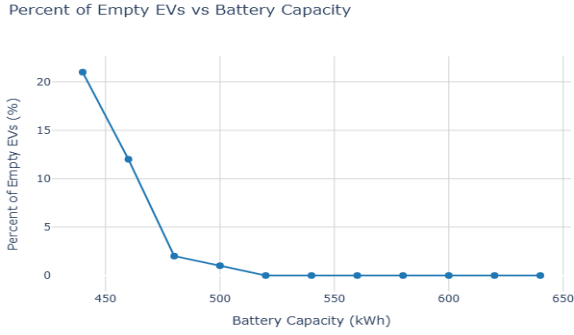


Fig. 4. Battery reduction for 175 AEVs in the network

These gains stem from effective proximity-based assignment, strategic termination of charging near parking facilities, and demand-aware redistribution that reduces empty movements and improves MAP utilization.

Despite these benefits, several limitations remain. The model is computationally intensive and requires prior knowledge of the maximum AEV fleet size, making retraining necessary when demand changes substantially. As a result, it is best suited to public transport and logistics applications with predetermined fleet sizes. Additionally, the use of deterministic SOC thresholds, idealized communication, and noise-free SUMO dynamics may limit real-world robustness. Furthermore, performance was benchmarked only against a simple greedy heuristic, indicating the need for comparisons with more competitive state-of-the-art heuristics.

Future work should incorporate additional objectives such as energy cost minimization, heterogeneous vehicle fleets, and stochastic demand. Extensions to multi-agent RL, where each MAP acts autonomously, or hierarchical RL architectures that separate high-level coordination from low-level control, could

improve scalability and responsiveness. Finally, applying the framework to larger or real-world networks and integrating it with optimization-based or pricing-based control strategies would further advance its practical applicability.

REFERENCES

- [1] D. Vignon, Y. Yin, S. Bahrami, and K. Laberteaux, "Economic analysis of vehicle-infrastructure cooperative approach for enabling automated driving," in *Transportation Research Part C: Emerging Technologies*, 2021.
- [2] C. Hecht, K. Victor, S. Zurmühlen, and D. U. Sauer, "Electric vehicle route planning using real-world charging infrastructure in germany," *eTransportation*, vol. 10, p. 100143, 2021.
- [3] M. A. Khan, W. Burghout, O. Cats, E. Jenelius, and M. Cebecauer, "Charge-on-the-move solutions for future mobility: A review of current and future prospects," *Transportation Research Interdisciplinary Perspectives*, vol. 29, p. 101323, 2025.
- [4] O. N. Nezamuddin, C. L. Nicholas, and E. C. d. Santos, "The problem of electric vehicle charging: State-of-the-art and an innovative solution," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4663–4673, 2022.
- [5] M. A. Khan, W. Burghout, O. Cats, E. Jenelius, and M. Cebecauer, "A simulation framework for evaluating mobile autonomous charging pod operations," *IEEE Open Journal of Intelligent Transportation Systems*, 2025.
- [6] —, "A comparative evaluation of mobile charging pods for electric bus operations (under review)," *Journal of Public Transportation*, 2025.
- [7] M. A. Khan, W. Burghout, E. Jenelius, O. Cats, and M. Cebecauer, "Evaluating the operational and economic feasibility of mobile charging pods for electric bus operations," in *2025 9th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2025, pp. 1–6.
- [8] S. Alaskar and M. Younis, "Scheduling a fleet of dynamic ev chargers for maximal profile," *Energies*, vol. 17, no. 23, p. 6009, 2024.
- [9] J. Gao, Y. Zhu, and O. Cats, "Uncertainties in shared mobility optimization problems: Survey and perspective," *Transportation Research Part E: Logistics and Transportation Review*, vol. 203, p. 104350, 2025.
- [10] S. Liu, Y. Wang, X. Chen, Y. Fu, and X. Di, "Smart-eflo: An integrated sumo-gym framework for multi-agent reinforcement learning in electric fleet management problem," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 3026–3031.
- [11] Y. Zhu, H. Zou, C. Liu, Y. Luo, Y. Wu, and Y. Liang, "Reinforcement learning for hybrid charging stations planning and operation considering fixed and mobile chargers," *arXiv preprint arXiv:2506.16764*, 2025.
- [12] A. C. Erüst, M. A. Beyazıt, F. Y. Taşçıkaraoğlu, and A. Taşçıkaraoğlu, "Deep reinforcement learning-based navigation strategy for a mobile charging station in a dynamic environment," in *2023 International Conference on Smart Energy Systems and Technologies (SEST)*. IEEE, 2023, pp. 1–6.
- [13] U. Qureshi, A. Ghosh, and B. Panigrahi, "Dynamic routing and scheduling of mobile charging stations for electric vehicles using deep reinforcement learning," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2024, pp. 1–5.
- [14] L. Liu, Z. Huang, and J. Xu, "Multi-agent deep reinforcement learning based scheduling approach for mobile charging in internet of electric vehicles," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 10 130–10 145, 2024.
- [15] K. N. Porfyri, E. Mintsis, and E. Mitsakis, "Assessment of acc and cacc systems using sumo," *EPiC Series in Engineering*, vol. 2, pp. 82–93, 2018.
- [16] P. Chakraborty, R. Parker, T. Hoque, J. Cruz, L. Du, S. Wang, and S. Bhunia, "Addressing the range anxiety of battery electric vehicles with charging en route," *Scientific reports*, vol. 12, no. 1, p. 5588, 2022.
- [17] Q. Chen, "Optimizing control strategies for mobile autonomous charging pods in electric vehicle networks: A simulation-based approach," 2025.
- [18] M. A. Khan, "RL model for maps." [Online]. Available: <https://github.com/AimanKhan1997/RL-model-for-MAPS>
- [19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.