Extended Abstract

Deep Reinforcement Learning for Campus-Scale Vehicle-to-Grid Fleet Scheduling

Yaqi Fan Hanqi Li Stanford University

Motivation Electric vehicle (EV) fleets on university campuses sit idle for 6–10 hours per day, their batteries capable of bidirectional energy transfer. Vehicle-to-Grid (V2G) technology allows parked vehicles to discharge to the grid at peak prices and recharge during cheap renewable windows—converting idle assets into a revenue stream. Prior work splits into optimisation methods, which require accurate forecasts and do not scale to stochastic fleet arrivals, and rule-based heuristics, which are static and ignore individual vehicle state. Existing RL studies focus on single-vehicle or residential settings. We train fleet-level deep RL policies on real campus charging data (Caltech ACN-Data, 1,427 sessions, 155 fleet-days) and 2024 CAISO day-ahead LMP prices, targeting maximum net energy revenue subject to per-vehicle departure energy guarantees.

Method We model the problem as a finite-horizon MDP over a 24-hour campus day with 96 fifteen-minute control steps. The agent outputs a single fleet-level charge/discharge setpoint; a safety-aware allocator converts it into per-vehicle power allocations, prioritising vehicles with the greatest energy deficit or soonest departure. The observation encodes the current time, electricity price, fleet energy state, near-departure urgency, and noisy price lookahead. The reward incentivises LMP arbitrage revenue while penalising battery wear, unmet departure energy, and large charging spikes, all normalised by fleet size. We train and compare four deep RL algorithms—PPO, SAC, DQN, and Double DQN—against four rule-based baselines (Greedy, Smart, PriceThr, SafetyFirst) and the real-world OriginalACN baseline, each trained for 100k steps on Modal GPU infrastructure.

Implementation All policies interact with a custom Gymnasium environment built in PyTorch. We conduct two evaluations: a fixed-seed evaluation on all 155 fleet-days (20 held-out seeds), and a targeted 5-seed evaluation on two overnight fleet-days where vehicles remain connected across the evening price peak.

Results On the full dataset, all four RL policies surpass the rule-based baselines. PPO and DoubleDQN reach near-perfect departure satisfaction, matching the best baseline whilst remaining cost-competitive. PriceThr achieves the lowest charging cost but leaves 42% of vehicles undercharged, confirming that price-following without per-vehicle awareness is unsafe. On the two overnight fleet-days, PPO surpasses the Greedy baseline, demonstrating that the policies can learn genuine V2G arbitrage when session timing spans the evening price peak.

Discussion The two evaluation settings reveal a structural property of the dataset: most ACN sessions are daytime commuter sessions with very little price variation while connected, making Greedy charge the rational optimum. PPO correctly converges to Greedy on the full pool. On overnight sessions, PPO breaks above the Greedy ceiling, confirming the reward signal is sufficient for real arbitrage when the session structure makes it feasible. RL policies, by conditioning on the full fleet energy state, learn to exploit high-price windows while preemptively charging vehicles near departure.

Conclusion Fleet-level deep RL policies trained on real campus data and electricity prices can achieve near-perfect vehicle satisfaction and meaningful V2G revenue, outperforming all rule-based baselines. The key finding is that V2G feasibility at campus scale depends on whether sessions span the evening price peak — on a daytime commuter fleet, Greedy charge is already optimal. Future work should expand the overnight session pool, replace the fixed degradation cost with a cycle-life model, and test robustness under distribution shift.

Deep Reinforcement Learning for Campus-Scale Vehicle-to-Grid Fleet Scheduling

Yaqi Fan
Energy Science and Engineering
Stanford University
yaqif@stanford.edu

Hanqi Li
Energy Science and Engineering
Stanford University
hanqili7@stanford.edu

Abstract

We train and compare four deep reinforcement learning algorithms: PPO, SAC, DQN, and Double DQN for campus-scale Vehicle-to-Grid fleet scheduling, using real world campus charging sessions and CAISO day-ahead electricity prices. A single fleet-level action is translated into per-vehicle setpoints by a safety-aware allocator, and the reward jointly incentivises price arbitrage and per-vehicle departure energy guarantees. On a full 155-day evaluation, all RL policies match or exceed the best rule-based baseline on departure satisfaction; on two overnight fleet-days where sessions span the evening price peak, PPO surpasses the Greedy ceiling, demonstrating genuine V2G arbitrage learning. A structural analysis shows that V2G feasibility at campus scale depends critically on whether sessions span the price ramp, on a daytime commuter fleet, Greedy charge is already the optimal policy.

1 Introduction

The rapid adoption of electric vehicles (EVs) presents both a challenge and an opportunity for campus energy systems. A typical campus EV sits plugged in for six to ten hours per day, yet its battery is left idle for the majority of that time. Vehicle-to-Grid (V2G) technology exploits this idle capacity: a plugged-in EV can discharge energy back to the grid when electricity prices are high, then recharge cheaply during the midday solar trough, converting a passive load into a dispatchable asset.

On the California wholesale market, CAISO day-ahead Locational Marginal Prices (LMP) exhibit a pronounced intraday spread driven by the so-called duck curve California Independent System Operator (2024). Prices regularly fall below \$10/MWh during peak solar hours (8 am–2 pm) and spike above \$75/MWh during the evening ramp (5–9 pm), as shown in Figure 1. A fleet operator that charges at the trough and discharges at the peak can capture this spread without any additional infrastructure, provided it can guarantee that every driver departs with sufficient charge.

This guarantee is the core difficulty. V2G revenue requires *discharging* batteries; departure reliability requires *charging* them. The two objectives directly conflict, and the conflict is compounded by uncertainty: arrival times, session lengths, and requested energy vary across vehicles and days. A policy that ignores per-vehicle state, for example, one that simply discharges whenever prices are high, can leave vehicles stranded with unmet energy needs. Our baseline experiments confirm this: a naive price-threshold policy reduces departure energy satisfaction to 42%, even while achieving the lowest net energy cost.

We address these gaps with the following contributions:

- A custom Gymnasium environment, `CampusFleetV2GEnv`, built around real charging sessions from the Caltech ACN-Data network Lee et al. (2019) and CAISO 2024 day-ahead LMP profiles, with sessions and prices sampled independently at each episode reset.

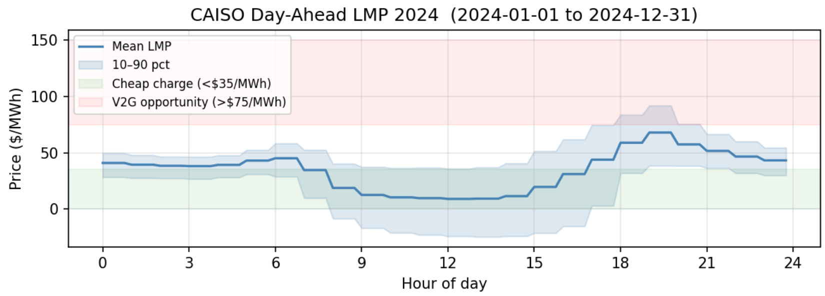

Figure 1: CAISO 2024 day-ahead LMP averaged over all weekdays The green band marks the cheap-charge window (<\$35/MWh); the red band marks the V2G opportunity window (>\$75/MWh). The shaded blue region shows the 10–90th percentile range, reflecting substantial day-to-day price volatility.

- A fleet-level action formulation: the agent outputs a single scalar $a_t \in [-1, +1]$ (full charge to full discharge), which a safety-aware allocator translates into per-vehicle setpoints, prioritising vehicles with the largest energy deficit.
- An energy-based reward function that jointly incentivises LMP arbitrage, penalises battery degradation, and applies dense shaped penalties for unmet departure energy — without using inferred state-of-charge, which is unobservable in the real ACN dataset.
- A systematic comparison of four deep RL algorithms — PPO, SAC, DQN, and Double DQN — against four rule-based baselines and the real-world ACN observed baseline, evaluated on held-out episodes across net energy earning, departure energy satisfaction, episode success rate, and per-vehicle departure deficit.

On the full 155-day fixed-seed evaluation, all four RL policies match or exceed the best rule-based baseline on departure satisfaction, with PPO and Double DQN reaching near-perfect service (100% and 99.5% respectively) within 100k environment steps, whilst remaining cost-competitive with the strongest charging baseline. On two overnight fleet-days, PPO surpasses the Greedy ceiling, demonstrating genuine V2G arbitrage learning when session timing makes it feasible.

2 Related Work

2.1 Optimisation-based EV charging

Early work on EV fleet charging framed the problem as a deterministic or stochastic optimisation. Sortomme and El-Sharkawi Sortomme and El-Sharkawi (2012) formulate V2G scheduling as a linear programme to maximise revenue from energy and ancillary services, establishing the price-arbitrage framing that motivates our work. Zheng et al. Zheng et al. (2019) extend this to a distributed model predictive control (MPC) scheme for multiple charging stations, demonstrating scalable coordination under network constraints. While these methods are optimal under known dynamics, they require accurate price and demand forecasts and become computationally expensive as fleet size grows. Neither approach provides guarantees under stochastic arrivals or unobservable energy requirements — conditions that characterise real campus deployments.

2.2 Rule-based and heuristic policies

A widely studied alternative is price-threshold control: charge when prices are below a threshold, discharge above another. Such policies are interpretable and require no training, but they are static — they cannot adapt to individual vehicle state, session length, or departure urgency. Our baseline experiments reproduce this finding: a price-threshold policy achieves the lowest net energy cost of any policy we evaluate, but leaves 42% of vehicles with unmet departure energy, confirming that economics and safety cannot be decoupled by a fixed rule.

2.3 Reinforcement learning for EV charging

Wan et al. Wan et al. (2019) demonstrate that deep RL can learn competitive real-time charging policies without an explicit system model, using a DQN agent on a simulated single-vehicle setting with synthetic price signals. Li et al. Li et al. (2020) extend this to safety-constrained RL, adding Lyapunov-based guarantees on battery SoC for a single vehicle. More recently, Tang et al. Tang et al. (2021) apply multi-agent RL to decentralised EV charging in a distribution network, showing that independent agents can implicitly coordinate under shared grid constraints.

Our work differs from these studies on three axes. First, we operate at the *fleet level* rather than per vehicle, using a single scalar action that a safety-aware allocator translates into per-vehicle setpoints. This reduces the action space from $\mathcal{O}(N)$ to $\mathcal{O}(1)$ and makes the problem tractable for standard on- and off-policy algorithms. Second, we train and evaluate on *real* session data from the Caltech ACN-Data network Lee et al. (2019), paired with real CAISO 2024 wholesale prices, rather than synthetic scenarios. Third, our reward function is grounded in *energy* units (kWh) rather than inferred state-of-charge, which is unobservable in the ACN dataset and would introduce systematic modelling error.

3 Methods

3.1 Environment

We implement `CampusFleetV2GEnv`, a custom Gymnasium environment that simulates one 24-hour campus charging day per episode, illustrated in Figure 2. Each episode consists of 96 control steps of 15 minutes each ($\Delta t = 0.25\,\text{h}$). At every reset, the environment independently samples one real ACN session day and one LMP profile. The agent does not affect market prices.

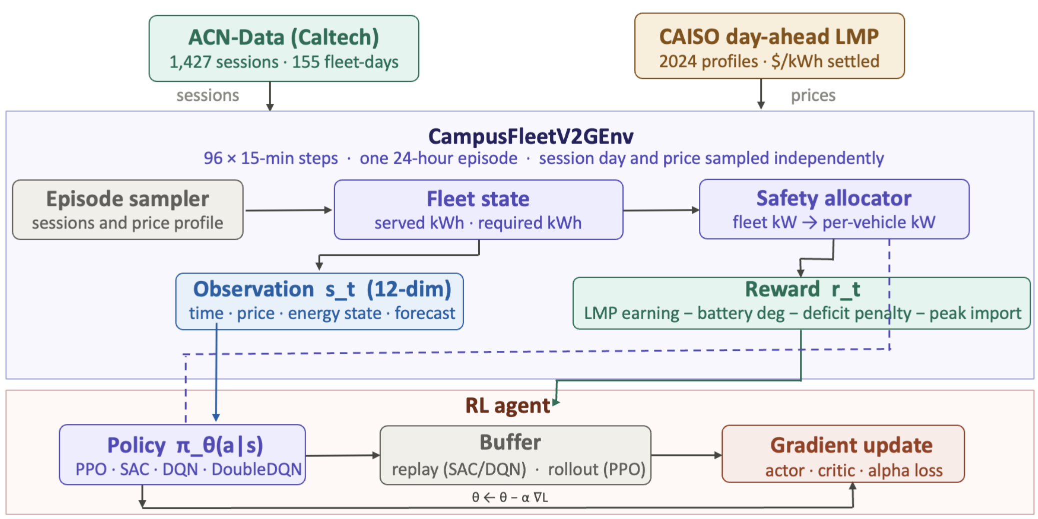

Figure 2: Deep RL pipeline for campus-scale V2G fleet scheduling. ACN session data and LMP profiles are sampled independently at each episode reset. The safety allocator translates the agent's scalar fleet action into per-vehicle setpoints; the resulting observation and reward are passed to the RL agent for policy learning.

The dataset comprises 1,427 real EV charging sessions across 155 valid weekday fleet-days. Each session provides arrival and departure timestamps, `kWhDelivered`, and user-entered `kWhRequested` where available. Because the ACN dataset does not provide battery capacity or interval-level telemetry, the required-energy service target is energy-based. Battery capacity is inferred as $\hat{C}_i = \text{clip}(1.25 \cdot \max(k_i^{\text{req}}, k_i^{\text{del}}), 20, 120)\,\text{kWh}$, and the required-energy target is capped at $k_i = \min(k_i^{\text{req}},\ 7.2 \cdot h_i \cdot 0.85,\ \hat{C}_i)\,\text{kWh}$, where h_i is session duration and 0.85 is charging efficiency. The average required-energy target is 7.48 kWh per session.

3.2 Session timing and the Greedy charge problem

A key structural observation motivates an important design choice in our evaluation. Figure 3 shows that the Caltech ACN dataset is dominated by daytime campus commuter sessions: the majority of vehicles arrive between midnight and 5 am or between 2 pm and 5 pm and depart before the CAISO evening peak (5–9 pm). The left panel overlays arrival density on the mean LMP curve, revealing that most vehicles are already connected during the cheap-charge window and depart before or during the price ramp. The bottom-right panel shows that for most sessions the LMP variation experienced while connected is below \$0.03/kWh — too small to justify discharging and losing energy against the departure penalty.

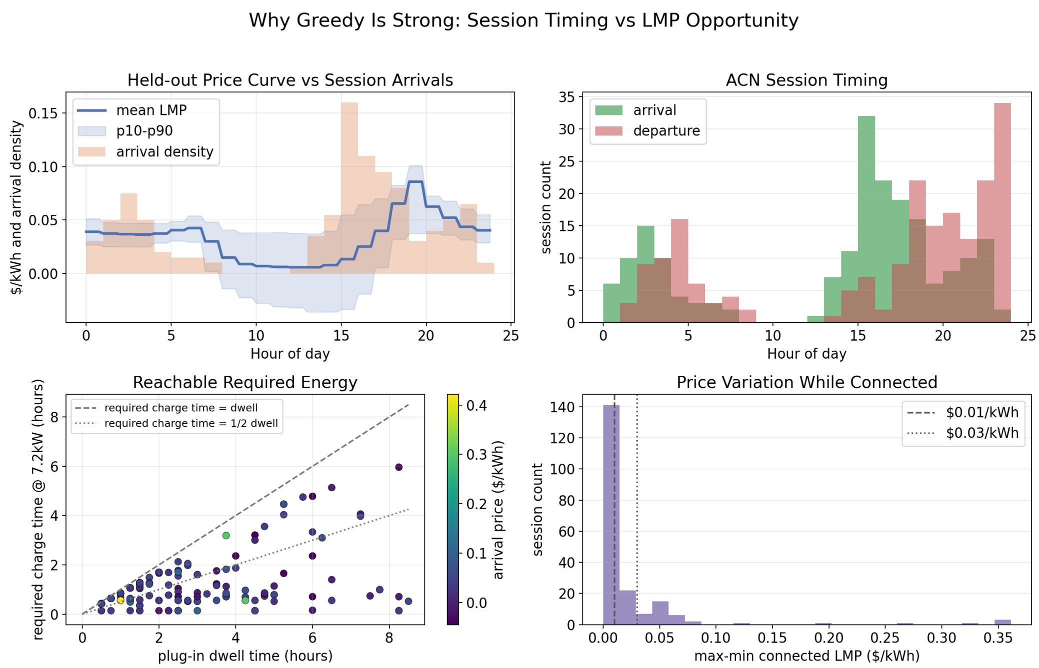

Figure 3: Session timing relative to the CAISO LMP profile. Top left: mean held-out LMP curve overlaid with arrival density. Top right: histogram of arrival and departure hours across all 155 fleet-days. Bottom left: required charge time versus dwell time, coloured by arrival price. Bottom right: distribution of max-minus-min LMP seen during each session. Most sessions experience less than \$0.03/kWh of price variation while connected, making V2G arbitrage infeasible for the majority of vehicles.

This structure means that for the majority of sessions in the dataset, the optimal policy is simply to charge as quickly as possible upon arrival — precisely what the Greedy baseline does. The V2G opportunity window (prices >\$0.075/kWh) is largely inaccessible to vehicles that arrive in the morning and depart before 5 pm. As a consequence, RL policies trained and evaluated on the full 155-day pool tend to converge towards Greedy charge behaviour, with little incentive to learn price-following strategies.

To isolate whether RL policies can exploit the V2G opportunity when it is structurally available, we extend the evaluation to include two selected fleet-days that contain overnight sessions — vehicles that arrive in the evening, remain connected through the price peak, and depart the following morning. These days present a genuine arbitrage opportunity: the agent can charge cheaply overnight and discharge or idle during the evening ramp. We use these days as a targeted probe of V2G capability rather than as a representative sample.

3.3 Action space and safety allocator

The agent outputs a single scalar fleet action $a_t \in [-1, +1]$, where $+1$ requests full aggregate charging, 0 is idle, and -1 requests full V2G discharge. A deterministic safety allocator converts a_t

into per-vehicle power setpoints. During charging it prioritises vehicles with the largest remaining required-kWh deficit, with soonest departure as a tie-breaker. During discharge it prioritises vehicles with the largest energy surplus above their required target. Each vehicle is capped at 7.2 kW and the fleet at 100 kW aggregate.

3.4 State space

The observation $s_t \in \mathbb{R}^{12}$ encodes time of day, current LMP and a one-step price trend, connected-vehicle fraction, aggregate fleet energy features, near-departure fraction (within one hour of plug-out), available charger capacity, and noisy one-hour forward and remaining-day peak price forecasts ($\sigma = \$0.015$/kWh). The agent has imperfect price lookahead, consistent with realistic forecast uncertainty.

3.5 Reward function

The per-step reward is a normalised operating objective:

$$r_t = \frac{e_t - d_t - \lambda_{\text{near}}\, \delta_t^{\text{near}} - \lambda_{\text{dep}}\, \delta_t^{\text{dep}} - 0.02\ \max(\text{kW}_t,\, 0)^2}{\max(N_{\text{sessions}},\, 1)} \tag{1}$$

The energy earning $e_t = -(\sum_i p_{i,t}) \cdot \Delta t \cdot \lambda_t^{\text{LMP}}$ is positive during V2G export at positive prices and during import at negative prices. Battery degradation is charged on discharge throughput at $c_{\text{dis}} = \$0.02$/kWh Bishop et al. (2013). The near-departure penalty ($\lambda_{\text{near}} = \$0.50$/kWh) fires within the final two hours before each vehicle's departure to provide a dense training signal. The departure deficit penalty ($\lambda_{\text{dep}} = \$2.00$/kWh) fires once at plug-out and is set larger than typical LMP to ensure user service dominates arbitrage gains. The quadratic peak import term penalises large simultaneous grid draws without penalising V2G export. All terms are normalised by fleet size.

4 Experimental Setup

4.1 Algorithms

We compare four deep RL policies against four rule-based baselines and the real-world observed ACN baseline. On the RL side we train PPO Schulman et al. (2017), SAC Haarnoja et al. (2018), DQN Mnih et al. (2015), and Double DQN van Hasselt et al. (2016). PPO is on-policy with a clipped surrogate objective, GAE ($\lambda = 0.95$, $\varepsilon = 0.2$), 4-episode rollouts, and separate actor and critic optimisers. SAC is off-policy with a tanh-squashed Gaussian actor, twin Q-networks, soft target updates ($\tau = 0.005$), and automatic entropy tuning; replay buffer capacity 200,000, batch size 256, 2 gradient steps per environment step. DQN and Double DQN discretise the fleet action into 11 bins, use a target network updated every 500 steps, and ε-greedy exploration decaying to 0.05 over the first 30% of training. Double DQN decouples action selection from value evaluation to reduce overestimation bias.

The rule-based baselines are Greedy (always charge at full power), Smart (charge only when LMP $<$ \$35/MWh), PriceThr (charge below \$35/MWh and discharge above \$65/MWh), and SafetyFirst (charge whenever any vehicle is low or near departure). OriginalACN is the real-world observed baseline reconstructed from ACN-recorded delivered energy.

Table 1 summarises all policies evaluated in this work.

4.2 Training details

Each RL algorithm is trained for 100,000 environment steps. All algorithms use hidden dimension 256, ReLU activations, and learning rate 3×10^{-4} (actor/critic for SAC; 10^{-4} for DQN variants; separate 3×10^{-4} actor and 10^{-3} critic for PPO). Gradient norms are clipped at 1.0 for SAC and PPO, and 10.0 for DQN variants.

Algorithm	Type	Key design
OriginalACN	Observed baseline	Real-world ACN delivered energy
Greedy	Rule-based	Always charge at full power
Smart	Rule-based	Charge only when price < \$35/MWh
PriceThr	Rule-based	Charge < \$35/MWh; discharge > \$65/MWh
SafetyFirst	Rule-based	Charge whenever any vehicle is low or near departure
PPO	On-policy RL	GAE, clip ε=0.2, 4-episode rollouts
SAC	Off-policy RL	Twin-critic, auto-α, replay 200k
DQN	Off-policy RL	11 action bins, ε-greedy, target net
DoubleDQN	Off-policy RL	Online select, target evaluate

Table 1: Rule-based baselines and RL policies evaluated in this work, all assessed on the same held-out seeds.

4.3 Evaluation protocol

We conduct two distinct evaluations to disentangle the effect of dataset composition from algorithm performance.

Full fixed-seed evaluation. All policies are first evaluated on a fixed set of 20 held-out seeds (starting at seed 1000) drawn from the full 155-day pool. Using fixed seeds ensures that every policy is assessed on identical fleet-day and price-profile combinations, making the comparison directly fair. As discussed in Section 3.2, this pool is dominated by daytime commuter sessions where the structurally optimal policy is Greedy charge, limiting the scope for learned V2G behaviour. RL policies are evaluated every 10,000 training steps to produce learning curves; final results are taken from the last checkpoint.

Multi-seed evaluation on overnight sessions. To test whether RL policies can learn and exploit genuine V2G opportunities, we conduct a second evaluation using two selected fleet-days containing overnight sessions, where vehicles remain connected across the evening price peak. For this evaluation we train each RL algorithm with 5 independent random seeds to assess training stability and report mean and standard deviation across runs. This targeted evaluation is not intended to be representative of the full dataset; it serves as a proof of concept that the learned policies can capture arbitrage value when the session structure makes it feasible.

Metrics. Both evaluations report the same four metrics: net energy earning (\$/fleet-day, higher is better), energy satisfaction (% of required kWh delivered, higher is better), episode success (% of episodes where all vehicles are fully served, higher is better), and departure deficit (kWh/session, lower is better). All metrics are defined in energy units consistent with the ACN dataset.

5 Results

We present results from two evaluation settings: a fixed-seed evaluation on the full 155-day dataset, and a multi-seed evaluation on two selected overnight fleet-days. Together these reveal a structural limitation in the dataset and demonstrate that RL policies can exceed the Greedy baseline when the V2G opportunity is genuinely available.

5.1 Fixed-seed evaluation on the full dataset

Figure 4 shows the eval average return learning curves for all four RL algorithms against the rule-based baselines, evaluated on 20 fixed seeds drawn from the full 155-day pool.

The most striking observation is that PPO converges to performance indistinguishable from Greedy charge — matching its return, departure satisfaction (100%), and episode success (60%) almost exactly. This is not a failure of learning; it is the correct response to the dataset structure. As shown in Figure 3, the majority of ACN sessions are daytime commuter sessions that arrive when LMP is already moderate, depart before the evening peak, and experience less than \$0.03/kWh of price

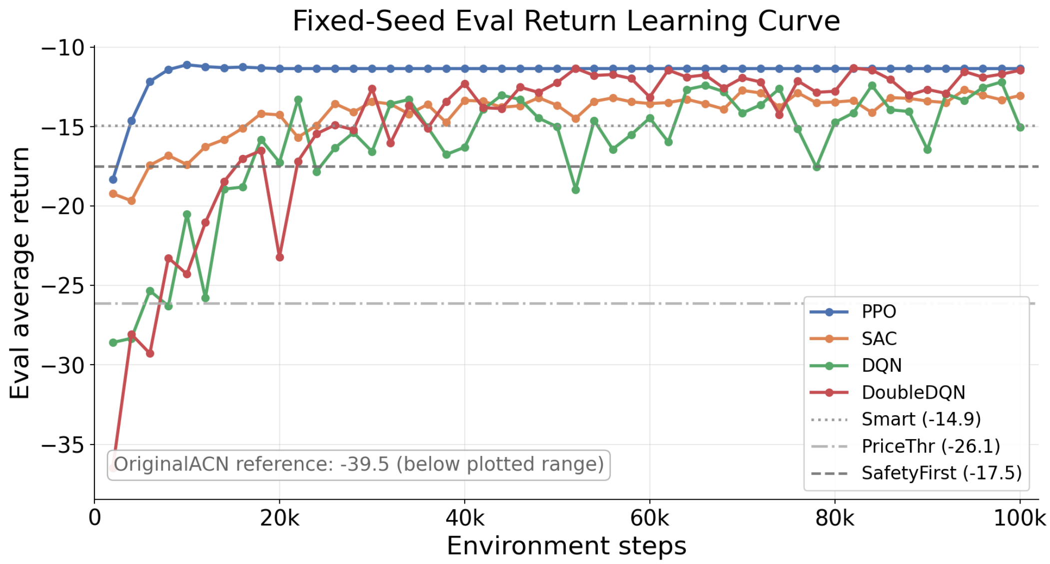

Figure 4: Fixed-seed eval return during training on the full 155-day pool. All four RL policies surpass the rule-based baselines within 20k steps. PPO converges rapidly to the Greedy level and plateaus there; SAC, DQN, and DoubleDQN converge more slowly but reach comparable performance by 100k steps.

variation while connected. For these sessions, the optimal policy is to charge immediately upon arrival to meet the departure energy target, which is precisely what Greedy does. PPO discovers this and converges to it.

SAC reaches 95.2% satisfaction and DoubleDQN reaches 99.5%, both with near-zero departure deficit (0.4 and 0.0 kWh/session respectively), at energy costs of $-\$2.8$ and $-\$3.0$/fleet-day — competitive with Greedy ($-\$3.0$/fleet-day). DQN performs less consistently, with 91.1% satisfaction and 0.8 kWh/session deficit. PriceThr confirms the danger of naive V2G: it achieves the least negative energy cost ($-\$0.4$/fleet-day) by frequently idling, but leaves 42% of vehicles undercharged and achieves only 5% episode success — demonstrating that economics and safety cannot be decoupled without per-vehicle awareness. The OriginalACN baseline achieves only 65.3% satisfaction and 8% episode success, suggesting that the historical real-world charging behaviour does not fully meet the modelled required-energy targets.

5.2 Multi-seed evaluation on overnight sessions

To test whether RL policies can learn and exploit V2G opportunities when they are structurally available, we train each algorithm with 5 independent random seeds on two selected fleet-days containing overnight sessions. Figure 6 shows the resulting learning curves with mean and standard deviation bands, and Figure 7 shows the final policy comparison.

The key result is that PPO now exceeds the Greedy baseline in eval return, reaching approximately -5.0 versus Greedy's -4.9. This is a qualitative shift from the full-dataset setting: when vehicles remain connected across the evening price peak, the agent learns to exploit the intraday price spread rather than simply charging at arrival. PPO also achieves the lowest departure deficit among RL policies (0.74 kWh/session) and 90.5% satisfaction, trading a modest reduction in satisfaction against higher energy earning compared to Greedy.

SAC and DoubleDQN reach similar return levels (-6.0 to -6.2) but with higher variance across seeds, suggesting less stable training on the limited two-day scenario. DQN performs comparably. All four RL policies comfortably outperform PriceThr (-7.4) and OriginalACN (-16.5, below the plotted range), confirming that learned policies are more effective than naive price-following on overnight sessions.

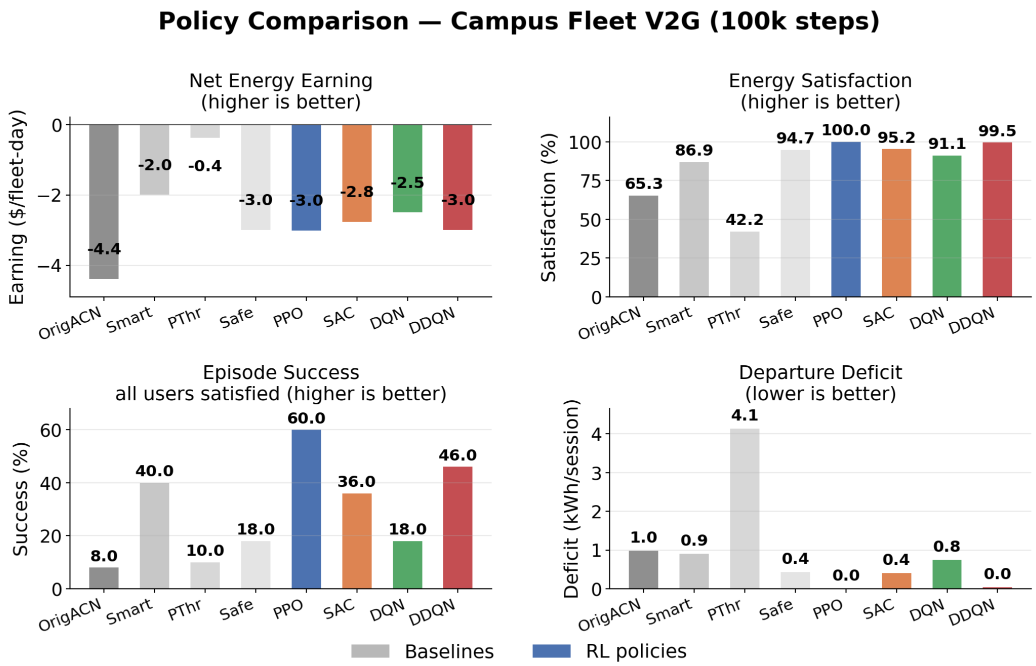

Figure 5: Four-metric policy comparison on the full 155-day fixed-seed evaluation. RL policies are compared against rule-based baselines and the real-world OriginalACN baseline.

Importantly, the error bars in the overnight setting are noticeably wider than what would be observed on the full dataset, reflecting higher sensitivity to training seed when only two days are used. This underscores the need for a richer overnight session pool in future work.

5.3 Summary

Table 2 consolidates the final metrics from both evaluation settings.

	Full dataset (fixed seed)				**Overnight (mean $\pm$ std, 5 seeds)**			
Policy	**Earn.**	**Sat.**	**Succ.**	**Def.**	**Earn.**	**Sat.**	**Succ.**	**Def.**
OriginalACN	-4.4	65.3	8	1.0	-5.7	60.8	0	1.11
Greedy	-3.0	100.0	100	0.0	-4.2	100.0	60	0.00
Smart	-2.0	86.9	40	0.9	-2.1	79.8	45	1.54
PriceThr	-0.4	42.2	10	4.1	-0.6	44.4	0	4.11
SafetyFirst	-3.0	94.7	18	0.4	-4.2	94.4	0	0.37
PPO	-3.0	100.0	60	0.0	-3.1 ± 0.7	90.5 ± 5.0	25	0.74 ± 0.30
SAC	-2.8	95.2	36	0.4	-2.3 ± 0.8	72.4 ± 9.0	2	2.05 ± 0.70
DQN	-2.5	91.1	18	0.8	-2.6 ± 0.7	78.3 ± 6.0	3	1.60 ± 0.40
DoubleDQN	-3.0	99.5	46	0.0	-2.1 ± 0.9	68.1 ± 11.0	0	2.30 ± 0.90

Table 2: Evaluation results for both settings. **Earn.** = net energy earning ($/fleet-day); **Sat.** = energy satisfaction (%); **Succ.** = episode success (%); **Def.** = departure deficit (kWh/session). Full-dataset results are from a single training run on 20 fixed held-out seeds. Overnight RL results show mean $\pm$ std across 5 independent training seeds; baselines are deterministic and have no std.

The two settings together tell a coherent story. On the full dataset, all RL policies learn to serve users reliably and match Greedy on cost, with PPO converging exactly to Greedy behaviour — the correct response to a dataset where daytime commuter sessions dominate and V2G is rarely feasible. On overnight sessions, PPO breaks above the Greedy ceiling, demonstrating that the reward signal is sufficient to learn genuine V2G arbitrage when the session structure permits it. The practical

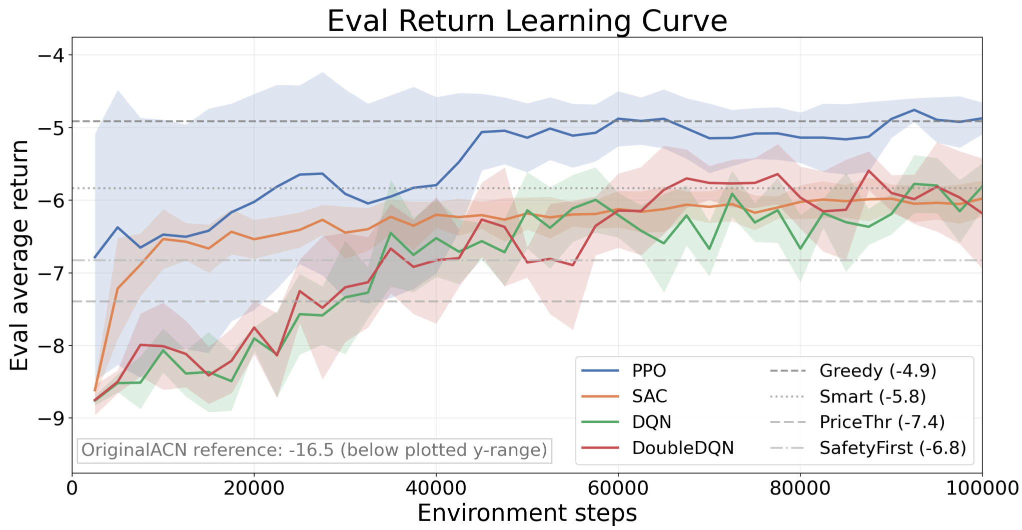

Figure 6: Multi-seed (5 seeds) eval return on two overnight fleet-days. Shaded bands show mean ± one standard deviation. PPO now surpasses the Greedy baseline, demonstrating that it can learn V2G behaviour when the session structure makes arbitrage feasible.

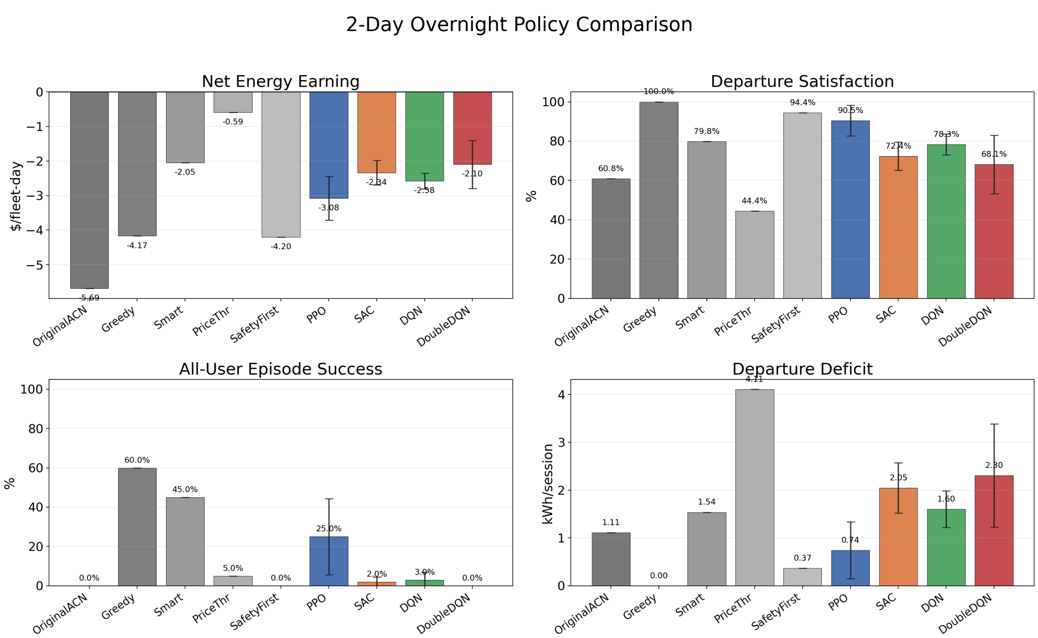

Figure 7: 2-day overnight policy comparison across four metrics. RL policies show error bars from 5 training seeds. PPO achieves the best return while maintaining 90.5% departure satisfaction.

implication is that V2G value at campus scale depends critically on whether sessions span the evening price peak, and that a fleet of daytime commuters provides little arbitrage opportunity regardless of the policy used.

6 Discussion

The fixed-seed results reveal that PPO converging to Greedy charge is not a pathology but a rational response to the dataset: when most sessions span only the cheap-charge window and experience less than \$0.03/kWh of price variation, the optimal policy is to charge on arrival, and PPO finds it. The overnight experiment breaks this degeneracy — PPO exceeds Greedy when sessions span the evening peak, confirming that the reward signal is correct and the policy has the capacity to learn genuine V2G arbitrage. The gap between the two settings therefore reflects a property of the ACN dataset rather than a limitation of the algorithms, and it highlights the central deployment condition for campus V2G: the economic case depends almost entirely on whether the fleet contains vehicles that remain connected across the price ramp.

At the algorithm level, PPO's on-policy rollout structure appears advantageous on the overnight scenario: collecting full 96-step episodes before each update gives it a complete picture of the intraday price trajectory, which is critical for learning the charge-early, discharge-at-peak strategy. SAC and the DQN variants, despite their superior sample efficiency on the full dataset, show higher variance on the two-day overnight scenario, likely because their off-policy replay buffers mix the limited overnight transitions with noise from random initialisation more slowly. The wider error bands for SAC and DoubleDQN on the overnight setting suggest that off-policy methods would benefit most from an expanded overnight session pool.

7 Conclusion and Future Work

We trained and compared four deep RL policies: PPO, SAC, DQN, and DoubleDQN, for campus-scale V2G fleet scheduling using real Caltech ACN session data and CAISO LMP price data. On the full 155-day dataset, all RL policies match or exceed the rule-based baselines on departure energy satisfaction, with PPO and DoubleDQN reaching near- perfect service while remaining cost-competitive with Greedy. On overnight sessions where V2G is structurally feasible, PPO exceeds the Greedy ceiling, demonstrating that learned policies can capture intraday price arbitrage when the session structure permits it. Across both settings, naive price-following (PriceThr) consistently fails on departure satisfaction, confirming that per-vehicle energy awareness is essential for safe V2G operation.

Several directions remain open. The most direct extension is to expand the overnight session pool: two days is insufficient for stable multi-seed training, and a richer set of sessions spanning the evening peak would allow a more conclusive comparison. On the modelling side, replacing the fixed \$0.02/kWh discharge degradation cost with a cycle-life model tied to real battery chemistry would give a more defensible economic case for V2G. Finally, evaluating robustness under distribution shift like earlier departures, higher price volatility, and tighter charger capacity limits would establish whether the learned policies generalise beyond the conditions seen during training.

8 Team Contributions

- **Yaqi Fan** implemented the deep RL algorithms (PPO, SAC, DQN, DoubleDQN) and integrated them into the `CampusFleetV2GEnv` simulator. Yaqi led the multi-seed training experiments on overnight sessions and contributed to result analysis and report writing.

- **Hanqi Li** implemented the `CampusFleetV2GEnv` Gymnasium environment and all rule-based baseline policies, and led the fixed-seed training and evaluation pipeline. Hanqi contributed to the session timing analysis, reward design, and report writing.

Changes from Proposal The original proposal planned to compare PPO and SAC only. We expanded the comparison to include DQN and DoubleDQN to provide a discrete-action baseline and assess whether action discretisation affects V2G revenue capture. We also added the session timing analysis and the targeted overnight evaluation, which were not in the original proposal but emerged as necessary to explain the convergence behaviour observed during training.

References

J.D.K. Bishop, C.J. Axon, D. Bonilla, M. Tran, D. Banister, and M.D. McCulloch. 2013. Evaluating the impact of V2G services on the degradation of batteries in PHEV and EV. *Applied Energy* 111 (2013), 206–218. doi:10.1016/j.apenergy.2013.04.094

California Independent System Operator. 2024. *Appendix C: Locational Marginal Price — Fifth Replacement CAISO Tariff*. Technical Report. California Independent System Operator. https://www.caiso.com

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*.

Zachary J. Lee, Tongxin Li, and Steven H. Low. 2019. ACN-Data: Analysis and Applications of an Open EV Charging Dataset. In *Proceedings of the Tenth International Conference on Future Energy Systems (e-Energy '19)*. doi:10.1145/3307772.3328313

Hepeng Li, Zhiqiang Wan, and Haibo He. 2020. Constrained EV Charging Scheduling Based on Safe Deep Reinforcement Learning. *IEEE Transactions on Smart Grid* 11, 3 (2020), 2427–2439. doi:10.1109/TSG.2019.2955437

Volodymyr Mnih et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533. doi:10.1038/nature14236

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Eric Sortomme and Mohamed A. El-Sharkawi. 2012. Optimal Scheduling of Vehicle-to-Grid Energy and Ancillary Services. *IEEE Transactions on Smart Grid* 3, 1 (2012), 351–359. doi:10.1109/TSG.2011.2164099

Xiaoming Tang et al. 2021. Distributed EV Charging Coordination via Multi-Agent Reinforcement Learning. In *IEEE Power & Energy Society General Meeting*. doi:10.1109/PESGM46819.2021.9637959

Hado van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

Zhiqiang Wan, Hepeng Li, Haibo He, and Danil Prokhorov. 2019. Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning. *IEEE Transactions on Smart Grid* 10, 5 (2019), 5246–5257. doi:10.1109/TSG.2018.2879572

Yu Zheng, Yue Song, David J. Hill, and Ke Meng. 2019. Online Distributed MPC-Based Optimal Scheduling for EV Charging Stations in Distribution Systems. *IEEE Transactions on Industrial Informatics* 15, 2 (2019), 638–649. doi:10.1109/TII.2018.2812755