

Article

Vehicle Heterogeneity-Aware Cooperative Dynamic Bus Control Based on Multi-Agent Reinforcement Learning for System–Individual Synergy

Hailong Zhang ¹, Haidi Wang ¹, Hanxuan Dong ^{2,*}, Zehui Ding ¹, Renjie Xiong ¹ and Hui Xu ¹

¹ The School of Mechanical and Electrical Engineering, North University of China, Taiyuan 030051, China

² The School of Electrical and Control Engineering, North University of China, Taiyuan 030051, China

* Correspondence: 20220144@nuc.edu.cn

Abstract

Under the trend of intelligent transportation and connected vehicles, real-time control plays a vital role in improving bus system efficiency. Existing bus control strategies typically treat buses as homogeneous points and achieve system equilibrium by maintaining consistent headways. However, this simplification overlooks differences in dynamic responses and the evolution of powertrain lifespan arising from vehicle heterogeneity. It converts the sparse constraint problem, which is intended to ensure timely arrival, into a hard constraint on the vehicle trajectory over the entire time horizon, thereby excessively restricting individual optimal evolutionary paths and causing the optimization process to become trapped in a local optimum. To this end, this paper proposes SMATD3, a multi-agent cooperative control algorithm that accounts for vehicle heterogeneity. By adopting a centralized training and decentralized execution paradigm and avoiding the specification of a fixed inter-vehicle spacing target, the algorithm enables each vehicle to adaptively adjust its speed control strategy according to its own dynamic characteristics, thereby achieving the coordinated optimization of system equilibrium and individual objectives. The simulation results indicate that the proposed method can effectively suppress bus tailgating and achieve the coordinated multi-objective optimization of operational stability, passenger travel efficiency, energy consumption, and battery health. From a sustainability perspective, improved headway regularity and service reliability can enhance public transit attractiveness and support mode shift, while smoother energy use and reduced battery degradation lower lifecycle impacts.

Keywords: bus bunching; reinforcement learning; adaptive cruise control; energy management



Academic Editor: Juneyoung Park

Received: 7 May 2026

Revised: 29 May 2026

Accepted: 3 June 2026

Published: 5 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Public transportation is a core component of multimodal urban transportation systems and plays a crucial role in improving the efficiency and sustainability of urban mobility [1]. Advanced public transportation systems can reduce the number of private vehicles, thereby alleviating traffic congestion and lowering transportation-related emissions [2]. Driven by electrification and intelligent technologies, electric buses are rapidly replacing conventional fuel-powered buses and are gradually becoming the mainstay of urban public transportation [3]. While reducing greenhouse gas emissions, this transition also significantly decreases the transportation sector's reliance on fossil fuels, providing an important pathway for the low-carbon transition and sustainable development of urban

transportation. Beyond operational efficiency, bus operation regularity is closely tied to sustainability outcomes. Stable and reliable headways reduce passenger waiting time uncertainty, improve perceived service quality, and help increase the competitiveness of public transit, which is essential for reducing private car dependence and related urban emissions. For electric buses, smoother speed control also reduces traction power fluctuations and battery stress, which can improve energy economy and extend battery lifetime, thereby lowering lifecycle environmental impacts and replacement-related resource use.

Public transport efficiency is strongly influenced by the regularity and coordination of bus operations [4,5]. From a system-level perspective, inefficient bus control not only causes deviations from the timetable but also reduces the effective utilization of fleet capacity. When headways become uneven, delayed buses accumulate more passengers and require longer dwell times at stops, while following buses may carry fewer passengers and operate with underutilized capacity. This imbalance propagates along the route, resulting in service gaps, bus bunching, longer passenger waiting times, increased travel times, and reduced route turnover efficiency [6]. Therefore, improving public transport efficiency requires a control strategy that can simultaneously enhance service regularity, passenger service quality, vehicle operating speed, and resource utilization, rather than merely enforcing uniform vehicle spacing.

Quantitative evidence further demonstrates the practical importance of electric bus operation control. In a representative high-frequency bus route study, uncontrolled operation resulted in the first bus bunching event at 27 min, an average bus travel time of 47 min, and a headway-deviation-based instability parameter of 4.08. By contrast, real-time control prevented bunching during the evaluated operation period, reduced the average travel time to 35 min, and decreased the instability parameter to 0.05 [7]. These findings indicate that uncontrolled bus operations can rapidly produce bunching, irregular headways, and substantial passenger travel time loss. From the electric bus operation perspective, real-world evaluations also show that battery electric buses incur non-negligible electricity costs, reaching USD 0.45 per mile for a fast-charge fleet and USD 0.42 per mile for a depot-charge fleet [8]. Battery replacement is also an important lifecycle cost component, and battery life variations can substantially influence fleet-level economic performance [9]. Therefore, electric bus operation control is practically important not only for improving headway regularity and reducing passenger delay but also for supporting energy-efficient and battery-health-aware fleet operation.

However, in mixed traffic environments, time-varying disturbances in inter-stop travel times interact with stochastic passenger arrivals at stops, inevitably causing deviations in bus arrival times [10]. If a bus is delayed, a longer passenger queue will have accumulated by the time it arrives at the stop, resulting in increased boarding demand and longer dwell time, which in turn further exacerbates the delay. At the same time, because the leading bus has already absorbed most of the passenger demand, following buses have shorter dwell times and operate relatively faster, gradually catching up with the bus ahead. This ultimately results in bus bunching and service gaps, significantly increasing passengers' waiting and travel times [11]. This phenomenon, known as bus bunching, has been widely discussed by scholars, leading to two main categories of control methods: station-based control [12] and inter-stop control [13]. Station-based control strategies include holding control [14] and stop skipping [15]. Holding control adjusts the relative spacing between vehicles by extending or shortening dwell times, thereby preventing bus bunching [16–19]. The stop skipping strategy restores fleet balance by allowing some buses to skip certain stops, thereby reducing dwell time [20,21]. However, the Holding strategy prolongs dwell time, which may reduce operational efficiency and easily lead to service gaps, especially during high-demand periods. The stop skipping strategy may degrade the passenger

experience at skipped stops. Overall, while these strategies work well in mitigating bus bunching, they increase passenger travel time and lower passenger satisfaction. Furthermore, as these methods are essentially reactive measures, they can hardly provide proactive intervention for delays that have already occurred.

From the passenger behavior perspective, holding control may improve headway regularity but introduces additional delay for onboard passengers, whereas stop skipping can shorten in-vehicle time while increasing waiting time or changing passengers' service choices. Lee et al. [22] incorporated smartcard-derived train choice behavior into a skip stop strategy, demonstrating that service control should be evaluated in terms of passenger responses as well as operational efficiency. A related passenger-oriented perspective was provided by Wen et al. [23], who incorporated dynamic passenger demand and passenger-related costs into a Soft Actor–Critic-based collaborative timetable optimization framework for large-scale urban rail transit networks. Although developed for rail transit, this study reinforces the need to assess real-time public transport control based on its passenger service consequences in addition to operational stability.

With the advancement of vehicle-to-everything (V2X) technology, V2V and V2I communications have made it possible for vehicles in public transportation systems to exchange real-time information and cooperate with one another [24,25]. On this basis, inter-stop speed control strategies dynamically regulate headways by adjusting vehicle speeds along the route, enabling a more timely response to traffic fluctuations and avoiding the delays associated with station-based control. Accordingly, such strategies, as well as hybrid methodologies incorporating inter-stop speed control, constitute more effective approaches for the coordinated control of public transportation systems [7,26–29]. Subsequently, researchers have proposed various speed planning methods to pursue green, ecological, energy-saving, and safe driving objectives.

However, existing methods transform the sparse constraint problem, which should ensure arrival timing, into a hard constraint on the vehicle trajectory over the entire time domain. Lacombe et al. proposed a bilevel optimization method that converts sparse arrival and headway constraints into hard constraints on the full-time-domain speed trajectories of electric bus fleets [30]. Jin et al. developed an energy-optimal speed control method that transforms sparse station arrival constraints into hard constraints on full-time-domain eco-driving trajectories considering passenger load [31]. Han et al. presented a speed planning approach based on multi-objective nonlinear model predictive control (NMPC), which translates sparse punctuality constraints into hard constraints on the full-journey speed trajectories of connected electric buses [32]. Shi et al. established a distributed deep reinforcement learning-based dynamic bus control system, which fuses station-based and inter-stop control strategies and converts sparse arrival timing constraints into hard constraints on full-domain vehicle speed and dwelling time trajectories [33]. These methods fail to retain the flexible characteristics of sparse timing constraints and instead impose rigid restrictions on the full-time-domain trajectory, thereby excessively restricting the optimal evolution path that each vehicle should select according to its own characteristics. The fundamental requirement of bus operations is to ensure that vehicles arrive at stops on schedule; however, this excessive constraint prevents the system from approaching a globally optimal solution, resulting in an equilibrium that is superficial and local rather than truly global.

A further critical limitation is that virtually all inter-stop control strategies implicitly assume that buses can be treated as homogeneous particles, thereby ignoring the objective reality of vehicle heterogeneity. In fact, differences in passenger load lead to significant variations in vehicle dynamic response. The energy consumption of electric vehicles changes over the course of their service life; even when the same speed control commands

are applied, the operating conditions of each vehicle's powertrain and the rate of battery health degradation exhibit individual differences [34]. Although this issue has not been fully addressed in studies of inter-stop control, it has remained a key challenge in research on vehicle powertrain control. For example, Wei et al. conducted a study on energy management strategies that account for passenger load, using power distribution between the engine and the drive motor in hybrid vehicles to adapt to the varying effects of different passenger loads on the powertrain [35]. Jia et al. further examined differences in battery life degradation under various speed profiles and power distribution schemes in the energy management of fuel cell hybrid buses [36]. It is worth noting that these energy management strategies all rely on adjusting the operating conditions of the powertrain under specific vehicle speed profiles, whereas inter-stop control can fundamentally improve the operating conditions of the vehicle's powertrain through speed regulation. Therefore, achieving the coordinated optimization of system-wide balance and individual objectives while respecting vehicle heterogeneity remains a key challenge that existing research has yet to resolve.

To address these limitations, this paper incorporates the real-time dynamic state of each vehicle into the multi-vehicle collaborative scheduling optimization problem of the bus system, enabling each vehicle to adaptively adjust its speed according to its own heterogeneity. On this basis, a SMATD3-based multi-agent reinforcement learning algorithm is proposed to autonomously facilitate inter-vehicle cooperation through self-learning. During policy parameter optimization, a sparse spacing penalty mechanism is introduced, in which penalties are imposed only when vehicle spacing becomes severely imbalanced; otherwise, the optimization focuses on passenger waiting time, speed efficiency, and practical battery health performance indicators. This approach allows vehicles to freely select the optimal evolution path according to their own heterogeneity, thereby enabling the coordinated optimization of system equilibrium and individual objectives and avoiding the local optimum problem caused by traditional continuous spacing control. At the same time, by adopting a twin-critic architecture, target policy smoothing, and delayed policy updates, the approach effectively mitigates value overestimation and policy oscillation without altering the distributed execution structure, thereby significantly enhancing the stability and robustness of the training process. The contributions of this paper can be summarized as follows:

1. A multi-agent reinforcement learning framework integrating vehicle heterogeneity is constructed, enabling each vehicle to adaptively regulate its inter-station travel speed based on its own dynamic attributes, thereby achieving self-learning multi-vehicle coordination on a closed-loop bus route.
2. A sparse spacing penalty strategy is proposed to relax overly restrictive adjustment constraints on vehicle systems, thereby preventing bus bunching while improving bus system operational efficiency and extending vehicle service life, ultimately achieving the coordinated optimization of system-level and individual-level objectives.
3. A twin-critic architecture, target policy smoothing, and delayed policy update mechanisms are introduced to mitigate value function overestimation and policy oscillations, thereby enhancing training stability and robustness.

This paper is organized as follows: Section 2 presents the problem statement and system modeling; Section 3 introduces the proposed SMATD3 framework and algorithm design; Section 4 describes the experimental setup and analysis of the results; and Section 5 concludes this paper and discusses future research directions.

2. Experimental Modeling

This study develops a closed-loop circuit simulation environment based on multi-agent reinforcement learning to systematically evaluate the comprehensive performance of the proposed cooperative dynamic control strategy under dynamic stochastic perturbations. The simulation scenario comprises a closed circular circuit of total length $L = 20,000$ m, uniformly partitioned into $S = 20$ stations with an inter-station spacing of $d_s = 1000$ m. The system incorporates five identical electric buses, each initialized at a full state of charge (i.e., initial SoC = 1), operating continuously along the closed circuit. Vehicle parameters are provided in Table 1. The simulation spans a complete representative operational day, with a total duration of $T_{\text{sim}} = 16$ h, corresponding to 57,600 s. A discrete time step of $\Delta t = 1$ s is adopted to ensure the high-fidelity characterization of high-frequency events, including start–stop maneuvers, station entry and exit, and acceleration and deceleration profiles, thereby satisfying the data volume and temporal resolution requirements essential for reinforcement learning training.

Table 1. Main parameters of power-split PHEB.

Symbol	Parameters	Values
Driving motor	Maximum power	178 kW
	Maximum torque	830 Nm
	Maximum speed	2000 rpm
Battery	Type	Li-ion
	Capacity	44 Ah
	Voltage	592 V
Transmission	Ratio of final single gear	6.43
Vehicle	Curb weight	18,000 kg
	Air resistance coefficient	0.65
	Cross-section	7.48 m ²
	Rolling resistance coefficient	0.0095

For clarity, we define the set of vehicles and the set of stations as follows:

$$\mathcal{B} = \{B_1, B_2, \dots, B_5\}, \quad \mathcal{S} = \{S_0, S_1, \dots, S_{19}\} \quad (1)$$

Specifically, the arc-length coordinate of station S_i along the closed circuit is given by

$$x_i = i \cdot d_s, \quad i \in \{0, 1, \dots, S - 1\}. \quad (2)$$

For each vehicle $B_b \in \mathcal{B}$, its arc-length position at discrete time step k is denoted by $x_b(k) \in [0, L)$, with its velocity and acceleration denoted by $v_b(k)$ and $a_b(k)$, respectively. To accommodate the cyclic nature of the closed circuit, the vehicle position is updated via a modulo operation to enforce periodic boundary conditions:

$$x_b(k+1) = (x_b(k) + v_b(k)\Delta t) \bmod L \quad (3)$$

where $(\cdot) \bmod L$ denotes the modulo operation with respect to L , which constrains the vehicle position to the interval $[0, L)$, thereby accurately capturing the return to the origin upon the completion of a full circuit.

2.1. Passenger Demand Model

In public transit systems, passenger arrival patterns exhibit pronounced randomness and temporal variability. Notably, during peak periods, such as morning and evening rush hours, short-term fluctuations in passenger flow at stops substantially amplify the uncertainty of vehicle dwell times and may, through delay propagation, precipitate bus clustering. To characterize the spatiotemporal dynamics of passenger demand, this study employs a non-homogeneous Poisson process to model passenger arrival behavior at bus stops. It should be noted that the passenger demand parameters in this study are not calibrated from a specific AFC or APC dataset. Instead, they are designed as normalized scenario parameters to generate a representative and repeatable demand environment for algorithm evaluation. The model aims to reproduce typical urban bus demand patterns, including stochastic arrivals, morning and evening peaks, school-related demand concentration, lower demand during low-demand periods, directional OD asymmetry, and short-to medium-distance trip dominance.

The arrival intensity of passengers departing from station s_i and destined for station s_j at time t is denoted by $\lambda_{i,j}(t)$. For any sufficiently small time interval $[t, t + \Delta t)$, the number of arrivals $A_{i,j}(t)$ associated with the origin destination (OD) pair (i, j) follows a Poisson distribution:

$$A_{i,j}(t) \sim \text{Poisson}(\lambda_{i,j}(t)\Delta t). \quad (4)$$

To reflect the spatiotemporal heterogeneity inherent in actual transit demand, this study formulates the intensity function $\lambda_{i,j}(t)$ as being jointly governed by the baseline demand level, time-of-day effects, station functional characteristics, and distance decay effects, thereby capturing the stochasticity, periodicity, and spatial distributional patterns of passenger demand at the modeling level.

$$\lambda_{i,j}(t) = \lambda_{\text{base}}(t) \cdot \delta_{i,j} + \lambda_{\text{peak}}(t) \cdot I_{\text{peak}}(i, j) + \lambda_{\text{school}}(t) \cdot I_{\text{school}}(i, j) \quad (5)$$

where $\lambda_{\text{base}}(t)$ denotes the baseline demand component:

$$\lambda_{\text{base}}(t) = \lambda_0 \cdot \eta(t) \quad (6)$$

where λ_0 denotes the baseline intensity constant, and $\eta(t)$ represents the intraday fluctuation factor, which captures the diurnal variation in demand. In this study, $\eta(t)$ is assigned a value of 1.2 during daytime hours and 0.8 during nighttime hours, thereby reflecting the periodic nature of demand fluctuations without introducing superfluous parameters.

The term $\lambda_{\text{peak}}(t)$ denotes the additional demand intensity during morning and evening peak periods; in this study, $\lambda_{\text{peak}}(t)$ is assigned a value of 0.1. This component is activated for specific OD directions via the indicator function I_{peak} to characterize the directional asymmetry inherent in peak period commuting flows.

The term $\lambda_{\text{school}}(t)$ denotes the additional demand intensity during school arrival and dismissal periods; in this study, $\lambda_{\text{school}}(t)$ is assigned a value of 0.3. This component is activated for specific demand segments via the indicator function I_{school} .

The term δ_{ij} denotes the distance decay function, which serves to attenuate the proportion of implausibly long-distance trips, thereby reflecting the empirical regularity that demand intensity diminishes with increasing travel distance. This study adopts a decay formulation based on the shortest circular distance:

$$\delta_{ij} = \frac{1}{1 + 0.2 \text{dist}(i, j)}, \quad \text{dist}(i, j) = \min(|j - i|, S - |j - i|) \quad (7)$$

where $\text{dist}(i, j)$ denotes the shortest circular distance between stations i and j .

In this formulation, the daytime/nighttime coefficients of 1.2 and 0.8 represent relative demand differences between active service periods and low-demand periods. The peak demand and school demand terms are used to generate additional passenger flows during commuting and school-related periods, respectively. The distance decay function is introduced to reduce the probability of unrealistic long-distance OD trips. Therefore, the OD matrix used in this study is a synthetic OD distribution generated from time-of-day factors, station functional assumptions, directional commuting patterns, and the distance decay rule, rather than an AFC/APC-calibrated OD matrix.

To more accurately capture the temporal patterns characteristic of actual urban transit travel, this study partitions intraday demand into three distinct periods, morning peak, evening peak, and school commute, and assigns a dominant travel direction to each. The morning peak is defined as $t \in [7:00, 9:00)$, during which demand predominantly comprises commuter flows from residential areas to commercial districts and schools. The evening peak is defined as $t \in [17:00, 19:00)$, during which the directional pattern reverses, with demand consisting primarily of return flows from commercial districts to residential areas. Additionally, a school period is defined as $t \in [8:00, 16:00)$ to account for concentrated travel demand associated with school schedules. This formulation effectively captures the characteristic peak to off-peak demand differentials and directional asymmetries without introducing undue complexity, thereby furnishing representative demand inputs for the subsequent assessments of control strategy stability and robustness under time-varying perturbations.

Figure 1 depicts the spatial distribution of passenger OD demand together with the associated marginal statistics for the complete route, which comprises $S = 20$ stations. The central matrix employs stations S_1 through S_{20} as both row and column indices, with color intensity encoding the simulated trip volume from origin station S_i to destination station S_j . Diagonal entries are left blank, signifying that trips originating and terminating at the same station are excluded from consideration. A pronounced band of elevated intensity is evident adjacent to the diagonal, indicating that short-distance trips predominate, whereas long-distance trips spanning multiple stations occur infrequently, a pattern consistent with the distance decay principle. The upper bar chart summarizes the total number of alighting passengers at each station, while the right-hand bar chart summarizes the total number of boarding passengers. The marginal distributions reveal pronounced heterogeneity in boarding and alighting demand across stations, reflecting the spatial variability in station functionality and passenger flow distribution.

2.2. Public Transportation Dispatch Model

This study formulates a unified discrete-time state transition framework to characterize the bus system's scheduling dynamics, with the longitudinal acceleration $a_k(t)$ of any vehicle k serving as the control variable. Notably, this formulation does not explicitly distinguish between inter-stop speed regulation and dwell time management at stops; rather, the system state is fully parameterized by vehicle acceleration alone, with station holding behavior realized by simply setting $a_k(t) = 0$. For any vehicle B_k , its position and velocity at time t are denoted by $p_k(t)$ and $v_k(t)$, respectively. The vehicle state evolves according to the following discrete state equation:

$$\begin{aligned} p_k(t+1) &= p_k(t) + v_k(t) \cdot \Delta t \pmod{L} \\ v_k(t+1) &= \Pi_{[v_{\min}, v_{\max}]}(v_k(t) + a_k(t) \cdot \Delta t) \end{aligned} \quad (8)$$

where $\text{mod } L$ denotes the modulo operation that enforces the closed-loop position wrap around, thereby ensuring that $p_k(t) \in [0, L)$. The operator $\Pi_{[v_{\min}, v_{\max}]}$ represents the interval projection, which constrains the velocity to the admissible range $[v_{\min}, v_{\max}]$, reflecting

both roadway speed restrictions and vehicular dynamic limitations. Furthermore, to preclude unrealistic abrupt acceleration or deceleration maneuvers and to satisfy passenger comfort requirements, acceleration is subject to the following bounds:

$$a_k(t) \in [a_{\min}, a_{\max}] \quad (9)$$

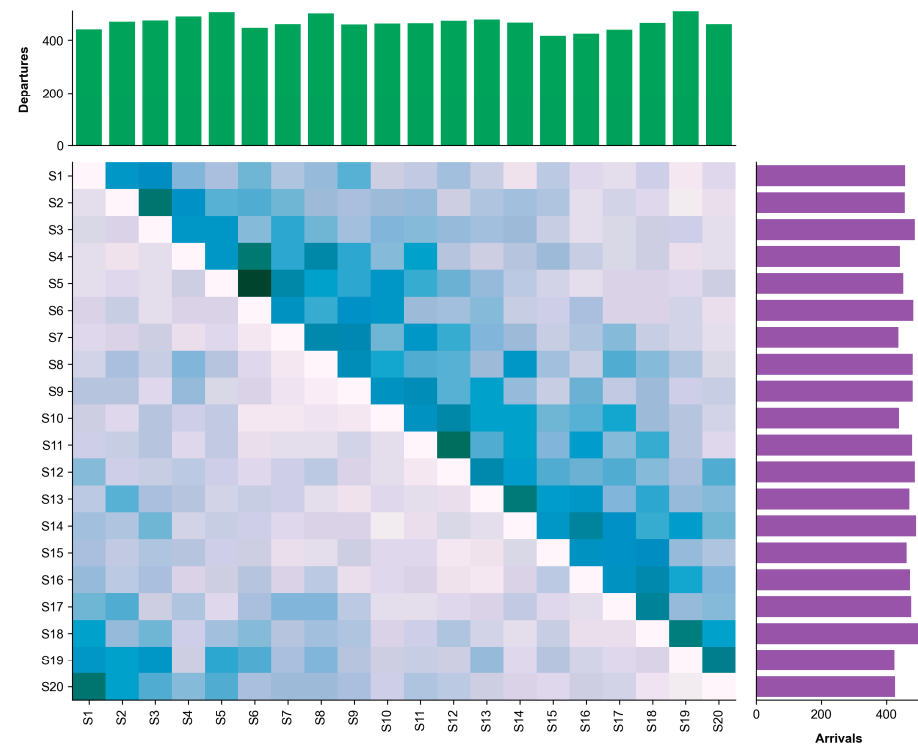


Figure 1. Origin destination demand heatmap with boarding/alighting marginals.

To satisfy the safety and comfort requirements inherent in bus operations, this study further incorporates a car-following safety constraint derived from a discrete-time longitudinal kinematic model, thereby suppressing unreasonable or excessive acceleration maneuvers. In addition to the global maximum speed $v_{\max}$, the admissible upper bound of a vehicle's velocity at time t is further constrained by a local safe speed, denoted by $v_{\max}^{\text{local}}(t)$. First, the circular headway between vehicle B_k and its immediate predecessor B_{k-1} is defined as follows:

$$\Delta p_k(t) = (p_{k+1}(t) - p_k(t)) \bmod L \quad (10)$$

where $p_k(t)$ denotes the position of the vehicle along the closed circuit, and L represents the total circuit length.

To preclude rear-end collisions and to accommodate a safety margin during braking maneuvers, the upper bound of the local safe speed, accounting for the minimum safe distance d_{safe} , is defined as follows:

$$v_{\max}^{\text{local}}(t) = \min \left(v_{\max}, \sqrt{2a_{\text{dec}} \cdot \max(\Delta p_k(t) - d_{\text{safe}}, 0)} \right) \quad (11)$$

where $a_{\text{dec}} = 1 \text{ m/s}^2$ denotes the smoothed deceleration profile, which characterizes comfortable braking performance. By actively reducing the admissible upper bound on velocity as the vehicle approaches its predecessor and the effective following distance diminishes, this constraint compels the agent to explicitly account for car-following safety

during the learning of control policies, thereby yielding control behaviors and velocity trajectories that more faithfully reflect real-world bus operating conditions.

Upon vehicle B_k arriving at station s_i , the station service phase is initiated, whereupon the system sequentially executes passenger alighting, passenger boarding, and dwell time updating procedures.

Passenger boarding model: For any destination station $S_i (j \neq i)$, the number of newly arriving passengers traveling from S_j to S_i within the time interval $[t, t + \Delta t]$ is denoted by $A_{i,j}(t)$.

Passenger alighting model: Suppose vehicle B_k arrives at station S_i at time t . The number of alighting passengers, denoted by $o_i(t)$, is determined by the number of onboard passengers whose intended destination is S_i :

$$o_i^{(k)}(t) = n_i^{(k)}(t) \quad (12)$$

where $n_i^{(k)}(t)$ denotes the number of passengers on vehicle B_k at time t whose destination is stop S_i . After alighting is completed, the onboard passenger load decreases accordingly, thereby freeing capacity for the subsequent boarding process.

Dwell time model: To account for the combined influence of passenger boarding and alighting on station dwell time, this study adopts the following formulation:

$$\tau_{\text{stop}} = \max(2 \cdot \sum_j A_{i,j}(t), 4 \cdot o_i(t)) + \tau_{\text{hold}} \quad (13)$$

where $\sum_j A_{i,j}(t)$ represents the total number of passengers alighting at stop S_i , and $o_i(t)$ denotes the total number of passengers boarding at the same stop. Specifically, τ_{hold} refers to the duration during which the vehicle at the stop is controlled by the dispatching system with its acceleration maintained at zero. The model assumes that the average passenger spends approximately 2 s on boarding and 4 s on alighting. Since these two processes can occur in parallel or alternately, the dwell time is determined by the dominant one among the two. This dwell time is subsequently incorporated into the vehicle's total operating time and the evolution process of headway.

2.3. Vehicle Longitudinal Dynamic Model

This study centers on bus operational stability and headway pattern evolution. Accordingly, a discrete-time longitudinal kinematic model is adopted to characterize the motion dynamics of each electric bus traversing a closed circuit, wherein vehicle steering dynamics and the ancillary energy consumption attributable to the air conditioning system are disregarded. Based on the longitudinal kinematics and resistance formulations, the traction or braking power demanded at the drive axle is expressed as follows:

$$T_{\text{axle}} = \frac{R_w}{i} \left(C_f mg + \frac{1}{2} \rho C_d A v^2 + ma \right) \quad (14)$$

where i denotes the transmission ratio, a represents the longitudinal acceleration, m is the equivalent vehicle mass (defined as the sum of the unladen vehicle mass and the passenger payload), R_w denotes the wheel rolling radius, C_f is the rolling resistance coefficient, ρ represents air density, and $C_d A$ corresponds to the product of the aerodynamic drag coefficient and the frontal area.

For battery electric buses, the drive torque at the drive axle is supplied by the traction motor, i.e., $T_{\text{mot}} = T_{\text{axle}}$. The motor rotational speed is directly proportional to the vehicle velocity and is given by $W_{\text{mot}} = v \frac{i}{R_w}$. The motor efficiency $\eta_{\text{mot}} = f(T_{\text{mot}}, W_{\text{mot}})$ is determined from the motor efficiency map, as illustrated in the accompanying Figure 2.

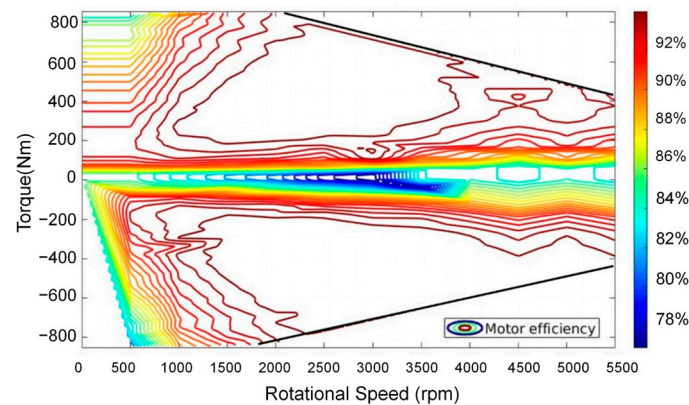


Figure 2. MAP chart of motor efficiency for all-electric buses.

Furthermore, this study adopts an equivalent circuit model (ECM) for lithium iron phosphate (LFP) batteries, wherein the state of charge (SoC) $z \in [0, 1]$ serves as the core state variable. This formulation allows the open-circuit voltage and equivalent internal resistance to vary with the SoC, thereby capturing the performance variations in the battery across different SoC regimes with greater fidelity. An LFP equivalent circuit model and terminal voltage/current calculations are employed. Both the open-circuit voltage and the internal resistance of the battery are modeled as piecewise linear functions of the SoC:

$$\begin{aligned} R_{\text{dis}}(z) &= f_R^{\text{dis}}(z) \\ R_{\text{chg}}(z) &= f_R^{\text{chg}}(z) \\ V_{\text{oc}}(z) &= f_v(z) \end{aligned} \quad (15)$$

where $z \in [0, 1]$ denotes the battery SoC; $V_{\text{oc}}(z)$ represents the open-circuit voltage; and $R_{\text{dis}}(z)$ and $R_{\text{chg}}(z)$ correspond to the equivalent internal resistances during discharging and charging (regenerative braking), respectively. In the simulation, the circuit toggles between R_{dis} and R_{chg} according to the sign of the battery power or the direction of the battery current, thereby capturing the influence of the disparity in internal resistance between traction and energy recuperation modes on both the terminal voltage and the power loss. The battery-side power output P_{batt} is then expressed as follows:

$$P_{\text{batt}} = \frac{T_{\text{mot}} W_{\text{mot}}}{\eta_{\text{mot}}} = \eta_{\text{drv}} \cdot (V_{\text{oc}} I_{\text{batt}} - R I_{\text{batt}}^2) \quad (16)$$

where I_{batt} denotes the battery current, and $R(z) \in \{R_{\text{dis}}(z), R_{\text{chg}}(z)\}$ is selected according to the charging or discharging state. The parameter η_{drv} represents the energy conversion efficiency, which is assigned a value of $\eta_{\text{drv}} = 1$ under traction conditions and $\eta_{\text{drv}} = 0.98$ during regenerative braking. From the above expression, $I_{\text{batt}}(t)$ can be determined, and the terminal voltage may subsequently be derived as follows:

In terms of SoC evolution and energy accumulation, the battery state of charge is updated via discrete-time coulomb counting as follows:

$$z(t+1) = z(t) - \frac{I_{\text{batt}}(t) \cdot \Delta t}{3600 \cdot C_n} \quad (17)$$

where C_n denotes the nominal battery capacity, Δt represents the simulation time step, and the cumulative Ah-throughput quantifies the cycling load experienced by the battery:

$$\text{Ah}(t+1) = \text{Ah}(t) + \frac{|I_{\text{batt}}(t)| \Delta t}{3600} \quad (18)$$

To characterize the impact of battery degradation on vehicle energy availability and operational capability over extended simulation horizons, this study adopts a hybrid degradation model that integrates electrochemical mechanisms with empirical terms. Let SoH denote the ratio of the battery's available capacity to its initial capacity; the single-step degradation is then defined as follows:

$$\Delta SoH = (\alpha z + \beta) \exp\left(\frac{-31700 + 163.3C_{rate}}{RT}\right) \cdot Ah^{0.57} \quad (19)$$

where α, β denote undetermined coefficients; R represents the universal gas constant; T denotes the battery temperature; and C_{rate} signifies the rate index, which is computed according to $C_{rate} = \frac{|I_{batt}|}{400}$.

In this study, identical degradation coefficients and temperature settings are adopted for all compared control strategies to ensure a fair and consistent comparison. The Ah -throughput term is accumulated according to Equation (18), thereby representing the cumulative cycling load induced by traction and regenerative braking processes. Therefore, the adopted SoH degradation model is primarily used to evaluate the relative impact of different control strategies on battery health under identical simulation conditions.

Vehicle heterogeneity is defined as the time-varying agent-level operational heterogeneity that gradually emerges among bus agents during operation. All buses are assumed to have the same nominal vehicle parameters at initialization. However, due to stochastic passenger demand, different boarding and alighting processes, and different speed control histories, each bus gradually evolves into a distinct operating state. Therefore, the heterogeneity considered in this study is mainly reflected in passenger load, equivalent vehicle mass, battery SoC , battery SoH , battery internal resistance, motor operating efficiency, and effective acceleration response. The representation of vehicle heterogeneity is summarized in Table 2.

Table 2. Representation of agent-level vehicle heterogeneity.

Factor	Representation in This Paper	Main Effect
Passenger load	$m_p N_i^p(t)$	Dwell time and vehicle mass
Vehicle mass	$m_i(t) = m_0 + m_p N_i^p(t)$	Longitudinal dynamics and power demand
SoC	$SoC_i(t)$	Voltage, resistance, and available energy
SoH	$SoH_i(t)$	Battery aging and health-related reward
Battery internal resistance	$R_{dis}(SoC_i), R_{chg}(SoC_i)$	Current, power loss, and degradation
Motor efficiency	$\eta_{mot} = f(T_{mot}, W_{mot})$	Energy conversion efficiency
Acceleration performance	$a_i^{eff}(t)$ under dynamic and safety constraints	Speed evolution and feasible control

For bus agent i at time t , agent-level heterogeneity can be summarized as follows:

$$\bar{z}_i^{het}(t) = [m_p N_i^p(t), m_i(t), SoC_i(t), SoH_i(t), R_i(t), \eta_{mot}, a_i^{eff}(t)], \quad (20)$$

where $m_p N_i^p(t)$ denotes the onboard passenger load; $m_i(t)$ is the equivalent vehicle mass; $SoC_i(t)$ and $SoH_i(t)$ denote the battery state of charge and state of health, respectively; $R_i(t)$ is the charge/discharge internal resistance determined by the battery state and current direction; η_{mot} is the motor efficiency corresponding to the current torque–speed operating point; and $a_i^{eff}(t)$ represents the effective acceleration response under the combined

influence of mass, power demand, battery state, and safety constraints. Thus, even with identical nominal vehicle parameters, different bus agents exhibit heterogeneous dynamic and energy responses during operation.

2.4. Question Analysis

As can be seen from the above modeling, the core control variable in the bus scheduling problem can be defined as the longitudinal acceleration $a_k(t)$ of vehicle k at time t . This control variable directly determines the evolution of the vehicle's longitudinal motion state, including the dynamic updates of vehicle position $x_k(t)$, velocity $v_k(t)$, and arrival time $T_{k,i}$, and further affects the headway between adjacent vehicles at stop s_i . Therefore, $a_k(t)$ is essentially the key decision variable for regulating the operational rhythm of the route, suppressing headway imbalance, and avoiding the bus bunching phenomenon. Meanwhile, the longitudinal control input also alters the vehicle's traction demand and the operating point of the powertrain, thereby causing coupled variations in the battery output power $P_{b,k}(t)$, current $I_k(t)$, magnification $C_k(t)$, and health degradation rate $\text{SoH}_k(t)$. The above relationship indicates that the bus scheduling problem is essentially a multivariable, time-varying, and strongly coupled collaborative optimization problem jointly shaped by passenger demand, vehicle motion, and powertrain degradation, rather than a conventional single-layer scheduling problem concerned only with headway equilibrium. The coupling framework of the passenger demand, vehicle motion and powertrain in the bus system is presented in Figure 3.

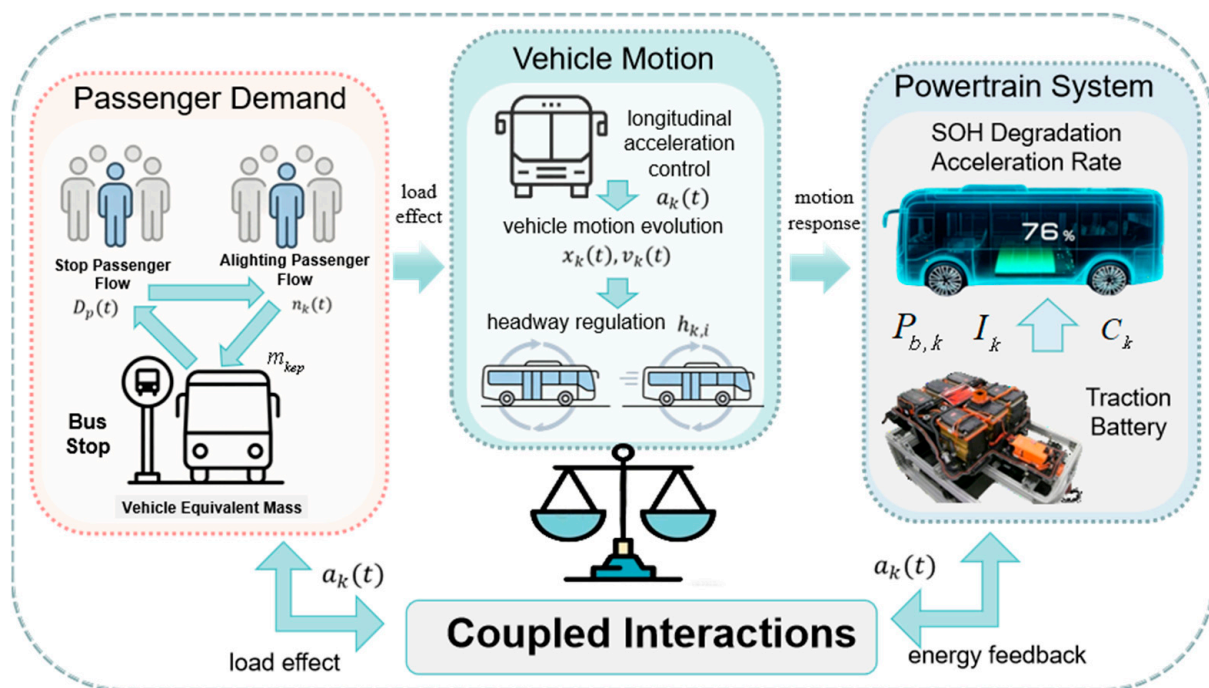


Figure 3. Coupling interaction framework diagram of passenger demand–vehicle motion–powertrain system for bus operations.

Conventional scheduling methodologies regard buses merely as undifferentiated rigid bodies in motion, thereby entirely disregarding the underlying vehicle dynamics and battery degradation phenomena. In essence, such an approach constitutes a superficial equilibrium that overlooks vehicle heterogeneity:

1. Such an approach proves incapable of accommodating variations in passenger load and the consequent divergence in dynamic performance attributable to battery aging.

The resulting mismatch between scheduling strategies and actual vehicle capabilities renders the optimization highly susceptible to entrapment in local optima.

2. By privileging superficial operational stability at the expense of operational efficiency, energy economy, and battery longevity, such methodologies prove inherently incapable of attaining global system optimization or meaningful multi-objective coordination.
3. An excessive emphasis on uniformity in vehicle spacing exemplified by the adoption of consistent headway as an optimization objective imposes unduly stringent constraints on the entirety of bus operations, thereby severely curtailing the latitude of each vehicle to select the optimal trajectory in accordance with its individual load and powertrain characteristics.

In summary, bus scheduling transcends mere traffic flow equalization; rather, it constitutes a multi-agent cooperative optimization problem that explicitly accounts for vehicle heterogeneity and jointly addresses passenger service quality alongside battery health preservation. Extant methodologies prove inadequate in reconciling system-level equilibrium with vehicle-level optimality, thereby underscoring the pressing need for a cooperative control framework capable of accommodating dynamic vehicle attributes, relaxing unduly rigid constraints, and attaining meaningful multi-objective coordination.

3. Multi-Agent Reinforcement Learning Framework

To address the aforementioned issues, this paper formulates the cooperative dynamic control problem for electric buses as a discrete-time multi-agent Markov game and proposes a Shared Multi-Agent Twin Delayed Deep Deterministic Policy Gradient algorithm (SMATD3). The proposed framework follows the centralized training and decentralized execution (CTDE) paradigm. During training, centralized critics use the global system state and joint actions to capture inter-vehicle interactions and alleviate the non-stationarity caused by multi-agent coupling. During execution, each bus independently determines its continuous control action based only on local observations, enabling distributed and real-time implementation. Unlike conventional Multi-Agent Deep Deterministic Policy Gradient (MADDPG) methods that assign separate actor networks to different agents, SMATD3 introduces a shared policy mechanism for isomorphic bus agents. Since buses share similar observation structures, action semantics, and control objectives, a shared actor network allows all agents to learn common decision-making knowledge from collective interaction data while still responding to their individual operating states. This design reduces redundant policy parameters, improves sample efficiency, and enhances policy generalization across vehicles. To further improve training stability in continuous action spaces, SMATD3 incorporates the main stabilization mechanisms of TD3, including twin critics, target policy smoothing, and delayed policy updates. These mechanisms help reduce value overestimation and policy oscillation. Therefore, the proposed SMATD3 framework is well suited for cooperative electric bus control under dynamic passenger demand, vehicle heterogeneity, and coupled energy health constraints. The framework of the SMATD3 algorithm is shown in Figure 4.

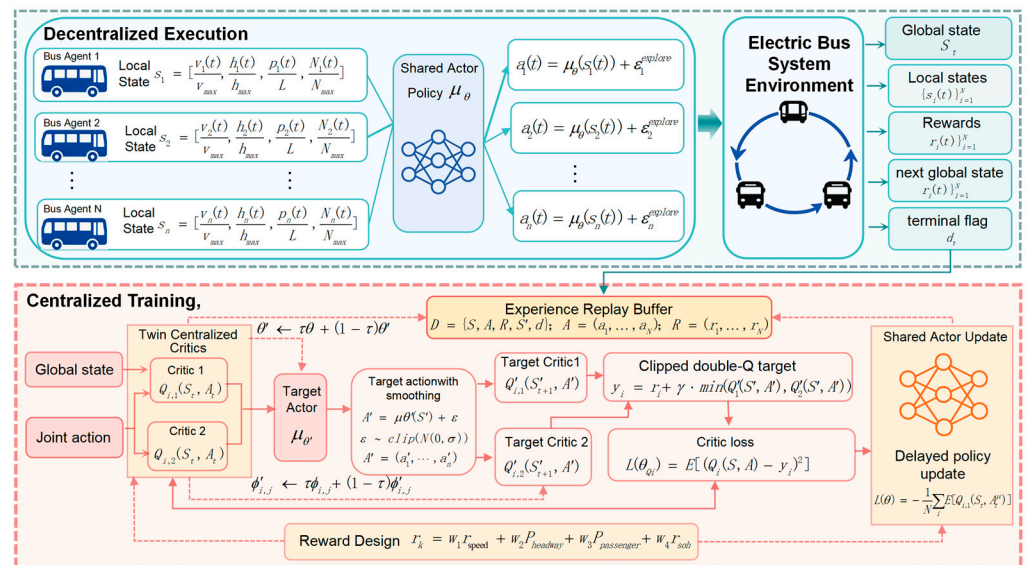


Figure 4. Framework of proposed SMATD3 algorithm for cooperative electric bus control.

3.1. State Space and Action Space

During the distributed execution phase, each bus B_k makes decisions based solely on locally available information. In this paper, the local state vector of agent k is defined in a four-dimensional normalized form:

$$s_k = \left[\frac{v_k}{v_{max}}, \frac{h_k}{h_{max}}, \frac{p_k}{L}, \frac{N_k}{N_{max}} \right]^T \quad (21)$$

where $v_k(t)$ and $p_k(t)$ denote the vehicle's speed and position on the closed-loop route, respectively; $v_{\max}$ is the global maximum speed; and L is the total route length. To characterize the evolution of the running intervals between vehicles, the spatial interval between vehicle k and the vehicle ahead is defined as

$$h_k = (p_{k+1} - p_k) \bmod L \quad (22)$$

where the modulo operation with respect to L ensures that the shortest forward distance remains non-negative on a closed-loop circuit. In this paper, $h_{\max} = 6000$ m is used for normalization.

In addition, to reflect the service pressure at each stop, we define $N_k(t)$ as the total number of passengers waiting to be served at the next stop for vehicle k :

$$N_k = \sum_j n_{i+1,j}(\mathbf{t}) \quad (23)$$

where $i + 1$ denotes the index of the next stop for vehicle k , and $n_{i+1,j}(t)$ represents the number of passengers waiting at stop s_{i+1} with destination s_j . In this paper, $N_{\max} = 50$ is adopted as the normalization upper bound for this term.

Global joint state: During the centralized training phase, to mitigate the non-stationarity of the multi-agent environment, the critic network takes the global state as input. The system’s global state is constructed by concatenating the local states of all vehicles:

$$S(t) = [s_1^T(t), s_2^T(t), \dots, s_5^T(t)]^\top. \quad (24)$$

Define the kinematic space as the continuous acceleration of each vehicle:

$$a_k(t) \in [a_{min}, a_{max}]. \quad (25)$$

where $a_{min} = -1 \text{ m/s}^2$, $a_{max} = 1 \text{ m/s}^2$. The action updates the vehicle's velocity using discrete-time dynamics and projects it onto the feasible region by incorporating the velocity upper bound and the following safety constraint, thereby ensuring motion feasibility and operational safety. Specifically, when the vehicle is at a bus stop and $v = 0$, it will only start moving if $a > 0.5$. The advantage of this approach is that it does not introduce an additional control variable for holding time into the control strategy; instead, it integrates inter-stop and intra-stop control through a unified acceleration mechanism.

3.2. Reward Function

From a transportation engineering perspective, the reward function is designed to reflect four major operational objectives: route operating efficiency, service regularity, passenger service quality, and battery health protection. The speed reward encourages buses to maintain efficient inter-stop movement and avoid overly conservative driving. The headway-related penalty is introduced to suppress bus bunching and excessive service gaps, because headway regularity is a key indicator of bus service reliability. The passenger-related penalty reflects waiting pressure at stops and is used to improve passenger service quality. The SoH-related penalty accounts for the lifecycle cost of electric bus operation, because aggressive speed regulation may increase battery current stress and accelerate battery degradation.

The reward function is formulated as a weighted scalarization of multiple objectives, which is a common treatment in multi-objective reinforcement learning. In this study, the reward weights do not represent physical parameters with direct engineering units. Instead, they encode the relative priority among operating efficiency, service regularity, passenger service, and battery health.

To balance bus operational stability, efficiency, passenger service quality, and battery health protection within a multi-agent reinforcement learning framework, this paper constructs a weighted multi-objective reward function for each agent k :

$$r_k = \omega_1 r_{\text{speed}} + \omega_2 r_{\text{headway}} + \omega_3 r_{\text{passenger}} + \omega_4 r_{\text{soh}} \quad (26)$$

The weights in Equation (25) are determined according to three principles. First, each reward or penalty component is normalized or scaled to a comparable numerical range so that no term dominates the learning process only because of its numerical magnitude. Second, the weights reflect transportation engineering priorities: headway stability and passenger service quality are treated as primary public transport objectives, while speed efficiency and battery health protection are used to improve route turnover efficiency and reduce long-term vehicle operation cost. Third, the weights are selected to maintain training stability and balanced performance across average waiting time, average travel time, average holding delay, average speed, energy efficiency, and cumulative battery degradation.

Operational efficiency incentive: This incentive encourages buses to travel at the maximum feasible speed between stops, avoiding overly conservative driving that would increase travel time, thus improving overall operational efficiency. The maximum speed is influenced by traffic congestion and speed limits. In this paper, we assume free-flow traffic conditions on bus routes; therefore, the maximum speed depends solely on the speed limit, which is set to 14.5 m/s for urban areas and 10 m/s for suburban areas:

$$r_{\text{speed}}(t) = \left(\frac{v_k}{v_{\text{max}}} \right)^2 \quad (27)$$

Bus tailgating penalty: To mitigate issues such as “bus tailgating” caused by uneven headways in the bus system, a quadratic penalty constraint is applied to deviations from the

target headway $h' = 4000$ m. Unlike the virtual queue method, this paper imposes penalties only when the inter-vehicle distance exceeds 6000 m or falls below 2000 m. For the 20 km circular route with five buses, the nominal equal spacing is $L/N = 4000$ m. Therefore, the lower and upper thresholds of 2000 m and 6000 m correspond to approximately 50% and 150% of the nominal spacing, respectively. This sparse penalty structure is designed to prevent severe bus bunching and excessive service gaps without forcing all buses to maintain identical spacing at every time step. Within the remaining range, bus spacing is not treated as an optimization objective; instead, actual system objectives such as system efficiency, service intelligence and service capacity are used as rewards. This approach restores the system's true optimization objectives rather than merely ensuring the convergence of the control strategy.

Given the circular route length and fleet size specified in Table 3, the benchmark spatial headway is determined by uniform fleet distribution, i.e., $d' = L/|B| = 20,000/5 = 4000$ m. The penalty-triggering bounds are set to $0.5d' = 2000$ m and $1.5d' = 6000$ m, forming a $\pm 50\%$ tolerance band around uniform spacing: a spacing below 2000 m indicates excessive vehicle convergence and an increased risk of bunching, whereas a spacing above 6000 m indicates a substantial service gap. Within this band, no spacing penalty is imposed, allowing the agents to optimize passenger waiting time, operational efficiency, and battery health rather than unnecessarily tracking perfectly equal spacing. In temporal terms, the associated nominal service headway is $H' = d'/\bar{v} = L/(|B|\bar{v})$, where $\bar{v}$ denotes the route-level average operating speed; thus, the 4000 m spatial benchmark is the distance-domain representation of an evenly distributed service headway.

$$r_{\text{headway}}(t) = \begin{cases} -\left(\frac{h_k}{h'} - 1\right)^2, & h_k > 6000 \text{ or } h_k < 2000 \\ 0, & 6000 \geq h_k \geq 2000 \end{cases} \quad (28)$$

Table 3. Training hyperparameters.

Parameter	Value
Replay buffer size	1×10^6
Batch size	1024
Actor learning rate	1×10^{-4}
Critic learning rate	1×10^{-3}
Discount factor (γ)	0.99
Target update rate (τ)	0.005
Exploration noise (σ_{explore})	$0.5 \rightarrow 0.1$
Target noise (σ_{target})	0.2
Noise clip range	$[-0.5, 0.5]$
Training steps	500,000
Update frequency	Every 100 steps

Service quality penalty term: This term aims to reduce passenger waiting times in the bus system, alleviate congestion and queuing pressure at stops, and improve passenger experience. In this paper, the number of passengers waiting at the next stop for each bus $N_k(t)$ is adopted as the instantaneous penalty term for each agent, rather than the total number of passengers at all stops $n_{k,i}(t)$. The advantage of this approach is that it establishes a more direct mapping between each agent's control action and its reward, ensuring that the training process effectively guides policy parameter updates. Furthermore, leveraging the near-equilibrium spatial distribution characteristic of multiple buses in the public transportation system, the overall passenger state is approximated through sampling,

ultimately enabling multi-agent collaboration via the CTDE mechanism. A piecewise penalty function is implemented using the floor function to enhance robustness:

$$r_{\text{passenger}}(t) = - \left\lfloor \frac{N_k}{20} \right\rfloor \quad (29)$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

Battery health protection: Operational control not only affects changes in the system's external state but also influences the lifespan evolution of each bus's powertrain. Furthermore, bus operators are encouraged to ensure uniform performance degradation across all buses, thereby enabling the replacement of vehicles from the same batch. As the component most critical to the lifespan of new energy vehicles, the power battery is the primary focus of the incentive design in this section. $\Delta\text{SoH}_k(t)$ represents the incremental degradation of battery health within the current time step. A scale factor of 10^7 is applied to ensure that the degradation cost remains numerically comparable to the other incentive terms, preventing it from being overshadowed.

$$r_{\text{soh}} = -\Delta\text{SoH} \cdot 10^7 \quad (30)$$

3.3. Shared Policy Multi-Agent Twin Delayed Deep Deterministic Policy Gradient Algorithm

In multi-agent cooperative control tasks, traditional methods typically maintain a separate set of actor network parameters for each agent. However, when all agents perform isomorphic decision-making tasks sharing consistent action semantics, similar observation structures, and identical control objectives, independent learning not only leads to redundant parameter scales and reduces sample efficiency but also exacerbates the non-stationarity issues inherent in multi-agent training. To address this problem, this paper introduces a shared policy mechanism into the SMATD3 framework and proposes the SMATD3 algorithm. This algorithm enables all agents to share a single actor network and uses interaction samples collected from all vehicles to jointly train the shared policy. As a result, while preserving each agent's ability to perceive individual state differences, it extracts common control knowledge across agents, thereby improving parameter utilization efficiency and policy generalization capabilities.

To clarify the logical relationship among decentralized execution, reward feedback, centralized training, and shared policy deployment, the overall workflow of the proposed SMATD3 algorithm is first illustrated in Figure 5. The workflow contains four main stages. During decentralized execution, each bus agent observes only its local state and generates a continuous acceleration action through the shared actor network. The environment then produces the state transition and multi-objective reward, which jointly considers operational efficiency, passenger service quality, sparse spacing regulation, and battery health preservation. The collected transitions are stored in the experience replay buffer and further used for centralized training. Finally, policy gradients from all bus agents are aggregated to update the shared actor, which is then deployed to all buses for decentralized real-time control.

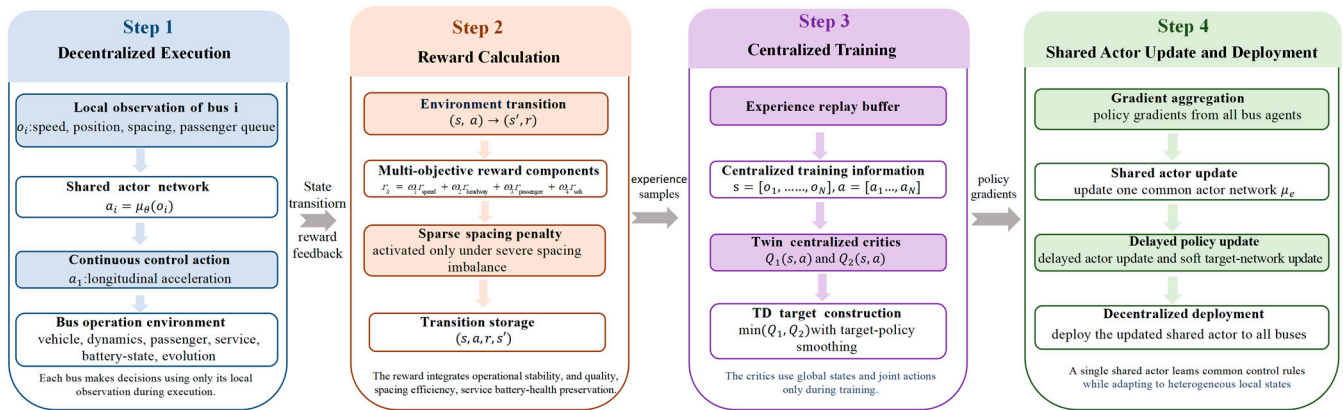


Figure 5. Overall workflow of proposed SMATD3 algorithm for cooperative electric bus control.

3.3.1. Shared Policy Mechanism

Suppose the system consists of N isomorphic agents, and let the local observation of the k -th agent at time t be o_k^t . All agents share the same deterministic policy network $\pi_\theta(\cdot)$, whose action outputs can be uniformly expressed as follows:

$$a_k^t = \pi_\theta(o_k^t), k = 1, 2, \dots, N \quad (31)$$

where θ denotes the network parameters of the shared actor. Since each bus agent performs an isomorphic control task, the samples generated by different vehicles can be stored uniformly in a centralized experience replay pool. To maintain consistency with the input format of the centralized critic, this paper adopts the following joint sample representation:

$$\mathcal{D} = \{(s_t, o_t, a_t, r_t, s_{t+1}, o_{t+1}, d_t)\}_{t=1}^T \quad (32)$$

where

$$o_t = (o_1^t, \dots, o_N^t), a_t = (a_1^t, \dots, a_N^t), r_t = (r_1^t, \dots, r_N^t)$$

where s_t denotes the global state, and d_t is the episode termination flag. This shared experience mechanism enables the interaction data from any single bus to be used not only for updating its own decision-making rules but also for optimizing the actor network shared by all agents, thereby significantly improving sample efficiency and policy generalization capabilities.

3.3.2. Dual-Critic Structure and Centralized Training Mechanism

To explicitly model the coupling relationships among multiple agents, this paper retains the CTDE architecture of SMATD3 and configures two centralized critic networks $Q_{\phi_1^1}$ and $Q_{\phi_1^2}$ for each agent. The i -th critic takes the global state s_t and the joint action a_t as input to evaluate the long-term cumulative reward for the corresponding agent. To mitigate the overestimation bias in the action value function, this paper adopts the clipped double-Q objective of TD3. First, the target action for the next time step is constructed based on the target actor network μ_{θ^-} :

$$a_j' = \text{clip}(\mu_{\theta^-}(o_j') + \epsilon, a_{\min}, a_{\max}), j = 1, \dots, N \quad (33)$$

$$\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma^2), -c, c)$$

where θ^- denotes the target actor parameters, σ is the standard deviation of the target smoothing noise, and c is the noise clipping threshold. This yields the joint target action:

$$a' = (a_1', a_2', \dots, a_N') \quad (34)$$

The TD target value for the i -th agent is defined as

$$y_i = r_i + \gamma(1 - d) \min_{m \in \{1,2\}} Q_{\phi_i^{m-}}(s', a') \quad (35)$$

where γ is the discount factor, r_i is the immediate reward of the i -th agent, and ϕ_i^{m-} denotes the parameters of the target critic network. Thus, the two critic networks of the i -th agent are updated by minimizing the following loss:

$$\mathcal{L}_{\text{critic}}^{(m,i)}(\phi_i^m) = \mathbb{E}_{(s,o,a,r,s',o',d) \sim \mathcal{D}} \left[\left(Q_{\phi_i^m}(s, a) - y_i \right)^2 \right], m \in \{1, 2\}. \quad (36)$$

The aforementioned dual-critic minimum mechanism is used solely to construct the target value, thereby reducing the systematic overestimation bias of the value function.

3.3.3. Optimizing Shared Actor Updates and Delay Strategies

Under the shared policy mechanism, all agents share the same actor network $\mu_\theta(\cdot)$. Therefore, the actor's update objective is no longer tailored to a single agent but rather involves joint optimization based on the sample contributions of all agents. Following the standard TD3 approach, the actor update uses only the first critic network to provide policy gradients, and its loss function can be written as

$$\mathcal{L}_{\text{actor}}(\theta) = -\mathbb{E}_{(s,o) \sim \mathcal{D}} \left[\frac{1}{N} \sum_{i=1}^N Q_{\phi_i^1}(s, \mu_\theta(o_1), \mu_\theta(o_2), \dots, \mu_\theta(o_N)) \right] \quad (37)$$

This loss function learns a shared control policy applicable to all buses by averaging the outputs of all agent critics.

To reduce gradient noise and training oscillations caused by overly frequent actor updates, this paper adopts the delayed update mechanism of TD3. Specifically, the critic network is updated at each training step, while the shared actor and its target network are updated only every d_{delay} steps. When the conditions for delayed updating are met, the target network parameters are updated using a soft update method:

$$\begin{aligned} \theta^- &= \tau\theta + (1 - \tau)\theta^- \\ \phi_i^{m-} &\leftarrow \tau\phi_i^m + (1 - \tau)\phi_i^{m-}, m \in \{1, 2\} \end{aligned} \quad (38)$$

where τ is the soft update coefficient. This mechanism enhances the stability of policy updates while ensuring that the critic networks have converged sufficiently.

3.3.4. Explore Strategies

During the training phase, to enhance the policy's exploration capabilities, this paper adds Gaussian noise to the actor's output actions:

$$\begin{aligned} \tilde{a}_i^t &= \text{clip}(\mu_\theta(o_i^t) + \xi_t, a_{\min}, a_{\max}) \\ \xi_t &\sim \mathcal{N}(0, \sigma_{\text{explore}}^2) \end{aligned} \quad (39)$$

where σ_{explore} gradually decays over the course of training. This design helps expand the exploration range in the early stages of training while gradually enhancing policy convergence in the later stages.

The main hyperparameters of the SMATD3 algorithm are shown in Table 4. By employing a large capacity experience replay buffer, larger batch sizes, and the target policy smoothing and delayed update mechanisms from TD3, the algorithm significantly improves training stability while ensuring efficient sample utilization.

Table 4. Key simulation data.

Parameter Category	Symbol	Value/Range	Description
Route Length	L	20,000 m	Medium-scale circular/trunk bus corridor
Number of Stations	S	20	Inter-station distance 1000 m
Number of Buses	B	5	
Simulation Duration	T_{sim}	16 h	Main daily service span, 06:00–22:00
Time Step	Δ_t	1 s	High-resolution state update
Traffic Condition	—	Free flow	Used to isolate the effect of bus control strategy
Maximum Speed	V_{max}	60/3.6 m/s	Approximately 60 km/h
Battery Capacity	C_n	600 Ah	Lithium iron phosphate
Passenger Mass	m_p	65 kg/person	Used for mass calculation

During the evaluation phase, we first load the optimal policy parameters θ^* obtained from training and initialize the environment under a fixed test configuration. Subsequently, under conditions free of exploration noise, agent i generates an action based on its local observation s_i using the deterministic policy. The actions of all agents form a joint action a and are applied to the environment simultaneously. The environment then returns to the next state and the termination flag d . During the evaluation process, complete trajectory data are recorded. If $d = 1$, the round is terminated early. Finally, performance metrics such as AWT, ATT, AOD, and AHD are computed from trajectory statistics to quantify the control effectiveness and stability of the method in the test scenario. The specific implementation procedure is summarized in Algorithm 1, and the pseudocode is presented as follows:

Algorithm 1 Policy Evaluation

```

1: Load trained policy parameters  $\theta^*$ 
2: Initialize environment with test configuration
3:  $S \leftarrow \text{env.reset}(\text{train} = \text{False})$ 
4: for  $t = 1$  to  $T_{\text{eval}}$  do
5:   for each agent  $i \in \{1, \dots, N\}$  do
6:      $a_i = \pi_{\theta}(o_i), i = 1, \dots, N.$ 
7:   end for
8:   Execute the joint action  $a$  in the environment
9:   Observe reward  $r$ , next state  $S'$ , and terminal flag  $d$ 
10:  Record trajectory data
11:  if  $d$  is True then
12:    break
13:  end if
14:   $S \leftarrow S'$ 
15: end for
16: Compute performance metrics (AWT, ATT, AOD, AHD)

```

In summary, SMATD3 builds on the dual critic, target policy smoothing, and delayed update mechanisms of SMATD3 by introducing a shared policy prior, which allows all bus agents to share the same actor network and jointly utilize interaction samples for training. This design aligns with the fundamental characteristics of closed-loop bus scheduling homogeneous tasks but heterogeneous states and effectively improves sample efficiency, training stability, and policy generalization performance.

4. Simulation Experiments and Analysis of Results

4.1. Simulation Environment Setup

To validate the effectiveness of the proposed SMATD3 control strategy, the simulation is conducted on a 20 km closed-loop bus route. The route is uniformly divided into 20 stops, with an average inter-stop spacing of 1 km. The system comprises five electric buses operating from 06:00 to 22:00, corresponding to a 16 h simulation horizon. The simulation time step is set to 1 s.

To clarify the representativeness of the simulation scenario, the parameter settings are designed to describe a medium-scale urban trunk or circular bus service rather than a direct replication of a specific real-world route. The 20 km closed-loop route represents a complete circular service corridor and avoids terminal boundary effects, thereby enabling continuous vehicle circulation and repeated inter-vehicle interactions. The 20 stops correspond to an average stop spacing of 1 km. This spacing is longer than that of dense local bus services, but it is suitable for trunk, arterial, rapid, or suburban–urban connecting bus services where operational efficiency and relatively long inter-stop movements are emphasized. Therefore, the route setting provides a simplified but representative benchmark for studying bus bunching, headway evolution, and cooperative speed control.

The fleet size of five buses corresponds to a frequent-service scenario on the 20 km loop. Under the simulated operating speeds, this configuration produces headways on the order of several minutes to approximately ten minutes, which is consistent with high-frequency urban bus services. The 16 h operation period from 06:00 to 22:00 covers the main daily service span of urban bus systems, including morning peak, evening peak, school-related demand periods, and off-peak intervals. Therefore, the simulation captures the main intraday passenger demand variations that may trigger dwell time fluctuations and bus bunching. The main parameters used in the simulation are listed in Table 3.

4.2. Comparative Control Strategy Settings

To systematically evaluate the effectiveness and advantages of the proposed SMATD3-based cooperative control method, this paper establishes and compares the following three representative control strategies:

No Control Strategy: Under this baseline scenario, each bus operates independently without any coordinated control. On inter-stop segments, buses travel at the maximum speed permitted by road regulations. After completing passenger boarding and alighting, each bus departs immediately without any additional holding time and continues along the route.

Holding Strategy: Buses regulate their departure times by imposing additional holding time at stops, thereby reducing fluctuations in inter-vehicle headways. This strategy represents a typical stop-based control method and generally does not involve explicit speed regulation on inter-stop segments. Therefore, buses are assumed to operate at the maximum allowable speed under the prevailing road conditions between stops.

SpaceHeadway Control: This strategy employs a virtual-queue-based method to regulate bus speeds on inter-stop segments, with the objective of maintaining a nearly uniform spatial spacing between adjacent vehicles.

4.3. Evaluation Metrics

To comprehensively evaluate the overall performance of different control strategies in the electric bus system, this study establishes an evaluation framework covering operational efficiency, passenger service quality, vehicle operation stability, energy efficiency, and battery health. The definitions of the selected indicators are presented as follows.

4.3.1. Average Holding Delay (AHD)

Average holding delay is used to quantify the additional stop-related waiting cost introduced by the control strategy.

Let the actual dwell time of vehicle i at the k stop event be $t_{i,k}^{dwell}$ and the baseline dwell time under uncontrolled conditions be $t_{i,k}^{base}$. Then, the additional dwell time is given by the following:

$$h_{i,k} = \max(0, t_{i,k}^{dwell} - t_{i,k}^{base}) \quad (40)$$

If there are N vehicle trips (or vehicles) in total, the average total stop delay per vehicle, calculated on a trip basis, is given by the following:

$$AHD = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{K_i} h_{i,k} \quad (41)$$

where K_i denotes the number of stops in trip i .

4.3.2. Average Waiting Time (AWT)

Average waiting time is used to measure the average time cost incurred by passengers at a stop from arrival to boarding.

Suppose there are N_p passengers in the evaluation interval, and the arrival time of passenger p is t_p^{arr} , while the actual boarding time is t_p^{board} . Then

$$AWT = \frac{1}{N_p} \sum_{p=1}^{N_p} (t_p^{board} - t_p^{arr}) \quad (42)$$

4.3.3. Average Travel Time (ATT)

Average travel time is used to measure the average time required for a vehicle to complete one loop.

Suppose N trip samples are collected within the evaluation interval, and the start and end times of the i trip are t_i^{start} and t_i^{end} , respectively. Then

$$ATT = \frac{1}{N} \sum_{i=1}^N (t_i^{end} - t_i^{start}) \quad (43)$$

4.3.4. Vehicle Average Speed (VAS)

Let the travel distance of the i trip be d_i and the corresponding travel time be t_i^{drive} . Then, the vehicle average speed is defined as

$$VAS = \frac{1}{N} \sum_{i=1}^N \frac{d_i}{t_i^{drive}} \quad (44)$$

4.3.5. Vehicle Energy Efficiency (VEE)

Let the energy consumption of the i trip be E_i and the travel distance be d_i . Then, the vehicle energy efficiency is defined as

$$VEE_{eff} = \frac{\sum_{i=1}^N d_i}{\sum_{i=1}^N E_i} \quad (45)$$

4.3.6. Battery Degradation (BD)

Battery degradation is used to quantify the long-term impact of operational control on battery life cost. In this paper, the SoH is used to characterize the battery health state. Let

the single-step SoH degradation of vehicle i within the evaluation interval be $\Delta\text{SOH}_i(t)$. Then, the cumulative degradation can be defined as

$$BD = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \Delta\text{SOH}_i(t) \quad (46)$$

where N denotes the number of vehicles, and T denotes the evaluation horizon. Since all control strategies are evaluated under the same battery degradation model and parameter settings, BD is used as a relative indicator for comparing their influence on battery health. A smaller BD value indicates less cumulative SoH loss during operation.

4.3.7. Average Overload Duration (AOD)

Average overload duration is used to quantify the cumulative duration for which the vehicle powertrain or battery system operates beyond its allowable power limit. Let $P_i(t)$ denote the power demand of bus i at time t , and let $P_i^{\max}$ denote its allowable power limit. The overload indicator function is defined as

$$I_i(t) = \begin{cases} 1, & P_i(t) > P_i^{\max} \\ 0, & P_i(t) \leq P_i^{\max} \end{cases} \quad (47)$$

For a discrete simulation with a time step of Δt , the overload duration of bus i is calculated as

$$D_i = \sum_{k=1}^K I_i(k) \Delta t. \quad (48)$$

Accordingly, the average overload duration over all buses is defined as

$$\text{AOD} = \frac{1}{N} \sum_{i=1}^N D_i = \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^N I_i(k) \Delta t \quad (49)$$

4.4. Simulation Experiments

To enable a unified assessment of multi-dimensional performance, this paper normalizes the seven evaluation metrics so that higher values indicate better performance and maps the results onto the interval $[0, 1]$. The comprehensive comparison of the various strategies is presented as a radar chart, as shown in Figure 6. This evaluation quantitatively verifies the overall advantages of the proposed strategy at a statistical level.

As shown in Figure 6, the overall performance of the No Control strategy is significantly lower than that of the other control strategies, and its area on the radar chart is the smallest. Specifically, this leads to lower average vehicle speeds and system energy efficiency, as well as significantly longer average travel times. These results further confirm the necessity and effectiveness of introducing active control mechanisms into public transit systems.

The Holding strategy achieves strong performance in terms of battery degradation mitigation, average waiting time, and average overload duration. However, this strategy performed significantly worse in terms of energy efficiency and achieved only average performance in speed and travel efficiency. This indicates that, although its control mechanism, based on stop-level load regulation, contributes to system stability and constraint satisfaction, the lack of joint perception and coordinated optimization of both traffic and energy states often causes the operating point to deviate from the energy-efficient optimal range, thereby reducing energy utilization efficiency.

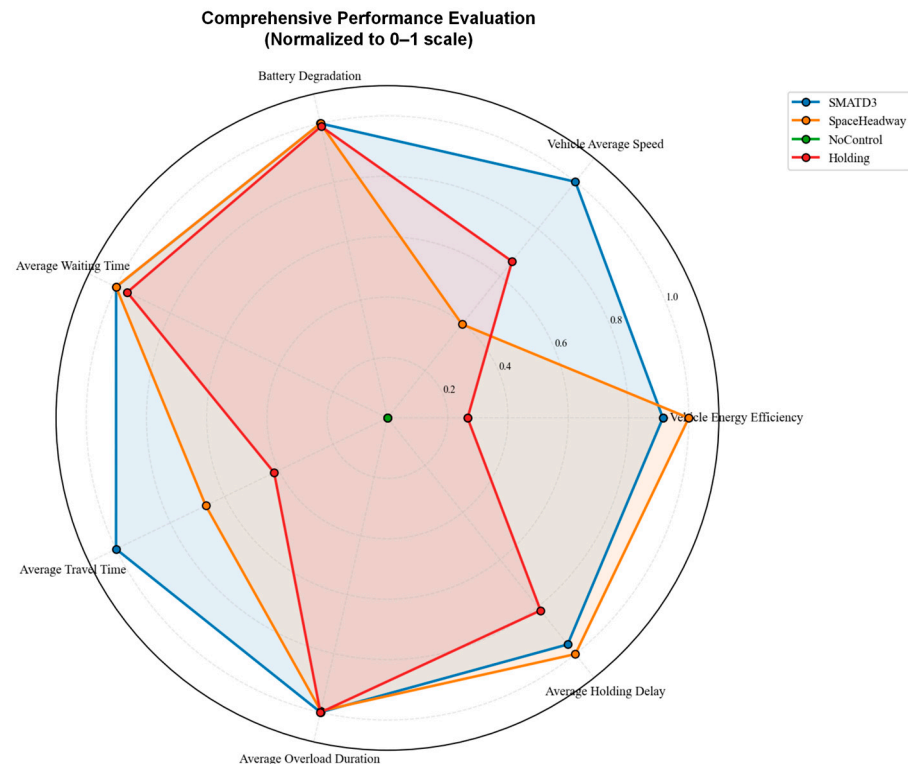


Figure 6. Normalized radar plot of various performance metrics.

SpaceHeadway achieves the highest normalized performance in terms of energy efficiency and performs comparably to the optimal strategy with respect to battery degradation and average waiting time. However, this strategy showed significantly poorer performance in terms of average vehicle speed and recorded the lowest scores on delay related metrics, including average stop delay, average overload duration, and average travel time. Space-Headway is a local car-following control rule based on maintaining a fixed inter-vehicle distance. It primarily improves energy efficiency by suppressing acceleration and deceleration variations, thereby reducing power fluctuations and energy loss. However, this rule lacks state feedback and coordinated scheduling that accounts for network traffic conditions and energy node capacity constraints. Consequently, it tends to promote conservative following behavior and increased headways, which reduces road capacity and translates efficiency losses into cumulative delays at stops and extended travel times. Furthermore, owing to the absence of active load balancing and overload constraint management at the node level, passenger demand becomes more concentrated in both time and space, rendering the effective control of overload duration difficult.

The SMATD3 strategy achieves the largest coverage area on the normalized radar plot, demonstrating strong overall trade-off capability. Specifically, SMATD3 achieves the highest normalized score in the average vehicle speed dimension and demonstrates the most significant improvement in traffic operational efficiency. Although its score in the energy efficiency dimension is slightly lower than that of SpaceHeadway, it still remains at a high level, indicating that the improvement in efficiency does not come at the expense of energy utilization performance. In terms of battery degradation and average waiting time, SMATD3 performs comparably to the optimal strategy, indicating that its speed gains are not achieved at the expense of significant battery health degradation or waiting costs. In addition, SMATD3 maintains high normalized scores on key constraint-related metrics, such as average stop delay and average overload duration. While its average travel time is slightly lower than that of No Control, it still outperforms both Holding and SpaceHeadway overall.

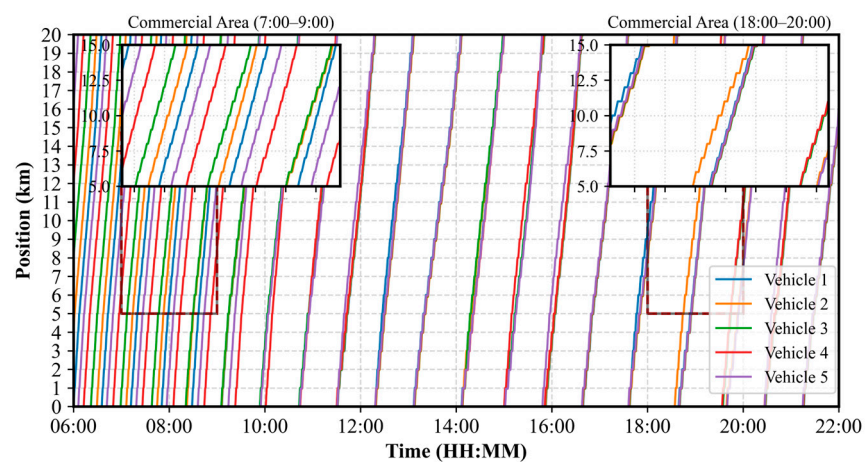
Overall, SMATD3 demonstrates a distinct advantage in key efficiency metrics and consistently maintains high or near-optimal performance across the remaining metrics. The above results confirm that the proposed SMATD3 strategy outperforms comparable methods in balancing vehicle operational efficiency, energy system performance, and service quality. The underlying logic of its multi-agent cooperative optimization provides a robust control framework for intelligent transportation and integrated energy management in real-world bus systems.

Table 5 summarizes the comparative performance of different control strategies in terms of service quality, operational efficiency, energy utilization, and battery health degradation. AWT, ATT, and AHD denote the average waiting time, average travel time, and average holding delay, respectively. Cumulative *dSoH* represents the accumulated battery health degradation over the simulation horizon, Energy Efficiency denotes the vehicle energy efficiency, and Speed denotes the average operating speed. For Cumulative *dSoH*, Energy Efficiency, and Speed, the results are reported in the form of the mean $\pm$ standard deviation to indicate the dispersion among vehicles or simulation runs.

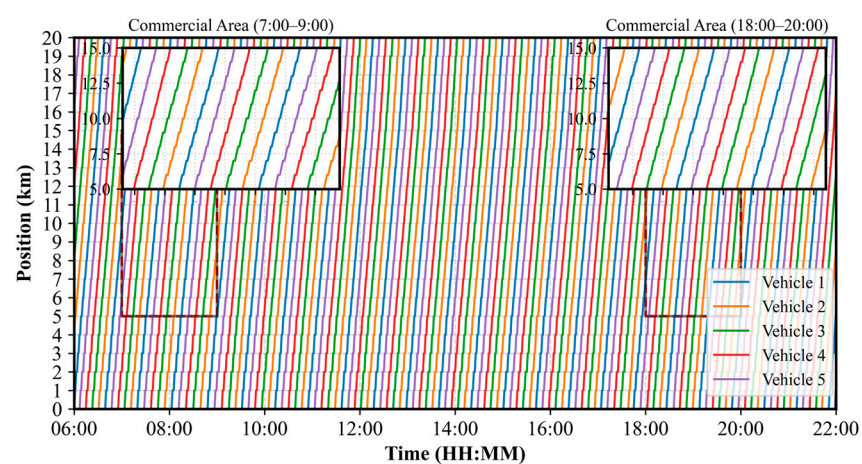
Table 5. Comparative performance of different control strategies under multiple evaluation metrics.

Algorithm	AWT(s)	ATT(s)	AHD(s)	Cumulative dSoH	Energy Efficiency	Speed (km/h)
No Control	1020.12	2788.04	26.73	1.24×10^{-7}	610.945868 ± 23.36487	24.725918 ± 0.40643
Holding	263.01	2570.51	17.34	1.59×10^{-9}	617.138358 ± 3.490732	29.055554 ± 0.013103
SpaceHeadway	262.09	2439.24	15.21	3.50×10^{-10}	634.243939 ± 3.197657	27.316257 ± 0.003587
SMATD3	229.69	2267.04	15.69	3.32×10^{-10}	632.241277 ± 5.105422	31.256087 ± 0.035838

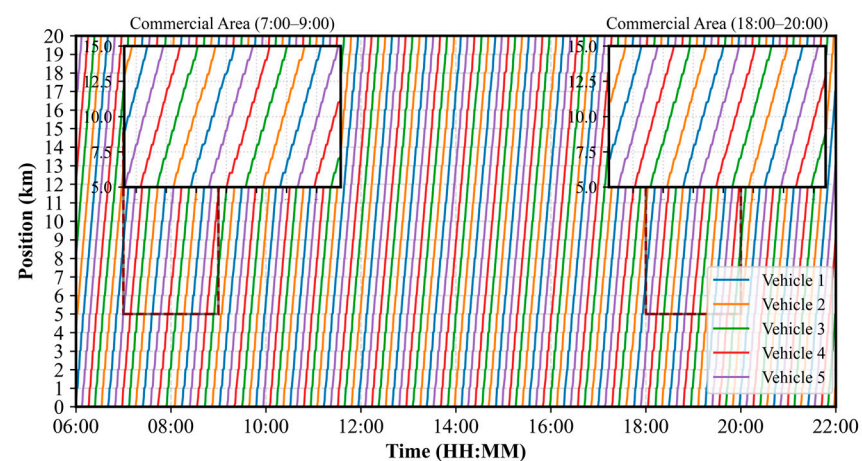
Trajectory data provide a direct representation of the operational state and dispatching effectiveness of a bus system. By analyzing headway stability, congestion propagation, and congestion accumulation, the key operational characteristics under different control strategies, such as bus bunching, headway regularity, and the spatiotemporal spread and accumulation of congestion, can be directly observed. Figure 7 presents the spatiotemporal trajectories of a bus fleet on a closed-loop route under four different control strategies. The horizontal axis represents operating time from 6:00 a.m. to 10:00 p.m., and the vertical axis represents the vehicles' spatial positions along the route. Trajectories of different colors correspond to different vehicles. The commercial zone, spanning the 5 km to 15 km segment of the route, is prominently indicated in Figure 7, which also offers magnified insets for the morning rush hour (7:00–9:00 a.m.) and the evening rush hour (6:00–8:00 p.m.). By analyzing the spatiotemporal trajectory map together with the performance metrics in Table 5, it is evident that, without a control strategy, vehicle trajectories are highly clustered and exhibit significant tailgating. This leads to elevated AWT, ATT, and AHD, indicating a substantial waste of capacity. Holding controls alleviate congestion and significantly improve AWT but at the expense of increased stop delays and reduced turnaround efficiency. SpaceHeadway makes vehicle trajectories more uniform through speed adjustment, achieving the lowest AHD and the highest energy efficiency among all strategies, but its average speed is lower than that of Holding and SMATD3, and its travel efficiency is inferior to that of the SMATD3 method. As shown in the magnified views, the proposed method does not forcibly maintain a constant headway at all times. Instead, while preventing bus bunching, it dynamically adjusts vehicle behavior according to vehicle position, operating state, and passenger demand. As a result, SMATD3 achieves the highest average speed of 31.26 km/h, the lowest AWT of 229.69 s, and a moderate AHD of 15.69 s, leading to the best overall performance. These results indicate that SMATD3 provides stronger robustness and operational stability under complex and unpredictable passenger demand conditions.



(a) No Control



(b) Holding



(c) SpaceHeadway

Figure 7. Cont.

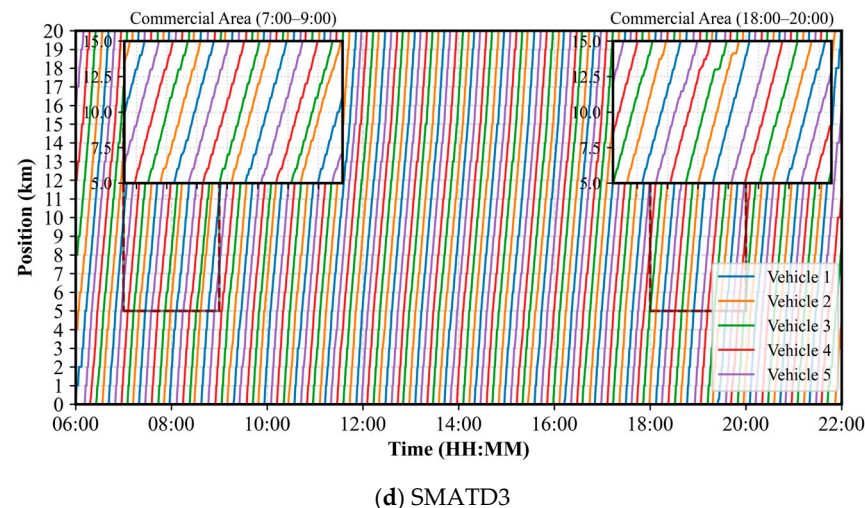


Figure 7. Spatiotemporal map of bus trajectories on same route under different control strategies.

Figure 8 further examines the service equity and resource utilization efficiency of the bus system in relation to the spatiotemporal distribution of passenger demand, presenting the dynamic evolution of passenger loads for five buses under four control strategies from 6:00 a.m. to 10:00 p.m. The solid colored lines represent the actual passenger volume of each bus, the black dashed line indicates the system-wide average passenger volume, and the gray shaded areas depict the range of passenger volumes for each bus at any given time. Overall, all strategies exhibit a typical intraday passenger flow pattern: the system shows a significant increase in average passenger volume during the morning rush hour (7:00–9:00 a.m.) and the evening rush hour (5:30–7:00 p.m.), accompanied by a marked widening of the disparity in passenger volume among vehicles. During midday and off-peak hours, passenger volume generally remains at a lower level and is distributed more evenly. However, different control strategies showed marked differences in their ability to handle peak hour demand and regulate vehicle load balancing. Without any control strategy, passenger loads during the morning and evening rush hours exhibited significant non-stationary fluctuations and polarization: some vehicles operated near full capacity, while others ran nearly empty, highlighting the imbalance in transport capacity and service quality caused by vehicle clustering. Although congestion control reduces peak congestion levels, passenger volume fluctuations remain significant throughout the day, and the disparity in vehicle loads widens again during the latter part of the peak period, indicating limited resilience to demand fluctuations. The SpaceHeadway strategy results in a high degree of overlap in vehicle passenger load trajectories, reduces variability, and significantly improves load balance; however, localized fluctuations and divergence still occur during the rapid rise phase of peak demand. Under the proposed SMATD3 strategy, passenger load increases are smoother, extreme peaks are suppressed, vehicle-to-vehicle variability is minimized, and capacity allocation is the most balanced. In summary, multi-agent reinforcement learning enables adaptive operation and offers the most significant overall advantages in suppressing vehicle queuing, balancing passenger loading, and improving service reliability.

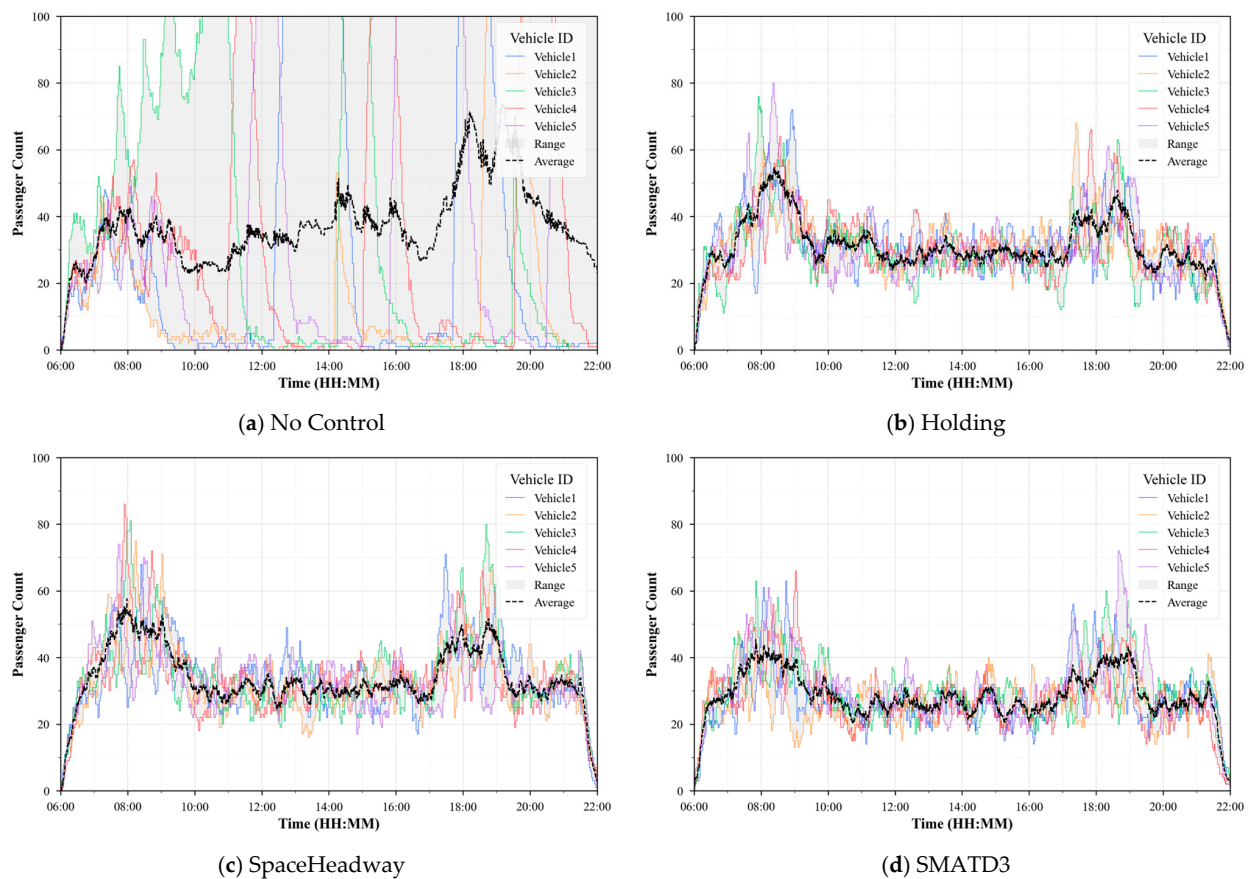


Figure 8. Changes in bus passenger volume over time under different control strategies.

To further elucidate the impact of different control strategies on bus operational mechanisms, this paper statistically analyzes the distribution of average operating speeds for each vehicle along the route, as shown in Figure 9. The No Control strategy lacks an inter-vehicle coordination mechanism. Although buses tend to operate at relatively high speeds, their overall performance is adversely affected by bus bunching, queue propagation, and stop service delays. As a result, the average speed is only 24.73 km/h, and the energy efficiency is the lowest among all strategies, at 610.95. Meanwhile, its standard deviation of energy efficiency reaches 23.36, which is substantially higher than that of the other strategies. This indicates that uncoordinated operation leads to larger fluctuations in vehicle energy consumption and poorer system-level operational stability. The Holding strategy regulates departure intervals by imposing additional dwell time at stops, thereby alleviating bus bunching to some extent. As a result, the average speed increases to 29.06 km/h, and the energy efficiency improves to 617.14. However, because this strategy primarily relies on stop-based holding for dispatching and lacks a proactive optimization of inter-stop speed and energy states, its improvement in energy efficiency remains limited. The SpaceHeadway strategy maintains balanced spatial spacing between adjacent vehicles through inter-stop speed control. As indicated by the speed distribution, this strategy imposes clear speed constraints on vehicles across several route segments, resulting in smoother vehicle operation. Consequently, it achieves the highest energy efficiency, at 634.24, with a standard deviation of only 3.20, indicating its advantages in reducing speed fluctuations and stabilizing energy utilization. However, because the SpaceHeadway strategy primarily aims to maintain balanced spatial spacing, it tends to produce conservative vehicle operation. As a result, its average speed is only 27.32 km/h, which is lower than those of the Holding and SMATD3 strategies. This indicates that solely pursuing uniform spatial spacing can

improve energy efficiency but may compromise vehicle throughput and route turnover efficiency. In contrast, the SMATD3 strategy demonstrates superior overall coordination capability. Rather than imposing a predefined fixed spatial spacing target, it adaptively adjusts inter-stop cruising speeds according to vehicle operating states, relationships with adjacent vehicles, passenger demand, and energy states. As shown in Figure 8, SMATD3 maintains relatively high operating speeds on most route segments while applying moderate speed adjustments on specific segments in response to system states. Quantitatively, SMATD3 achieves the highest average speed among all strategies, reaching 31.26 km/h, which is approximately 26.4%, 7.6%, and 14.4% higher than those of No Control, Holding, and SpaceHeadway, respectively. Meanwhile, its average energy efficiency reaches 632.24, only slightly lower than that of SpaceHeadway at 634.24 but substantially higher than those of No Control and Holding. These results indicate that SMATD3 significantly improves operational efficiency without noticeably compromising energy utilization performance.

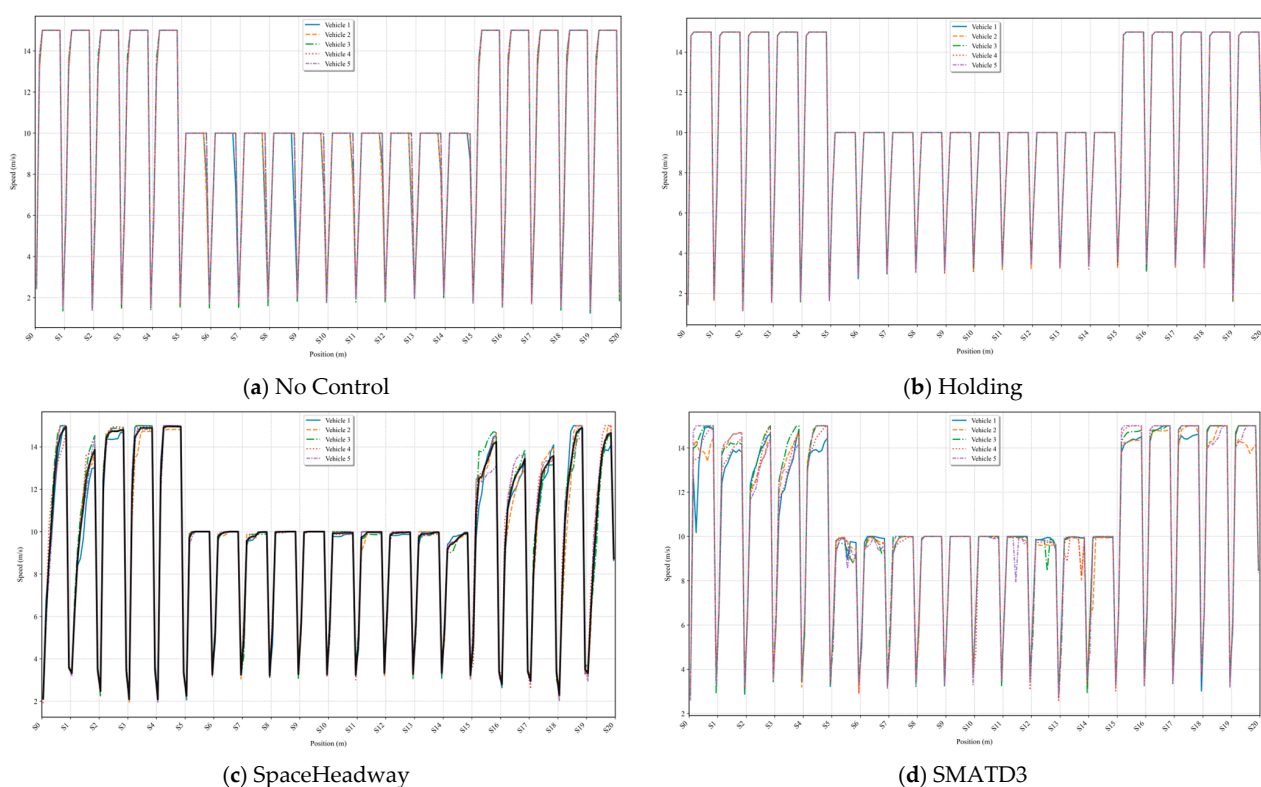


Figure 9. Comparison of speed distributions along route under different control strategies.

As shown in Figure 10, the SoC of the vehicles decreases with increasing cumulative mileage under all control strategies; however, noticeable differences exist in the smoothness of the SoC trajectories and their consistency across vehicles. Under the No Control strategy, the SOC trajectories exhibit a high degree of dispersion, and the cumulative battery degradation increases substantially. This indicates that, in the absence of coordinated control, vehicles are more susceptible to bus bunching, frequent stop-and-go operations, and localized high-load conditions, which intensify power demand fluctuations and accelerate battery health degradation. Table 5 further shows that the average cumulative $dSoH$ under the No Control strategy reaches 1.24×10^{-7} , which is significantly higher than that under the other control strategies. In addition, its standard deviation of energy efficiency is 23.36, indicating large inter-vehicle fluctuations in energy consumption and poor consistency in system operation. The Holding strategy alleviates spacing fluctuations by imposing additional dwell time at stops, resulting in smoother SoC trajectories and reduced inter-vehicle

differences compared with the No Control strategy. Its average cumulative $dSOH$ decreases to 1.59×10^{-9} , indicating that stop-based holding control can mitigate battery degradation caused by high-load operation to some extent. However, this strategy primarily relies on passive stop-level regulation and lacks a coordinated optimization of vehicle speed, power demand, and battery state. Consequently, its energy efficiency is only 617.14, which remains lower than those of SpaceHeadway and SMATD3. The SpaceHeadway strategy maintains balanced spatial spacing between adjacent vehicles through inter-stop speed regulation, resulting in smoother vehicle operation. Consequently, its SOC trajectories exhibit strong consistency across vehicles, the average cumulative $dSOH$ decreases to 3.50×10^{-10} , and the energy efficiency reaches the highest value of 634.24 with the smallest fluctuation. These results indicate that spatial-spacing-based control can effectively suppress speed variations and power shocks, thereby improving energy utilization and mitigating battery degradation. However, because this strategy primarily aims to maintain uniform spatial spacing, it tends to produce conservative vehicle operation. The preceding speed analysis also indicates that this strategy sacrifices operational efficiency to some extent. In contrast, the proposed SMATD3 strategy achieves a better balance between battery health protection and energy efficiency. As shown in Figure 8, the SoC trajectories of all vehicles under SMATD3 remain generally smooth and consistent, indicating that the strategy can effectively coordinate the energy consumption processes of different vehicles and prevent individual buses from operating under prolonged high-load or high-degradation conditions. Table 5 shows that the average cumulative $dSOH$ of SMATD3 is 3.32×10^{-10} , the lowest among all strategies. This represents reductions of approximately 99.7% and 79.1% compared with No Control and Holding, respectively, and is slightly better than SpaceHeadway. Meanwhile, the energy efficiency of SMATD3 is 632.24, only slightly lower than that of SpaceHeadway at 634.24. Considering the preceding speed analysis, SMATD3 also achieves the highest average operating speed, indicating that its battery health benefits are not obtained at the expense of operational efficiency.

Overall, the No Control strategy lacks coordinated control and is therefore prone to greater energy consumption fluctuations and accelerated battery degradation. The Holding strategy can mitigate battery degradation to some extent, but its optimization capability is limited by stop-level local regulation. SpaceHeadway achieves favorable energy efficiency; however, it tends to produce conservative vehicle operation. In contrast, SMATD3 leverages multi-agent cooperative decision-making to smooth power demand, reduce battery health degradation, and maintain high energy efficiency while preserving vehicle operating efficiency. Therefore, SMATD3 achieves a superior overall trade-off among battery health protection, energy efficiency, and operational efficiency.

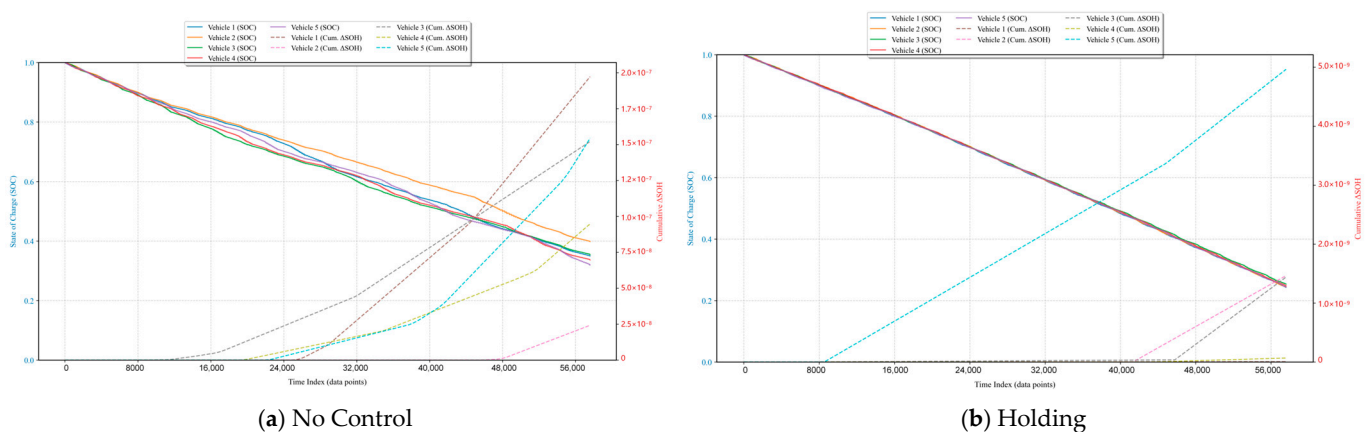


Figure 10. Cont.

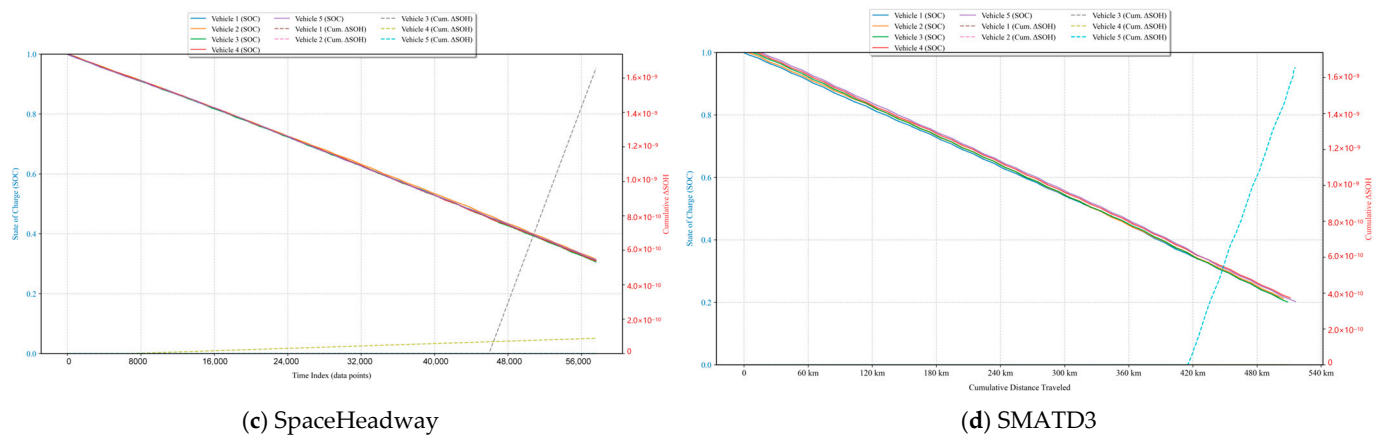


Figure 10. Comparison of SoC and battery lifetime degradation under different control strategies.

To further examine the performance of the proposed SMATD3 strategy beyond average-level indicators, Figure 11 presents a detailed analysis across temporal periods, stops, and individual vehicles. Specifically, Figure 11a compares the mileage accumulated over the full day and peak hour periods, Figure 11b compares the corresponding average speeds, Figure 11c illustrates the Δ SoH distribution during peak and off-peak periods, Figure 11d presents the stop-level dwell time distribution, Figure 11e shows the inter-stop speed distribution, and Figure 11f compares energy consumption between peak and off-peak operation. The results reveal that peak hour operation imposes a disproportionately high burden on both energy consumption and battery health. Although peak hour mileage accounts for only approximately one quarter of the total daily mileage, it contributes 30.7% of the total battery health degradation and 44.7% of the total energy consumption. This indicates that peak hour service is substantially more demanding than off-peak operation from the perspectives of both power demand and battery aging.

This phenomenon can be primarily attributed to the sharp increase in passenger demand during peak periods. Higher passenger loads increase the equivalent vehicle mass, thereby raising the traction power demand and battery current. Meanwhile, the average speed comparison shows that buses operate at noticeably lower speeds during peak hours than over the full-day period. Lower operating speeds, together with more frequent stopping, boarding, alighting, and acceleration events, intensify power fluctuations and accelerate battery degradation. These findings highlight the necessity of jointly considering passenger demand dynamics and battery health evolution in electric bus control.

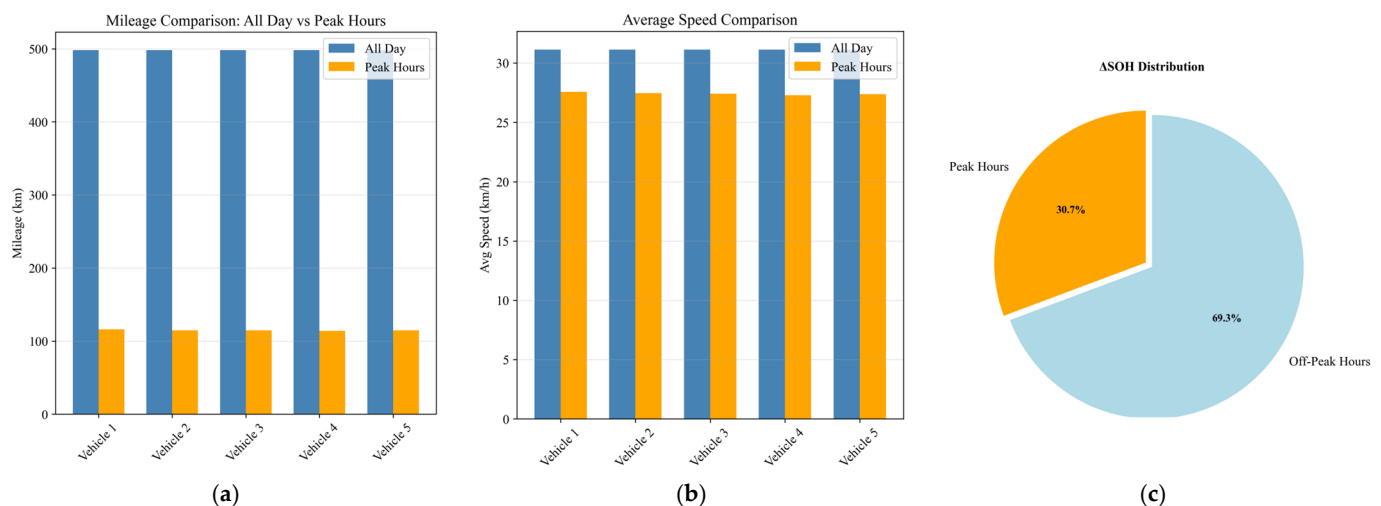


Figure 11. Cont.

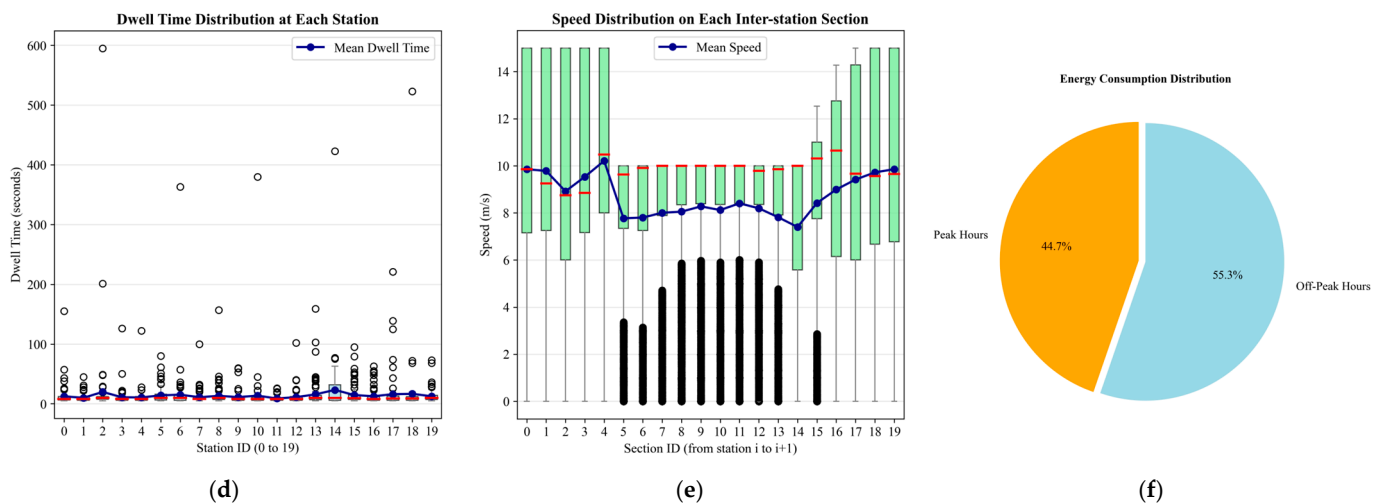


Figure 11. Detailed performance analysis of proposed SMATD3 strategy across time periods, stops, and individual vehicles. (a) Mileage Comparison: All Day vs Peak Hours, (b) Average Speed Comparison All, (c) Δ SOH Distribution, (d) Dwell Time Distribution at Each Station, (e) Speed Distribution on Each Inter-station Section, (f) Energy Consumption Distribution.

From the stop-level perspective, the inter-stop speed distribution reveals clear differences between high-demand and low-demand sections. Sections associated with high-demand stops generally exhibit lower mean speeds and larger speed dispersion, reflecting the combined effects of passenger accumulation, dwell time variation, and control intervention. Nevertheless, the stop-level dwell time distribution indicates that the proposed SMATD3 strategy maintains relatively balanced station service performance. The mean dwell time at most stops is generally controlled within approximately 2 min, suggesting that local passenger demand fluctuations can be effectively absorbed without developing into severe system-wide bus bunching.

From the vehicle-level perspective, the mileage and average speed of the five buses remain highly consistent under the proposed strategy. This indicates that SMATD3 does not improve system-level performance by sacrificing the operation of any particular vehicle. Instead, it coordinates the operating states of all buses and maintains balanced vehicle-level performance. The Δ SoH distribution further shows that peak hour operation contributes a substantial proportion of battery degradation, confirming the strong coupling among time-varying passenger demand, vehicle operating conditions, and battery aging. Overall, these detailed results demonstrate that the proposed SMATD3 strategy can effectively adapt to peak period disturbances, high-demand stops, and vehicle-level operational differences, rather than merely improving average system-level indicators.

5. Conclusions

A multi-agent reinforcement learning framework is proposed for the cooperative dynamic control of electric buses. Unlike existing methods that treat buses as homogeneous agents and impose fixed headway constraints, SMATD3 explicitly accounts for vehicle heterogeneity in passenger load, dynamic response, and battery health. The framework adopts centralized training with decentralized execution, a shared actor network for homogeneous agents under heterogeneous states, and twin critics with target policy smoothing and delayed updates to ensure stable learning. Simulated under time-varying passenger demand with full vehicle and battery dynamics, SMATD3 consistently outperforms No Control, Holding, and SpaceHeadway baselines. It achieves the highest average speed, the lowest passenger waiting time, and the lowest cumulative battery degradation while maintaining competitive energy efficiency. Further analyses confirm that SMATD3 adaptively

coordinates vehicle behavior under dynamic disturbances, effectively suppressing bus bunching and achieving a superior trade-off among operational efficiency, service quality, energy consumption, and battery health protection without requiring a fixed spacing target. Quantitatively, the No Control strategy produced an average passenger waiting time of 1020.12 s, an average travel time of 2788.04 s, and an average speed of 24.73 km/h, indicating serious passenger service deterioration under irregular and bunched bus operations. By contrast, SMATD3 reduced the average passenger waiting time to 229.69 s and shortened the average travel time to 2267.04 s while increasing the average speed to 31.26 km/h. Therefore, compared with SMATD3, uncontrolled operation led to an additional passenger waiting time of 790.43 s, equivalent to approximately 13.17 min per passenger, and an additional average travel time of 521.00 s, equivalent to approximately 8.68 min. Correspondingly, SMATD3 reduced passenger waiting time by 77.48% and average travel time by 18.69% while increasing the average operating speed by 26.41%. These results quantitatively confirm the practical value of electric bus operation control for mitigating bunching-related passenger delay and service deterioration.

Although the proposed SMATD3-based cooperative control strategy performs well in the studied closed-loop electric bus system, its applicability is mainly limited to high- and medium-frequency services where adjacent buses interact sufficiently. When the scheduled headway is too large, such interactions weaken, and the main control objective shifts from real-time bunching mitigation to timetable adherence, fleet size planning, and service frequency design. In this study, the average travel time under SMATD3 is 2267.04 s with five buses, corresponding to a nominal headway of approximately 7.56 min. Therefore, the proposed method is more suitable for services with headways around or below 8 min. For routes with headways substantially exceeding 8–10 min, it should be combined with timetable-based holding, fleet size adjustment, or frequency optimization to ensure service reliability. The proposed cooperative control offers clear sustainability benefits: it improves service regularity to support higher transit use and lower congestion emissions while reducing energy use fluctuation and battery degradation to lessen electricity demand and lifecycle impacts. Future work will extend SMATD3 to more realistic urban operating environments, including mixed traffic, signalized intersections, variable road capacity, charging constraints, and multiline networks. In addition, real AFC/APC data and bus operation records will be used to calibrate passenger arrival rates, OD distributions, time-of-day coefficients, peak demand intensity, school-related demand, distance decay parameters, and travel time uncertainty.

Author Contributions: Methodology, software, and writing—original draft and editing, H.Z.; validation and writing—review and editing, H.D.; investigation and data curation, H.W., Z.D., R.X. and H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 52402522), the Key Research and Development Program of Shanxi Province (Grant No. 202502010101013), and the 2025 Annual Postgraduate Academic Innovation Program of Shanxi Province (Grant Nos. 2025XS121 and 2025SJ300).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this analysis are publicly available, and access is provided in the text.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Melo, S.; Gomes, R.; Abbasi, R.; Arantes, A. Demand-Responsive Transport for Urban Mobility: Integrating Mobile Data Analytics to Enhance Public Transportation Systems. *Sustainability* **2024**, *16*, 4367. [\[CrossRef\]](#)
2. Alam, T.; Gupta, R.; Ahamed, N.N.; Ullah, A.; Almaghthwi, A. Towards Sustainable IoT-Based Smart Mobility Systems in Smart Cities. *GeoJournal* **2024**, *89*, 235. [\[CrossRef\]](#)
3. Ercan, T.; Onat, N.C.; Keya, N.; Tatari, O.; Eluru, N.; Kucukvar, M. Autonomous Electric Vehicles Can Reduce Carbon Emissions and Air Pollution in Cities. *Transp. Res. Part D Transp. Environ.* **2022**, *112*, 103472. [\[CrossRef\]](#)
4. Yang, C.; Jin, S.; Yao, W.; Rong, D.; Bai, C.; Alagbé, J.A. Collaborative Optimization of Intersection Signals and Speed Guidance for Buses Run on Overlapping Route Segments under Connected Environment. *Comput.-Aided Civ. Infrastruct. Eng.* **2024**, *39*, 3289–3316. [\[CrossRef\]](#)
5. Lin, H.; Tang, C. Analysis and Optimization of Urban Public Transport Lines Based on Multiobjective Adaptive Particle Swarm Optimization. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 16786–16798. [\[CrossRef\]](#)
6. Zhang, S.; Lo, H.K. Two-Way-Looking Self-Equalizing Headway Control for Bus Operations. *Transp. Res. Part B Methodol.* **2018**, *110*, 280–301. [\[CrossRef\]](#)
7. Xue, Y.; Zhong, M.; Xue, L.; Tu, H.; Tan, C.; Kong, Q.; Guan, H. A Real-Time Control Strategy for Bus Operation to Alleviate Bus Bunching. *Sustainability* **2022**, *14*, 7870. [\[CrossRef\]](#)
8. Jeffers, M.; Eudy, L. *Foothill Transit Battery Electric Bus Evaluation (Final Report)*; National Renewable Energy Laboratory (NREL): Golden, CO, USA, 2021; p. NREL/TP-5400-80022.
9. Johnson, C.; Nobler, E.; Eudy, L.; Jeffers, M. *Financial Analysis of Battery Electric Transit Buses*; National Renewable Energy Laboratory (NREL): Golden, CO, USA, 2020; p. NREL/TP--5400-74832.
10. Yu, M.; Yang, T.; Li, C.; Jin, Y.; Xu, Y. Mitigating Bus Bunching via Hierarchical Multi-Agent Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 9675–9692. [\[CrossRef\]](#)
11. Liu, H.; Qiu, H.; Lu, W.; Shan, X. A Multi-Objective Reinforcement Learning Framework for Energy-Efficient Electric Bus Operations. *Sustainability* **2025**, *17*, 10695. [\[CrossRef\]](#)
12. Laskaris, G.; Cats, O.; Jenelius, E.; Rinaldi, M.; Viti, F. A Holding Control Strategy for Diverging Bus Lines. *Transp. Res. Part C Emerg. Technol.* **2021**, *126*, 103087. [\[CrossRef\]](#)
13. Liang, S.; Ma, M.; He, S. Multiobjective Optimal Formulations for Bus Fleet Size of Public Transit under Headway-Based Holding Control. *J. Adv. Transp.* **2019**, *2019*, 2452348. [\[CrossRef\]](#)
14. Petit, A.; Lei, C.; Ouyang, Y. Multiline Bus Bunching Control via Vehicle Substitution. *Transp. Res. Part B Methodol.* **2019**, *126*, 68–86. [\[CrossRef\]](#)
15. Wu, W.; Liu, R.; Jin, W. Integrating Bus Holding Control Strategies and Schedule Recovery: Simulation-Based Comparison and Recommendation. *J. Adv. Transp.* **2018**, *2018*, 9407801. [\[CrossRef\]](#)
16. Zhou, X.; Wang, Y.; Ji, X.; Cottrill, C. Coordinated Control Strategy for Multi-Line Bus Bunching in Common Corridors. *Sustainability* **2019**, *11*, 6221. [\[CrossRef\]](#)
17. Li, C.; Shen, M.; Hu, Q.; Zhen, L.; Xiao, F. Dynamic Bus Holding Control for Multi-Line Busy Bus Corridors: Mitigating Bus Queues and Improving Headway Regularity via Graph-Aware Deep Reinforcement Learning. *Transp. Res. Part E Logist. Transp. Rev.* **2026**, *207*, 104603. [\[CrossRef\]](#)
18. He, S.-X.; Dong, J.; Liang, S.-D.; Yuan, P.-C. An Approach to Improve the Operational Stability of a Bus Line by Adjusting Bus Speeds on the Dedicated Bus Lanes. *Transp. Res. Part C Emerg. Technol.* **2019**, *107*, 54–69. [\[CrossRef\]](#)
19. Wang, J.; Sun, L. Dynamic Holding Control to Avoid Bus Bunching: A Multi-Agent Deep Reinforcement Learning Framework. *Transp. Res. Part C Emerg. Technol.* **2020**, *116*, 102661. [\[CrossRef\]](#)
20. Liu, D.; Xiao, F.; Luo, J.; Yang, F. Deep Reinforcement Learning-Based Holding Control for Bus Bunching under Stochastic Travel Time and Demand. *Sustainability* **2023**, *15*, 10947. [\[CrossRef\]](#)
21. Gkiotsalitis, K. Stop-Skipping in Rolling Horizons. *Transp. A Transp. Sci.* **2021**, *17*, 492–520. [\[CrossRef\]](#)
22. Lee, E.H.; Lee, I.; Cho, S.-H.; Kho, S.-Y.; Kim, D.-K. A Travel Behavior-Based Skip-Stop Strategy Considering Train Choice Behaviors Based on Smartcard Data. *Sustainability* **2019**, *11*, 2791. [\[CrossRef\]](#)
23. Wen, L.; Hu, L.; Zhou, W.; Ren, G.; Zhang, N. Soft Actor-Critic Deep Reinforcement Learning for Train Timetable Collaborative Optimization of Large-Scale Urban Rail Transit Network Under Dynamic Demand. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 7021–7035. [\[CrossRef\]](#)
24. Li, C.; Liu, H.; Jia, Q.; Xiong, L.; Wu, H. Intelligent Cooperative Perception Technology for Vehicles and Experiments Based on V2V/V2I Semantic Communication. *Electronics* **2025**, *14*, 4969. [\[CrossRef\]](#)
25. Dey, K.C.; Rayamajhi, A.; Chowdhury, M.; Bhavsar, P.; Martin, J. Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) Communication in a Heterogeneous Wireless Network—Performance Evaluation. *Transp. Res. Part C Emerg. Technol.* **2016**, *68*, 168–184. [\[CrossRef\]](#)

26. Treethidtaphat, W.; Pattara-Atikom, W.; Khaimook, S. Bus Arrival Time Prediction at Any Distance of Bus Route Using Deep Neural Network Model. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 988–992.
27. Bian, B.; Zhu, N.; Pinedo, M.; Ma, S.; Yu, Q. An Optimization-Based Speed-Control Method for High Frequency Buses Serving Curbside Stops. *Transp. Res. Part C Emerg. Technol.* **2020**, *121*, 102860. [[CrossRef](#)]
28. He, S.-X. An Anti-Bunching Strategy to Improve Bus Schedule and Headway Reliability by Making Use of the Available Accurate Information. *Comput. Ind. Eng.* **2015**, *85*, 17–32. [[CrossRef](#)]
29. Li, S.; Liu, R.; Yang, L.; Gao, Z. Robust Dynamic Bus Controls Considering Delay Disturbances and Passenger Demand Uncertainty. *Transp. Res. Part B Methodol.* **2019**, *123*, 88–109. [[CrossRef](#)]
30. Lacombe, R.; Gros, S.; Murgovski, N.; Kulcsar, B. Bilevel Optimization for Bunching Mitigation and Eco-Driving of Electric Bus Lines. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 10662–10679. [[CrossRef](#)]
31. Jin, K.; Li, X.; Wang, W.; Hua, X.; Long, W. Energy-Optimal Speed Control for Connected Electric Buses Considering Passenger Load. *J. Clean. Prod.* **2023**, *385*, 135773. [[CrossRef](#)]
32. Han, Y.; Ma, X.; Li, X.; Bian, J.; Wang, W.; Jiang, S. Speed Planning for Integrated Eco-Driving and Bus Bunching Mitigation in Connected Electric Buses. *IEEE Trans. Transp. Electrification* **2025**, *11*, 9165–9176. [[CrossRef](#)]
33. Shi, H.; Nie, Q.; Fu, S.; Wang, X.; Zhou, Y.; Ran, B. A Distributed Deep Reinforcement Learning-Based Integrated Dynamic Bus Control System in a Connected Environment. *Comput.-Aided Civ. Infrastruct. Eng.* **2022**, *37*, 2016–2032. [[CrossRef](#)]
34. Rücker, F.; Figgenger, J.; Schoeneberger, I.; Sauer, D.U. Battery Electric Vehicles in Commercial Fleets: Use Profiles, Battery Aging, and Open-Access Data. *J. Energy Storage* **2024**, *86*, 111030. [[CrossRef](#)]
35. Wei, H.; Zhang, N.; Liang, J.; Ai, Q.; Zhao, W.; Huang, T.; Zhang, Y. Deep Reinforcement Learning Based Direct Torque Control Strategy for Distributed Drive Electric Vehicles Considering Active Safety and Energy Saving Performance. *Energy* **2022**, *238*, 121725. [[CrossRef](#)]
36. Jia, C.; He, H.; Zhou, J.; Li, J.; Wei, Z.; Li, K. Learning-Based Model Predictive Energy Management for Fuel Cell Hybrid Electric Bus with Health-Aware Control. *Appl. Energy* **2024**, *355*, 122228. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.