energies

MDPI

Article

A Dual-Evaluation Collaborative Decision Deep Reinforcement Learning Strategy for Multi-Constraint Energy Management of Hybrid Electric Vehicles

Quanhong Zeng [1], Chun Wang [1,2,*], Jiahuan Lu [3,*], Hao Jin [1] and Hang Chen [1]

[1] School of Mechanical Engineering, Sichuan University of Science and Engineering, Yibin 644000, China; zengquanhong1412@163.com (Q.Z.); jinhao20161107@163.com (H.J.); chh_1113@126.com (H.C.)
[2] Sichuan Provincial Key Lab of Process Equipment and Control, Sichuan University of Science and Engineering, Yibin 644000, China
[3] College of Engineering, South China Agricultural University, Guangzhou 510642, China
* Correspondence: susewangchun@163.com (C.W.); jhlu@scau.edu.cn (J.L.)

Abstract

Coordinated control of fuel economy, state-of-charge (SOC), thermal state, and engine operation is central to hybrid electric vehicles (HEVs) energy management. In existing discrete-action deep reinforcement learning (DRL) methods, action selection usually relies on the Q-values predicted by the value network, whereas battery SOC and thermal constraints are often incorporated only through reward penalties after action execution. Accordingly, the short-term consequences of candidate actions cannot be compared directly during action screening, and the learned policy may exhibit SOC drift, thermal buildup, and frequent engine start-stop events. To address this issue, this paper proposes RAEP-KAN-DQN (Risk-Aware Electrothermal Prediction-KAN-DQN). The method retains the value-function framework and introduces executable-action mapping, one-step state prediction, and risk-aware action selection. Experimental results show that the main strength of RAEP-KAN-DQN lies in its balanced multi-objective performance. Under the standard condition, the method reduces fuel cost by 33.58% relative to Deep Q-network (DQN), while lowering the peak core and surface temperatures by 11.73% and 10.99%, respectively. When the initial temperature is raised to 40 °C, fuel cost decreases by 13.61% relative to the standard temperature case, and core and surface temperature rises are 62.13% and 62.66% lower than those of DQN. To further evaluate driving-cycle robustness, the HWFET driving cycle is adopted as a high-speed continuous-load condition. The proposed method limits SOC deviation to 0.111, which is 54.32% lower than that of DQN, while achieving a balanced trade-off among terminal SOC deviation, battery thermal loading, and engine operation. These results show that the proposed dual-evaluation decision strategy can improve the coordinated control capability of DRL-based energy management under multiple constraints.

Keywords: hybrid electric vehicle; energy management; deep reinforcement learning; dual-evaluation collaborative decision; risk assessment; battery thermal management

Academic Editor: Federico Barrero

Received: 20 October 2026
Revised: 3 November 2026
Accepted: 5 November 2026
Published: 9 November 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) license.

1. Introduction

Hybrid electric vehicles (HEVs) play an important role in the transition from internal combustion engine powertrains to fully electric propulsion systems. They offer a practical

 https://doi.org/10.3390/en19215381

balance among fuel economy, driving range, and infrastructure requirements [1–4]. More broadly, alternative drive systems and powertrain configurations have also been explored as approaches to improving the environmental sustainability of passenger transportation [5]. A typical HEV powertrain includes an engine, electric machines, and a lithium-ion traction battery. The energy management system must allocate power among these sources in real time. The goal is to reduce fuel use without violating drivability constraints [6–8]. However, HEV energy management is not a single-objective fuel-saving problem. Fuel economy, battery state-of-charge (SOC), battery thermal state, and engine operation all affect system efficiency, component durability, and control feasibility [9–11]. These objectives are strongly coupled under time-varying driving conditions. HEV energy management is therefore a constrained multi-objective optimization problem [12–14].

Deep reinforcement learning (DRL) has emerged as a promising data-driven approach beyond rule-based and model-based energy management strategy (EMS) design [15–17]. As DRL has been introduced into HEV energy management, it has been used to reduce the dependence of rule-based strategies on expert calibration and to address the limited real-time applicability of global optimization methods [18,19]. Early studies showed that deep Q-networks (DQN) could learn near-optimal power allocation policies in discrete action spaces [20,21]. DQN also laid the foundation for combining deep neural networks with reinforcement learning, thereby enabling policy approximation in complex state spaces [22]. Building on this framework, Qi et al. [23] proposed a hierarchical deep Q-learning (DQL-H) strategy to address insufficient long-term return feedback and slow training convergence. By decomposing complex decision-making into multiple levels, their method alleviated reward sparsity to some extent and improved both learning efficiency and fuel economy. Hierarchical DRL was further applied to the integrated optimization of adaptive cruise control (ACC) and EMS, allowing car-following control and energy allocation to be coordinated within a unified framework. This reduced the computational burden of coordinating multiple tasks and showed the potential of reinforcement learning for multi-objective vehicle control [24]. However, Deep Q-learning (DQL)-based methods are designed for discrete action spaces. They are not well suited to the fine control of continuous variables such as engine torque and motor power. To address this limitation, Ref. [25] introduced deep deterministic policy gradient (DDPG) and guided policy search with expert knowledge, thereby improving continuous action control, training stability, and learning efficiency. Subsequently, Twin delayed deep deterministic policy gradient (TD3) was further applied to HEV energy management together with driving-cycle recognition, enabling the agent to adapt its strategy to different driving conditions, shorten training time, and achieve fuel economy comparable to that of dynamic programming [26]. Overall, existing studies have demonstrated the effectiveness of DRL in HEV energy management. However, further improvements are still needed in generalization under complex driving conditions, multi-objective coordination, and training stability.

With the development of research, its focus has expanded from improving individual algorithms to designing network architectures, integrating control frameworks, and modeling multiple objectives jointly. To address the limited real-time stability of learning-based strategies, TD3 has been combined with model predictive control (MPC) [27]. The rolling optimization capability of MPC was used to help reinforcement-learning-based strategies balance fuel economy and system performance during real-time energy management. These studies showed that combining reinforcement learning with conventional optimal control can improve the practical use of intelligent EMSs in complex driving conditions. Training convergence and adaptability to varying driving conditions are still important for practical deployment. Accordingly, the ATSAC algorithm was proposed to improve policy updates by adaptively adjusting the learning rate and update frequency, and naturalistic driving data were used to construct training cycles [28]. This design

enabled the agent to learn EMSs from data distributions closer to real traffic conditions, thereby improving generalization and computational efficiency. As EMS objectives have expanded from fuel economy to battery health, driving-behavior adaptation, and overall =system performance, dual-layer energy management frameworks considering state of health and driving behavior have been developed. With an improved parallel DDPG method, this framework achieved adaptive multi-objective optimization [29], indicating a broader shift in intelligent EMSs from energy-consumption optimization toward coordinated multi-objective optimization.

Although DRL-based EMSs have achieved promising performance in HEV energy management, different approaches have been developed to address action feasibility and operational constraints. In conventional value-based DRL methods, action selection is generally governed by the estimated Q-values, while constraint information is often incorporated through state representation or reward design. To address action feasibility, Biswas et al. [30] proposed a physics-informed exploration (PIE) strategy combined with prioritized experience replay, which guides the agent toward feasible powertrain actions during training. This approach mainly focuses on improving exploration efficiency and reducing the selection of infeasible actions. In addition, model-based control methods have been integrated with DRL to improve predictive decision-making. Dong et al. [31] further proposed a predictive battery thermal and energy management strategy for connected and automated electric vehicles, which utilizes preview information and cloud-enabled predictive control to coordinate battery thermal and energy management.

Based on the above analysis, this study not only improves the approximation capability of the value network, but also refines the action decision mechanism for HEV energy management. RAEP-KAN-DQN (Risk-Aware Electrothermal Prediction-KAN-DQN) retains the discrete-action DQN framework and introduces a decision-stage dual-evaluation mechanism. The method retains the basic value-function-learning framework, while integrating executable-action mapping, one-step state-outcome evaluation, and risk-aware action selection into the decision process. This allows candidate actions to be evaluated before execution from both long-term value and short-term risk perspectives. As a result, constraint handling is shifted from ex post reward penalties to the action-selection stage. In addition, Kolmogorov–Arnold Networks (KAN) is incorporated into the action-value estimation module. It is further combined with a duel value decomposition structure to improve representation under complex nonlinear energy management. The main contributions of this study are as follows:

(1) An electro-thermal coupled one-step prediction model is developed for HEV online decision-making. It quickly estimates the next-step changes in battery SOC, core temperature, and surface temperature caused by each candidate action.
(2) KAN and dueling value decomposition are embedded in the Q network to improve nonlinear value approximation and to better distinguish between similar discrete actions.
(3) A dual-evaluation collaborative decision strategy is proposed for HEV energy management. Online action selection is determined jointly by long-term Q-value and short-term risk over the set of executable actions.

The remainder of this paper is organized as follows: Section 2 introduces the HEV powertrain model and the electro-thermal coupled battery model. Section 3 presents the Dueling-KAN network architecture and details the training and inference procedures of RAEP-KAN-DQN. Section 4 validates the effectiveness of the proposed method under various representative driving cycles and provides a comparative analysis with baseline energy management strategies. Section 5 concludes this paper and discusses future research directions.

2. Hybrid Electric Vehicle System Modeling

2.1. System Configuration

This study considers a typical series-parallel HEV. The powertrain consists mainly of an engine, a lithium-ion battery pack, a traction motor, and a generator. These components are mechanically coupled through a planetary gear set, which enables power transmission and distribution among the energy sources. The system configuration is shown in Figure 1, and the main model parameters are listed in Table 1 [32–33].

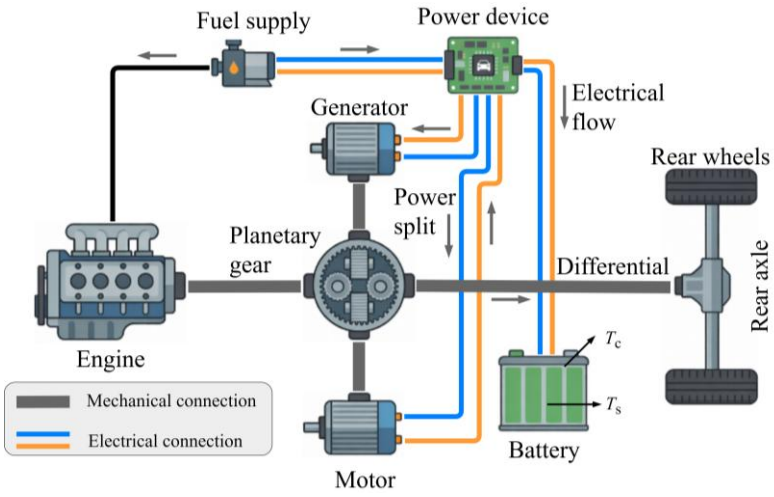


Figure 1. Schematic diagram of the powertrain architecture of the hybrid electric vehicle. (The arrows indicate the direction of energy flow.).

Table 1. Main vehicle and powertrain parameters.

Components	Parameters	Values
Vehicle	Vehicle mass	1449 kg
	Rolling resistance coefficient	0.013
	Air resistance coefficient	0.26
	Frontal area	2.23 m^2
Engine (ICE)	Maximum power	56 kW
	Maximum torque	120 Nm
	Maximum speed	4000 rpm
Traction motor (MG2)	Maximum power	50 kW
	Maximum torque	400 Nm
	Maximum speed	6000 rpm
Generator (MG1)	Maximum power	37.8 kW
	Maximum torque	75 Nm
	Maximum speed	10,000 rpm
Battery pack	Type	NMC
	Cell capacity	2 Ah
	Number of cells in series	56
	Number of cells in parallel	10
Transmission	Gear ratio	3.93
	Characteristic parameter	2.6

2.2. Vehicle Powertrain Modeling

Considering the effects of gravity, rolling resistance, aerodynamic drag, and vehicle mass variation [34], the longitudinal dynamics of the vehicle can be expressed as follows:

$$\begin{cases} F = F_{\mathrm{f}} + F_{\mathrm{w}} + F_{\mathrm{i}} + F_{\mathrm{a}} \\ F_{\mathrm{f}} = mg \cdot f \\ F_{\mathrm{w}} = \dfrac{1}{2} \cdot \rho \cdot A_{\mathrm{f}} \cdot C_{\mathrm{D}} \cdot v^2 \\ F_{\mathrm{i}} = mg \cdot i \\ F_{\mathrm{a}} = m \cdot a \end{cases} \tag{1}$$

where the vehicle mass and gravitational acceleration are represented by m and g, respectively. The rolling resistance coefficient, air density, frontal area, and aerodynamic drag coefficient are denoted by f, ρ, A_{f}, and C_{D}, respectively. In addition, v represents the vehicle speed relative to the air, i denotes the road grade, and a represents the longitudinal acceleration. According to the longitudinal dynamic relationship, the required power can be calculated from the product of F and v in the driving direction, with the driveline efficiency taken into account:

$$P_{\mathrm{req}} = \frac{Fv}{\eta} \tag{2}$$

where η denotes the driveline efficiency. The required power P_{req} can be obtained.

2.3. Modeling of the Engine and Traction Motor

The vehicle powertrain model was developed using quasi-static representations of the engine, traction motor, and generator to meet the computational requirements of online EMS implementation. The operating characteristics of the engine and electric machines were described by steady-state maps in the speed-torque plane. Interpolation was applied to the discrete speed-torque data to obtain the engine fuel consumption map and the electric-machine efficiency map, as shown in Figure 2 [35].

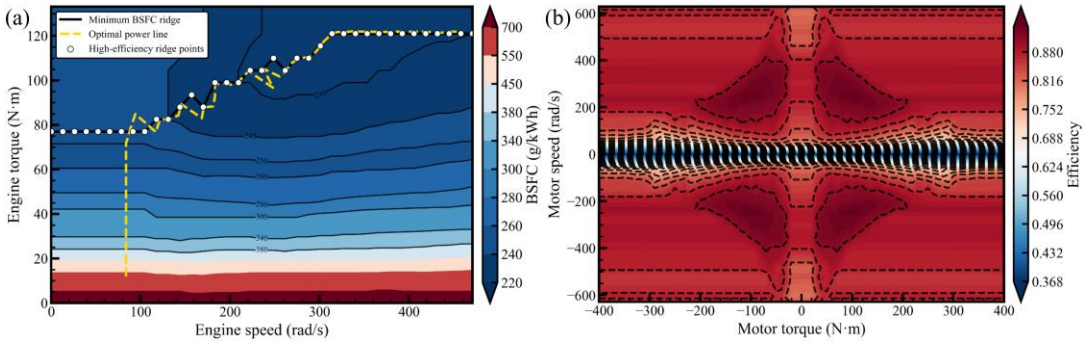


Figure 2. Characteristic maps of main powertrain components. (**a**) Engine BSFC map. (**b**) Motor efficiency map.

2.4. Electro-Thermal Modeling of Battery

Battery temperature is closely related to battery performance, aging, and safety. During charging and discharging, heat generation changes the internal temperature field and may promote the growth of the solid electrolyte interphase (SEI) layer, which affects battery lifetime and reliability [36]. In this way, real-time electro-thermal modeling is needed to support accurate EMS decision-making.

The battery model was formulated as a control-oriented coupled electro-thermal model for rapid evaluation of candidate action effects on battery states during action selection. The battery power request is used as the model input. It calculates the current response, SOC variation, polarization dynamics, heat generation, and temperature

evolution. Temperature feedback is further used to update the related electrical parameters. Based on this framework, the internal battery state is defined as

$$x_b = [SOC, U_1, U_2, \ldots, U_n, T_c, T_s]^T \tag{3}$$

where SOC denotes the battery state of charge. U_i represents the polarization voltage of the i-th RC branch (i = 1, 2, …, n). T_c and T_s denote the battery core temperature and surface temperature, respectively. For each candidate action, the model first calculates the current (I) from the corresponding battery power. The electrical states are then updated using I, and the battery heat generation rate is calculated. The temperature variation is further solved through a two-node thermal network. Finally, the updated temperature is used to correct the electrical parameters, which affects the electrical response in the next control step. Through this recursive process, the model outputs the predicted SOC, T_c and T_s associated with each candidate action within one control step.

An equivalent circuit model with n-th RC was adopted to describe the dynamic characteristics of the battery. To meet the computational requirements of real-time DRL simulation, the model order was set to (n = 1), resulting in a first-order RC equivalent circuit model. The battery current was defined as positive during discharge. Under this sign convention, the equivalent circuit structure is shown in Figure 3 [37]. The charging and discharging process of the battery can be described as follows:

$$\begin{cases} P_{\text{bat}} = IU_{\text{t}} \\ C_i \dfrac{\mathrm{d}U_i}{dt} + \dfrac{U_i}{R_i} = I \\ U_{\text{t}} = U_{\text{oc}} - IR_0 - \displaystyle\sum_{i=1}^{n} U_i, \quad i = 1, 2, \ldots, n \\ \dfrac{\mathrm{d}SOC}{\mathrm{d}t} = -\dfrac{\eta I}{C_{\text{a}}} \end{cases} \tag{4}$$

where U_{oc} is the open-circuit voltage, R_0 represents the ohmic resistance, and U_t is the terminal voltage. R_i and C_i are the resistance and capacitance of the i-th RC branch, respectively. C_a refers to the actual maximum available battery capacity, and η is the Coulombic efficiency. These electrical relationships indicate that the battery power request affects the SOC and polarization states through the current response, thereby changing the terminal voltage dynamics. This process provides the electrical basis for subsequent calculations of battery heat generation and temperature rise prediction [38].

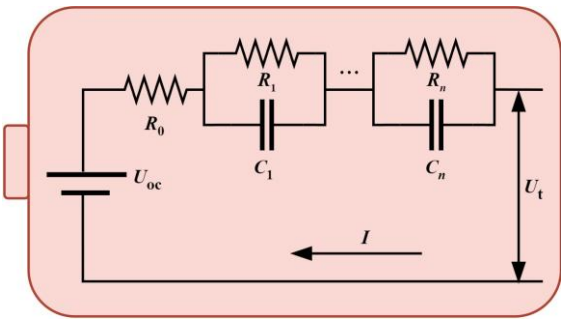


Figure 3. Equivalent circuit modeling of the lithium-ion battery.

After the current and polarization state are obtained, battery heat generation is approximated by irreversible losses. To balance physical interpretability and online computational efficiency, this study adopts a simplified heat source model consisting of ohmic heat and polarization heat:

$$Q_{\text{gen}} = I^2 R_0 + IU_1 \tag{5}$$

where I^2R_0 represents ohmic heat, and IU_1 is polarization heat. During online control, battery heat generation is approximated by ohmic and polarization-related losses, without explicitly accounting for reversible entropic heat or other heat sources. The resulting heat source term is constrained to be non-negative:

$$Q_{\text{gen}} = \max(0, I^2 R_0 + IU_1) \tag{6}$$

As a result, once a candidate action is selected, the corresponding battery power request can be determined, and the battery current response can be calculated accordingly. The heat generation rate is then obtained from the electro-thermal model. This heat source term is used as the input to the thermal state equations to predict the core and surface temperatures of the battery. To describe internal heat accumulation and surface heat dissipation, a two-node thermal network is adopted to model the thermal behavior of the battery. The battery is represented by a core node and a surface node, and the heat balance equations of the two nodes are given as follows:

$$\begin{cases} C_c \dfrac{\mathrm{d}T_{\mathrm{c}}}{\mathrm{d}t} = Q_{\text{gen}} + \dfrac{T_{\mathrm{s}} - T_{\mathrm{c}}}{R_{\mathrm{c}}} \\ C_s \dfrac{\mathrm{d}T_{\mathrm{s}}}{\mathrm{d}t} = \dfrac{T_{\mathrm{f}} - T_{\mathrm{s}}}{R_{\mathrm{s}}} - \dfrac{T_{\mathrm{s}} - T_{\mathrm{c}}}{R_{\mathrm{c}}} \end{cases} \tag{7}$$

where C_c and C_s are the equivalent heat capacities of the core and surface nodes, respectively. R_c represents the thermal resistance between the core and surface nodes, while R_s represents the thermal resistance between the surface node and the ambient environment. T_f is the ambient temperature. Consequently, the battery power request affects the SOC and terminal voltage through the current response, and it further influences the joint thermal evolution of the core and surface nodes through the heat source term.

In summary, this model represents the key physical relationships among SOC evolution, polarization response, heat generation, and temperature variation. Meanwhile, it satisfies the computational requirements of online action evaluation in reinforcement learning. However, the simplified heat-generation formulation and two-node thermal network omit reversible entropic heat and do not resolve detailed spatial temperature distributions. Accordingly, it enables one-step predictions of SOC, T_c and T_s for each candidate action.

3. Dual-Evaluation Collaborative Decision-Based Energy Management Strategy

3.1. Standard DQN Algorithm

DQN uses a deep neural network to approximate the Q-value function, thereby alleviating the curse of dimensionality encountered by conventional Q-learning in high-dimensional state spaces. By learning nonlinear state-action mappings, DQN improves policy learning efficiency and can reduce the number of training iterations to some extent [39]. Its key designs include experience replay and a fixed target network. In experience replay, transition samples (s_t, a_t, r_t, S_{t+1}) generated from agent-environment interactions are stored in a replay buffer. Mini-batches are then randomly sampled from the buffer for parameter updates. This process reduces the temporal correlation among samples and improves sample utilization and training stability [40].

During parameter optimization, DQN updates the Q-value function based on temporal difference (TD) learning. The training objective is to minimize the TD error:

$$\delta_t = r_t + \gamma \max_{a'} Q_{\theta^-}(s_{t+1}, a') - Q_\theta(s_t, a_t) \tag{8}$$

where γ is the discount factor, and θ and θ^- represent the parameters of the online network and target network, respectively. The online network is updated by minimizing the following loss function:

$$L(\theta) = \mathrm{E}[(\delta_t)^2] \tag{9}$$

The neural network parameters are updated iteratively, so that the Q-value function and the derived policy gradually approach the optimal solution. Although DQN is effective for decision-making tasks with high-dimensional state spaces, it still has several limitations in practical applications. First, overestimation bias in Q-value estimation can lead to inflated action values and reduce learning stability. Second, policy performance may fluctuate when the sample distribution shifts or the environment dynamics change. These limitations indicate that further improvement is needed in Q-value approximation accuracy and policy performance.

3.2. Principles of the KAN

To improve the approximation accuracy of the Q-value function and better distinguish discrete actions with similar values, this study incorporated a dueling value decomposition architecture into the KAN. The resulting Dueling-KAN value-function approximator was embedded into the DQN framework [41].

KAN is based on the Kolmogorov–Arnold representation theorem, which states that a continuous multivariate function can be represented by compositions of univariate functions. It assigns learnable nonlinear functions to network edges instead of using fixed activation functions at nodes, and it is different from that of the conventional MLP. This structure provides greater flexibility for approximating complex nonlinear functions [42].

The computation of the l-th KAN layer can be expressed by aggregating edge-wise nonlinear mappings:

$$x_l = \sum_{i=1}^{n_l} \phi_{l,i}(x_{l-1,i}), \quad l = 1, 2, \ldots, L \tag{10}$$

where $\phi_{l,i}(x_{l-1,i})$ is parameterized by B-spline basis functions. In contrast to multi-layer perceptrons (MLPs), which apply fixed nonlinear activation functions at nodes, KAN nodes mainly aggregate the outputs of connected edge functions. This structure allows the network to model high-dimensional nonlinear relationships with greater flexibility and provides a certain degree of interpretability.

The dueling architecture was further introduced to decompose the Q-value function into a state-value function (V (s)) and an action-advantage function (A (s, a)). A shared KAN backbone was used to extract state features. Two separate branches were then used to estimate V (s) and A (s, a), respectively. The two outputs were combined to obtain the action-value function:

$$Q(s,a) = V(s) + A(s,a) - \frac{1}{|A|} \sum_{a'} A(s,a') \tag{11}$$

This architecture separately models the value of the current state and the relative advantage of each action, thereby improving the stability of Q-value estimation. In HEV energy management, adjacent engine power actions in the discretized action space often have high similarity. The dueling structure helps the network capture fine-grained differences among these action values. In RAEP-KAN-DQN, the single-stream Q-network used

in conventional DQN is replaced with a shared KAN feature extraction layer and a dual-branch value decomposition structure.

3.3. Dual-Evaluation Collaborative Decision Framework

In the discrete engine-power control framework, the policy first outputs a raw action. This action indicates the intended engine power level. The actual engine command must satisfy practical constraints [43]. These constraints include the minimum engine on/off duration and the allowable power change between consecutive control steps. Based on this framework, a dual-evaluation energy management policy is developed. Conventional discrete-action DRL methods select actions mainly from Q-values. The proposed method introduces one-step state predictions of candidate actions into the decision-making stage. These predictions are jointly evaluated with the long-term return estimated by the value network. With this design, action selection becomes a coordinated trade-off between long-term value and short-term risk.

Specifically, at time step t, the Q-network first selects a preliminary discrete action $a_t^{\mathrm{raw}} \in \{0, 1, \ldots, N_a - 1\}$, where $N_a = 21$. This action corresponds to the intended engine power level. However, the final engine command is subject to practical constraints. Therefore, an execution-layer action mapping mechanism is applied to convert a_t^{raw} into an executable action a_t^{exec}. This mechanism is not merely used to correct infeasible actions. More importantly, it ensures that subsequent state evaluation, risk calculation, and online action selection are conducted within the physically executable action space. By this means, the decision target remains consistent with the actual control target.

For each a_t^{exec}, electro-thermal state prediction is performed. The vehicle power-flow model first calculates P_{bat}. P_{bat} is used as the input to the coupled electro-thermal battery model to estimate the battery state at the next control step:

$$x_{t+1}(a) = [SOC_{t+1}(a), T_{\mathrm{c},t+1}(a), T_{\mathrm{s},t+1}(a)] \tag{12}$$

However, these predicted variables have different physical dimensions and cannot be directly compared across candidate actions. In addition, action selection should consider not only the long-term action value estimated by the policy network, but also the short-term state cost caused by each candidate action at the current control step. Hence, the one-step predictions are further quantified as risk terms. SOC deviation, temperature load, and temperature rise impact are mapped into continuous risk indicators. A risk-aware function is then constructed for action ranking. First, the SOC deviation risk is defined as follows:

$$R_{\mathrm{soc}}(a) = w_{\mathrm{soc}}\left[\mathrm{ReLU}\left(\mathrm{SOC}_{\min} - SOC_{t+1}(a)\right)^2 + \mathrm{ReLU}\left(SOC_{t+1}(a) - \mathrm{SOC}_{\max}\right)^2\right] \tag{13}$$

where $SOC_{t+1}(a)$ is the one-step predicted SOC under candidate action (a), and w_{soc} is the weight assigned to the SOC deviation risk in the total risk function. When the predicted SOC lies within the reference interval, this risk term is set to zero. If the predicted SOC falls below the lower bound or exceeds the upper bound, the risk increases with the magnitude of the deviation. The temperature risk is then defined as follows:

$$R_{\mathrm{temp}}(a) = w_{\mathrm{temp}}\left[\mathrm{ReLU}\left(T_{\mathrm{c},t+1}(a) - T_{\mathrm{c,ref}}\right)^2 + \mathrm{ReLU}\left(T_{\mathrm{s},t+1}(a) - T_{\mathrm{s,ref}}\right)^2\right] \tag{14}$$

where $T_{\mathrm{c},t+1}(a)$ and $T_{\mathrm{s},t+1}(a)$ are the one-step predicted core and surface temperatures under candidate action (a). $T_{\mathrm{s},ref}$ reference threshold is set to 40.0 °C, referring to the commonly reported operating temperature range of lithium-ion batteries [44]. Considering that the core node in the two-node thermal model represents internal thermal accumulation and may exhibit a higher temperature than the surface node, $T_{\mathrm{c},ref}$ reference is set to 44.0 °C. The 4.0 °C difference is adopted as an engineering reference allowance in the present

model rather than a universal safety limit for NMC batteries. And w_{temp} determines the contribution of temperature risk to the total risk function. This risk term reflects the thermal load associated with a candidate action. When the predicted temperature approaches or exceeds the reference limit, the risk increases with the exceedance magnitude. To further describe the short-term thermal impact of a candidate action within one control step, the temperature rise risk is defined as follows:

$$R_{\text{rise}}(a) = w_{\text{rise}}\left[\text{ReLU}\left(T_{\text{c},t+1}(a) - T_{\text{c},t} - \Delta T_{\text{c,ref}}\right)^2 + \text{ReLU}\left(T_{\text{s},t+1}(a) - T_{\text{s},t} - \Delta T_{\text{s, ref}}\right)^2\right] \quad (15)$$

where $\Delta T_{c,ref}$ = 0.8 °C and $\Delta T_{s,ref}$ = 0.6 °C are the reference upper limits for the one-step temperature rises in the core and surface nodes, respectively. Unlike the absolute temperature thresholds, these parameters are intended to identify candidate actions that may cause relatively rapid temperature increases, thereby providing additional information for short-term thermal risk evaluation. w_{rise} determines the contribution of the temperature rise risk to the total risk function. This risk component reflects the short-term thermal variation induced by a candidate action and assigns a higher cost to actions that may lead to rapid temperature increases. It should be emphasized that these temperature thresholds serve as soft risk references rather than hard safety constraints. In practical HEVs, the BMS is responsible for battery-state monitoring and hardware-level protection, whereas the proposed EMS evaluates the potential electro-thermal consequences of candidate power-split actions at the vehicle level. Therefore, the proposed risk evaluation is intended to complement existing battery protection mechanisms by providing additional information for supervisory energy management, rather than replacing the BMS or its protection functions. An exceedance of a reference threshold increases the corresponding risk penalty but does not necessarily result in the exclusion of that candidate action. Based on the three risk components defined above, the total risk of executable candidate action (a) is expressed as follows:

$$R(a) = R_{\text{soc}}(a) + R_{\text{temp}}(a) + R_{\text{rise}}(a) \quad (16)$$

The total risk is further combined with the long-term action value (Q (s_t, a)) to construct the risk-aware utility function:

$$U(s_t, a) = Q(s_t, a) - \lambda \cdot R(a) \quad (17)$$

where λ controls the influence of the risk penalty. Accordingly, candidate actions are ranked by considering both the long-term return and the one-step state risk. It should be noted that these parameters serve as soft reference thresholds for risk evaluation rather than strict safety limits. Exceeding a reference threshold increases the corresponding risk penalty and affects the relative ranking of candidate actions, but does not automatically exclude the associated action from selection. The final decision is determined by the combined evaluation of the long-term Q-value and short-term risk.

Figure 4 presents the RAEP-KAN-DQN EMS framework, which summarizes the proposed dual-evaluation collaborative decision-making process. The framework consists of four components: execution-layer action supervision, one-step state prediction, risk quantification, and value learning. Execution-layer action supervision ensures that candidate actions satisfy practical control constraints. One-step state prediction and risk quantification provide forward-looking constraint information for action selection, while value learning estimates the long-term return. Unlike conventional strategies that rely only on value-function evaluation or reward penalties for post-execution correction, the proposed method selects actions from the executable action set by jointly considering long-term Q-value and short-term risk. Furthermore, SOC constraints, thermal safety risk, and execution consistency are incorporated directly into the decision-making stage.

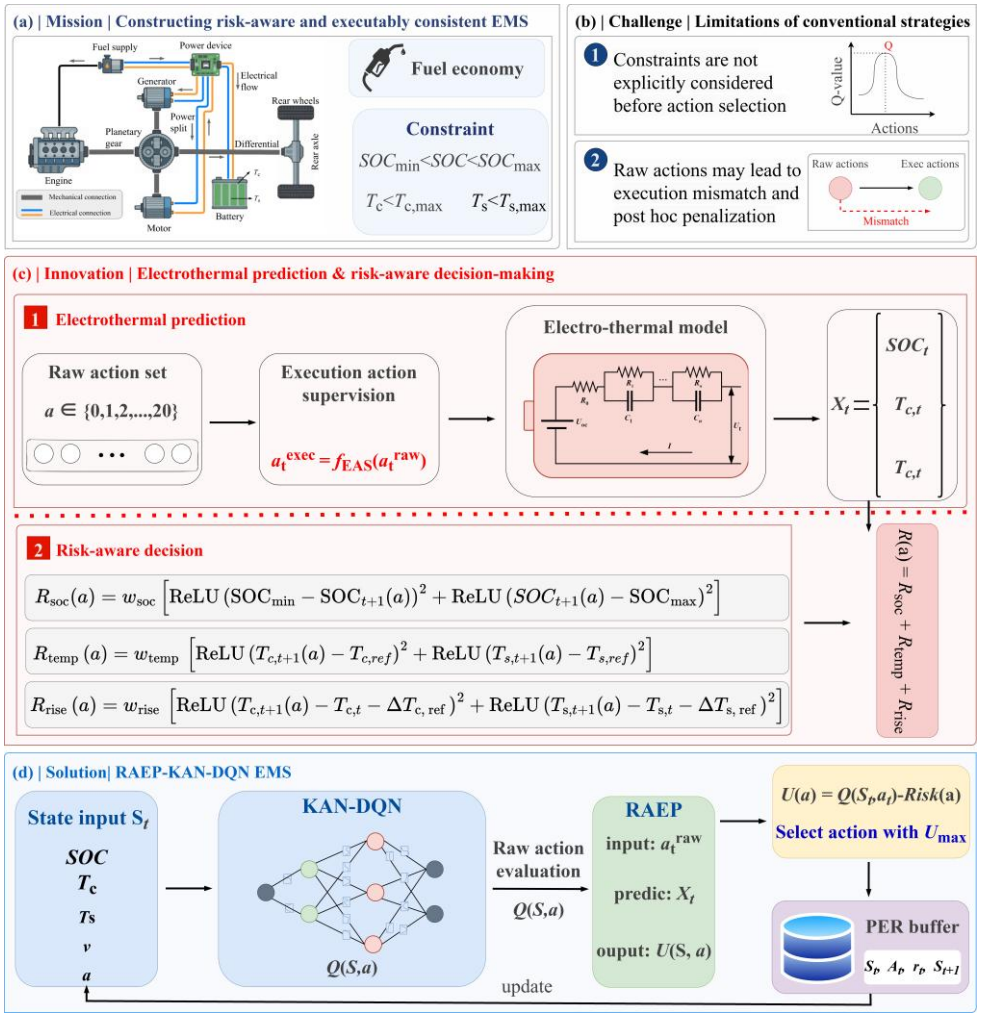


Figure 4. Agent-environment interaction framework for HEV energy management.

3.4. Formulation of the Energy Management Decision Problem

DQN, RAEP-DQN, KAN-DQN, and RAEP-KAN-DQN were compared under unified training settings to ensure experimental comparability. It is noteworthy that all algorithms were implemented under the same training settings. The main hyperparameters are listed in Table 2, and the network architecture used in KAN-DQN and RAEP-KAN-DQN is shown in Figure 5.

During training, a relatively large discount factor and multi-step returns were used to better account for long-term cumulative rewards. This setting is suitable for HEV energy management because fuel consumption and battery-state evolution are temporally correlated. Experience replay and the ε-greedy exploration strategy were applied to improve sample coverage and reduce premature convergence to local behavior patterns. Target network updates, mini-batch training, and gradient clipping were further adopted to stabilize value-function learning.

For KAN-DQN and RAEP-KAN-DQN, the same KAN-based value-network structure was used. This setting allows the performance difference between the two algorithms to mainly reflect the effect of the RAEP mechanism. Parameters were selected to balance nonlinear representation capability, model complexity, and training stability. Overall, these settings provide a consistent basis for comparing the four algorithms under the same experimental conditions.

Table 2. Hyperparameter configuration.

Parameter	Value
Training episodes	500
Discount factor	0.999
Learning rate	1×10^{-4}
Replay capacity	200,000
Warm-up size	20,000
Initial epsilon	1.0
Minimum epsilon	0.05
Epsilon decay	1×10^{-5}
Target update	100
Batch size	256
N-step horizon	20
Gradient clip	5.0
Hidden layer 1	128
Hidden layer 2	64
Grid size	4
Spline order	2
Grid epsilon	0.05
Base scale	1.0
Spline scale	0.1
Shared KAN hidden-layer widths	128, 64
Value-head architecture	Linear, 64
Advantage-head architecture	Linear, 6

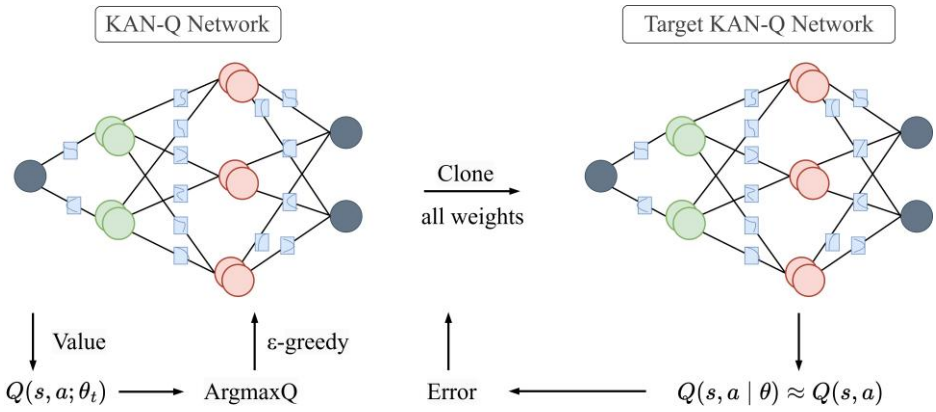


Figure 5. Algorithmic framework of KAN-DQN.

3.4.1. State Space

The state space describes the observable vehicle operating conditions and battery states used for agent decision making. Its design has an important influence on the learning performance of the EMS. The state space is defined as follows:

$$s = \{SOC, v, a, T_c, T_s\} \tag{18}$$

where v and a are the current vehicle speed and longitudinal acceleration, respectively. These two variables reflect the instantaneous power demand of the vehicle and influence the power split between the engine and the motor. SOC represents the battery state of charge, which indicates the remaining battery energy. T_c and T_s are the battery core and surface temperatures, respectively. Together, these variables describe the vehicle power

demand, battery energy state, and thermal state, providing the necessary information for EMS decision making.

3.4.2. Action Space

Engine power was selected as the control variable of the EMS. To match the discrete action space of DQN, the engine power range was discretized into a finite number of action levels. Each discrete action index is associated with a predefined engine power level. The action space and mapping to engine power are defined as follows:

$$A = \{0, 1, \ldots, N_a - 1\}, \quad N_a = 21 \tag{19}$$

$$P_{eng} \in [P_{min}, P_{max}] \tag{20}$$

where A is the discrete action set, and P_{eng} is the engine power obtained by mapping the selected discrete action to a power command. By adjusting P_{eng}, the EMS changes the power split between the engine and the battery. This enables coordinated optimization of fuel economy, SOC maintenance, and thermal constraints.

3.4.3. Reward Function

To highlight the effect of the proposed action decision mechanism, a simplified reward function was used in the main experiments. At non-terminal control steps, the reward is determined only by the fuel consumption at the current step. Thus, policy optimization mainly focuses on instantaneous fuel economy. To avoid apparent fuel savings caused by excessive battery depletion or charging, a terminal SOC equivalent fuel penalty was introduced at the end of the driving cycle. This penalty converts the deviation between the terminal SOC and the target SOC into an equivalent fuel cost, thereby constraining the battery energy state over the complete cycle. The reward function is defined as

$$r_t = \begin{cases} -m_{\text{fuel},t}, t < \mathrm{T} \\ -\left(m_{\text{fuel},t} + m_{\text{equivalent},t}\right), t = \mathrm{T} \end{cases} \tag{21}$$

where m_{fuel} is the fuel consumption cost, $m_{\text{equivalent}}$ is the equivalent fuel cost converted from the terminal SOC deviation, and t is the training step. The equivalent fuel penalty is applied only at the final step of the driving cycle. It corrects the influence of terminal battery energy deviation on the overall evaluation of EMS performance.

4. Results and Analysis

4.1. Overall Performance Under Standard Conditions

Under the standard experimental condition, the four methods showed different convergence behavior and cross-seed stability in the late training stage. As shown in Figure 6, RAEP-KAN-DQN and RAEP-DQN reached the high-reward region earlier than that of the other methods. This suggests that risk-aware action evaluation helps improve early policy learning. As training progressed, RAEP-KAN-DQN maintained a higher and more stable reward plateau. KAN-DQN converged more slowly, while DQN stayed at a lower reward level and showed larger fluctuations. Over the last 20 episodes, the average rewards of RAEP-KAN-DQN, KAN-DQN, RAEP-DQN, and DQN were −87.578, −99.995, −108.341, and −277.479, respectively. RAEP-KAN-DQN achieved the highest reward, indicating better late-stage convergence. It also showed the smallest cross-seed variation. The mean standard deviation over the last 20 episodes is only 2.266 for RAEP-KAN-DQN. However, the corresponding values are 9.876, 10.115, and 161.758 for KAN-DQN, RAEP-DQN, and DQN, which are substantially higher than that of RAEP-KAN-DQN. These

results indicate that RAEP-KAN-DQN achieves higher rewards and lower sensitivity to random initialization and experience sampling. It exhibits better training stability and convergence robustness.

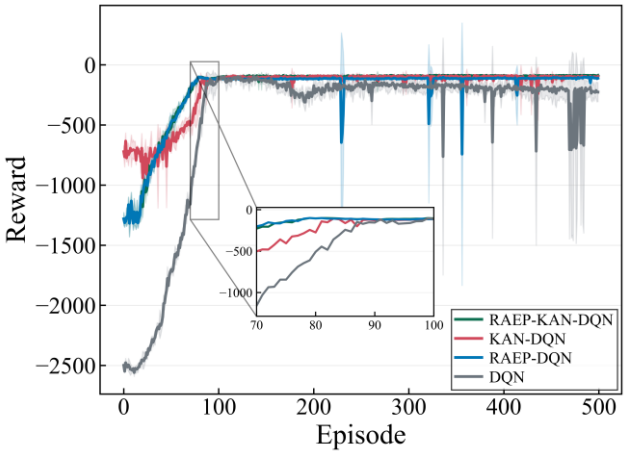


Figure 6. Training reward convergence curves.

To verify the effectiveness of the proposed RAEP-KAN-DQN, Table 3 and Figure 7 show the comparison of the performance with DQN, KAN-DQN, and RAEP-DQN under the standard ChinaCity driving cycle. The results indicate that the RAEP-KAN-DQN exhibits superior performance. Its fuel cost is 2.372 L/100 km, which is lower than those of KAN-DQN and DQN. Although RAEP-DQN achieves a lower fuel cost of 2.040 L/100 km, its terminal SOC decreases to 0.575, and the corresponding terminal SOC penalty reaches 49.286. Because the terminal SOC penalty reflects the deviation from 0.60, the lower fuel cost of RAEP-DQN is partly obtained by drawing more energy from the battery.

RAEP-KAN-DQN avoids this excessive battery-depletion tendency. As shown in Table 3, RAEP-KAN-DQN maintains its terminal SOC at 0.597. Its terminal SOC penalty is only 6.356 because the terminal SOC remains close to the charge-sustaining target of 0.60. The same tendency is observed in Figure 7a, where the SOC deviation of RAEP-KAN-DQN is 0.005. Compared with RAEP-DQN and DQN, the deviation is reduced by 87.18% and 88.10%, respectively. The results indicate that the proposed strategy maintains a favorable fuel-cost level while keeping the battery near the desired terminal energy state. Therefore, RAEP-KAN-DQN avoids improving fuel economy at the expense of terminal SOC balance.

The temperature results further support this interpretation. RAEP-KAN-DQN keeps the battery core temperature Tc and surface temperature Ts at 27.936 °C and 27.688 °C, respectively, which are the lowest among the four strategies. Such thermal behavior is important because good SOC regulation alone does not necessarily indicate effective energy management. For example, KAN-DQN achieves a slightly smaller SOC deviation of 0.004, but its Tc and Ts increase to 30.362 °C and 29.974 °C, respectively. Therefore, KAN-DQN reaches a similar terminal SOC state with a higher battery thermal burden. In contrast, RAEP-KAN-DQN maintains terminal SOC without increasing battery heat accumulation, indicating smoother energy distribution between the engine and battery.

The difference is also reflected in the engine operation shown in Figure 7b. The engine power of RAEP-KAN-DQN is concentrated in a lower and narrower range, indicating more stable engine intervention during the driving cycle. By comparison, KAN-DQN shows a wider engine power distribution. Its engine starts reach 91.183, which is about 1.93 times the value of RAEP-KAN-DQN. RAEP-DQN has fewer engine starts, but this behavior is accompanied by deeper battery depletion. Accordingly, RAEP-KAN-DQN

provides a more reasonable compromise by reducing excessive battery use and limiting frequent engine engagement.

Generally, Table 3 and Figure 7 indicate that RAEP-KAN-DQN improves the coordination among fuel cost, terminal SOC, battery thermal behavior, and engine operating smoothness. RAEP-DQN achieves better fuel economy, and KAN-DQN has a smaller SOC deviation, but both advantages are obtained with clear trade-offs. In contrast, RAEP-KAN-DQN guides the engine and battery toward a more coordinated power split, allowing the terminal energy state, thermal response, and engine behavior to remain consistent rather than being optimized in isolation. Such control behavior is more consistent with the practical objective of HEV energy management.

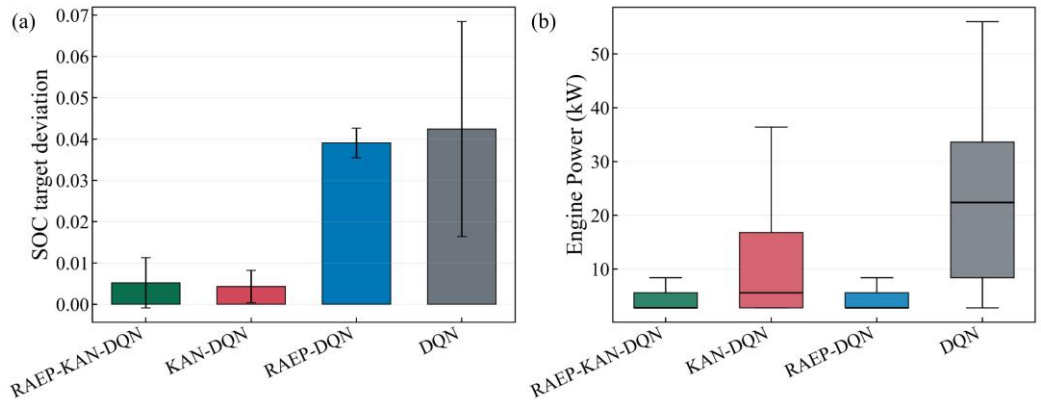


Figure 7. Comparison of performance of different strategies across 3 seeds. (**a**) SOC target deviation. (**b**) Engine power.

Table 3. Comparison of mean performance of different strategies across 3 seeds.

Strategy	Cost (L/100 km)	SOC	Tc (°C)	Ts (°C)	SOC Pen.	Starts
RAEP-KAN-DQN (ours)	2.372	0.597	27.936	27.688	6.356	47.3
KAN-DQN	2.591	0.603	30.362	29.974	6.998	91.2
RAEP-DQN	2.040	0.575	28.585	28.239	49.286	38.3
DQN	4.144	0.658	33.181	32.611	105.922	81.6

4.2. Effect of Temperature State on Battery Thermal Performance

Figure 8 shows the battery temperature rise in the four strategies under different initial temperature conditions. Here, T_c and T_s denote the battery core temperature and surface temperature, and the temperature rise is defined as the increase in terminal temperature relative to the ambient temperature. Since both experiments use the same ChinaCity driving cycle, the difference in temperature rise mainly reflects the influence of different EMSs on the power allocation between the engine and battery. At 25 °C, RAEP-KAN-DQN achieves the lowest temperature rise among the four strategies, with T_c and T_s increases of 2.903 °C and 2.672 °C. In contrast, KAN-DQN and DQN show considerably higher temperature rises, indicating that they impose a higher thermal load on the battery side. Although RAEP-DQN has lower temperature rises than KAN-DQN and DQN, its values remain higher than those of RAEP-KAN-DQN. This result indicates that a lower fuel cost does not necessarily correspond to better thermal-state control.

When the initial temperature increases to 40 °C, the difference in temperature rise becomes more pronounced. DQN shows the most evident heat accumulation, with T_c and T_s rises of 7.059 °C and 6.625 °C. By comparison, RAEP-KAN-DQN reduces the T_c and T_s rises by 62.13% and 62.66% relative to DQN. The results demonstrate that RAEP-KAN-

DQN can effectively suppress battery heat accumulation under high-temperature conditions.

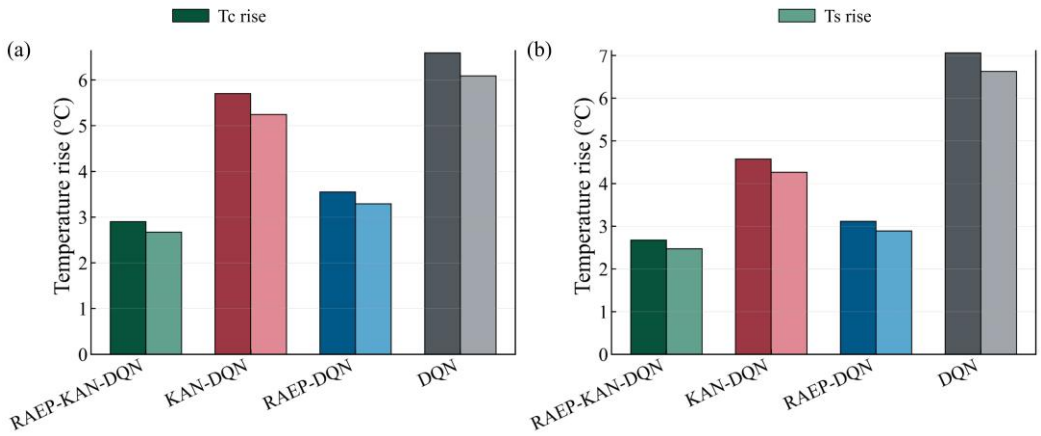


Figure 8. Temperature rise comparison of different strategies under different initial temperatures. (**a**) Temperature rise at 25 °C. (**b**) Temperature rise at 40 °C.

Table 4 and Figure 9 further reveal the trade-off between the thermal-state advantage and the fuel cost, SOC deviation, and Reward. At 25 °C, RAEP-KAN-DQN achieves a fuel cost of 2.395 L/100 km, which is slightly higher than that of RAEP-DQN but remains at a relatively low level. More importantly, its SOC deviation is only 0.002, indicating that the terminal SOC is maintained close to the charge-sustaining target of 0.60. Meanwhile, its Reward is −87.406, the highest among the four strategies, showing better overall performance in the composite objective. As shown in Figure 9, RAEP-KAN-DQN keeps the SOC trajectory closer to the target value, while the Tc and Ts curves remain at lower levels. Therefore, under the normal-temperature condition, the advantage of RAEP-KAN-DQN is not simply the pursuit of the lowest fuel cost, but its ability to achieve a more coordinated balance among low fuel cost, terminal SOC, and thermal state.

At 40 °C, the thermal-control advantage of RAEP-KAN-DQN becomes more evident. Under the elevated initial temperature condition (40 °C), RAEP-KAN-DQN achieves the lowest peak core and surface temperatures among the four strategies, reaching 42.753 °C and 42.474 °C, respectively. Compared with DQN, the core and surface temperature rises are reduced by 62.13% and 62.66%, respectively, demonstrating its effectiveness in mitigating battery heat accumulation. In terms of fuel economy, RAEP-KAN-DQN achieves a fuel cost of 2.069 L/100 km, which is substantially lower than that of DQN. Its terminal SOC deviation is 0.020, indicating a relatively balanced charge-sustaining performance. The results also show that the thermal response is influenced by the initial battery temperature and the sustained power demand during operation. Although the surface temperature exceeds its 40 °C reference value, the proposed strategy effectively limits the temperature increase compared with the other evaluated strategies. These findings demonstrate the benefit of incorporating short-term electro-thermal risk into action selection, particularly under elevated initial temperature conditions.

The dynamic trajectories in Figure 9 further confirm these results. At 25 °C, the SOC trajectory of RAEP-KAN-DQN is closest to the reference value of 0.60, and its temperature curves increase more slowly. At 40 °C, KAN-DQN tracks the SOC target more closely, but RAEP-KAN-DQN keeps T_c and T_s at lower levels throughout the cycle. As a whole, RAEP-KAN-DQN achieves better coordinated control among fuel economy, charge-sustaining performance, and thermal safety under different initial temperature conditions.

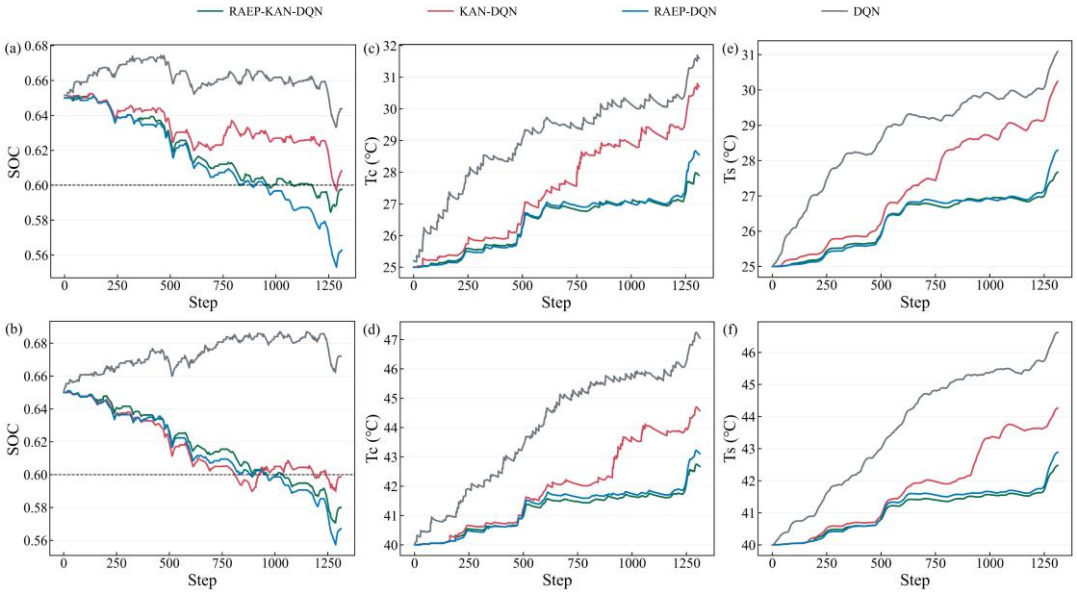


Figure 9. Comparison of SOC and temperature for different strategies under different initial temperatures. (**a**) SOC trajectories at 25 °C. (**b**) SOC trajectories at 40 °C. (**c**) Tc trajectories at 25 °C. (**d**) Tc trajectories at 40 °C. (**e**) Ts trajectories at 25 °C. (**f**) Ts trajectories at 40 °C. (the dotted lines represent the reference SOC.).

Table 4. (**a**). Quantitative comparison of performance metrics for different strategies at 25 °C. (**b**). Quantitative comparison of performance metrics for different strategies at 40 °C.

(a)					
Strategy	**Cost**	**SOC Deviation**	**Tc (max)**	**Ts (max)**	**Reward**
RAEP-KAN-DQN (ours)	2.395	0.002	27.980	27.672	−87.406
KAN-DQN	2.642	0.008	30.801	30.245	−104.176
RAEP-DQN	2.223	0.037	28.675	28.293	−106.330
DQN	3.606	0.044	31.699	31.089	−238.621
(b)					
Strategy	**Cost**	**SOC Deviation**	**Tc (max)**	**Ts (max)**	**Reward**
RAEP-KAN-DQN (ours)	2.069	0.020	42.753	42.474	−101.850
KAN-DQN	2.387	0.001	44.697	44.263	−85.666
RAEP-DQN	1.864	0.033	43.222	42.886	−110.521
DQN	4.401	0.072	47.246	46.625	−328.135

4.3. Supplementary Validation Under Different Driving Cycles

To further examine the adaptability of the proposed method to different driving cycles, the HWFET driving cycle was used for supplementary validation, as shown in Figure 10. ChinaCity and HWFET have different speed distributions, acceleration profiles, and power demand levels. Based on these two cycles, this section compares the control performance of RAEP-KAN-DQN under different load conditions.

Figure 11a and Table 4a show the engine power behavior under ChinaCity. RAEP-KAN-DQN mainly uses low-power and short-duration engine interventions. Its average number of engine starts is about 47.3. This is 49.92% lower than KAN-DQN and 49.81% lower than DQN. Thus, RAEP-KAN-DQN greatly reduces engine start-stop frequency. Figure 11c shows that the engine operating points of RAEP-KAN-DQN are concentrated in the low-to-medium power region. The distribution range is also narrower than those of KAN-DQN and DQN. This indicates moderate engine power scheduling. RAEP-KAN-DQN maintains SOC without frequent switching or high-power pulses. RAEP-DQN

achieves lower fuel consumption, but its SOC deviation and SOC penalty are much higher. This suggests stronger battery energy depletion.

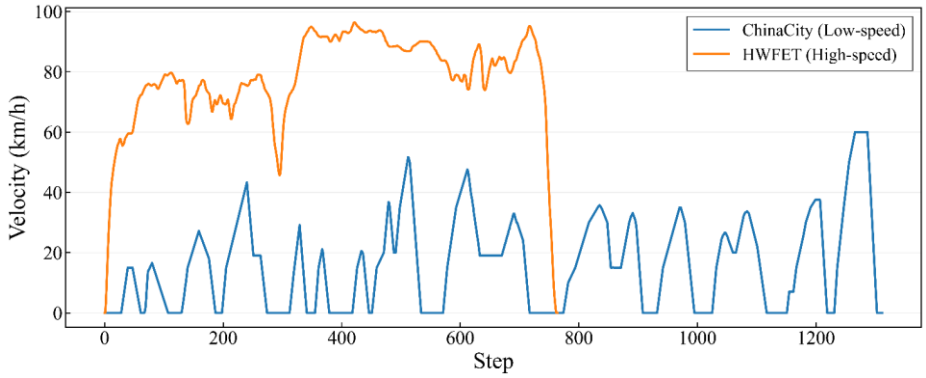


Figure 10. Speed profiles of the ChinaCity and HWFET.

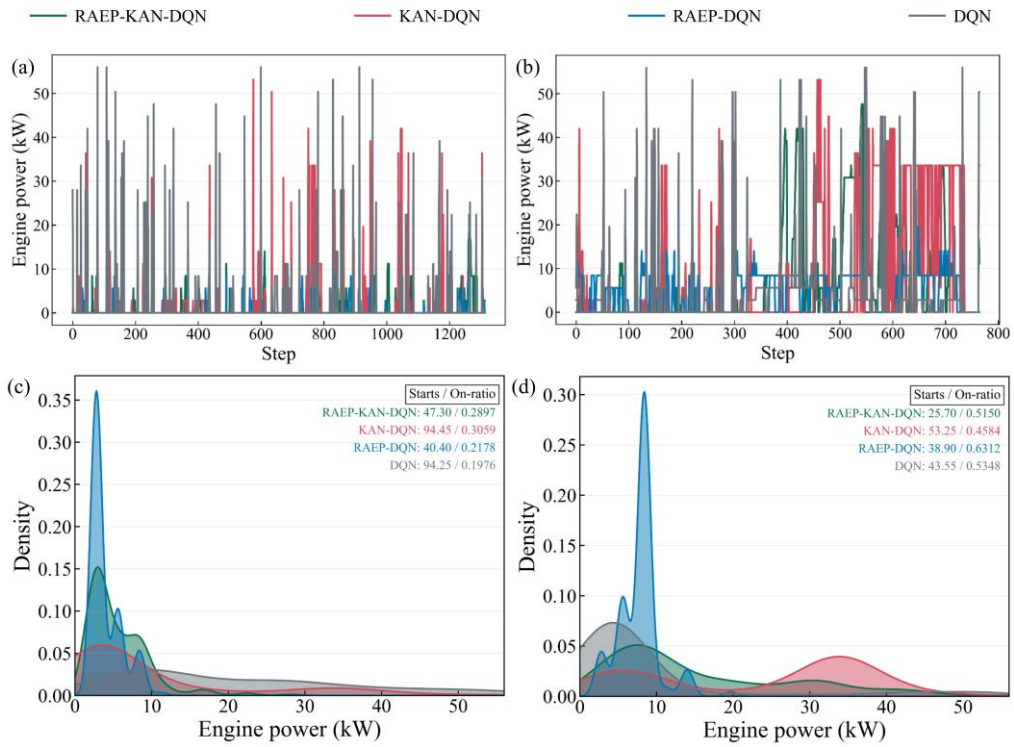


Figure 11. Engine power trajectories and distribution characteristics under different driving cycles. (**a**) Engine power trajectory (ChinaCity). (**b**) Engine power trajectory (HWFET). (**c**) engine power density distribution (ChinaCity). (**d**) Engine power density distribution (HWFET).

RAEP-KAN-DQN is additionally validated on the HWFET driving cycle to assess its robustness across different driving cycles. As shown in Table 5, RAEP-KAN-DQN does not achieve the lowest fuel cost under HWFET. Its fuel cost is 3.439 L/100 km, but it is 15.42% lower than that of KAN-DQN. This result indicates that RAEP-KAN-DQN avoids the fuel penalty caused by the excessive engine load of KAN-DQN in the high-speed scenario. Meanwhile, its SOC deviation is 0.111. Although this value is higher than that of KAN-DQN, it is clearly lower than those of RAEP-DQN and DQN, indicating that RAEP-KAN-DQN does not rely on deep battery depletion to meet the high-speed power demand. In addition, its T_c (max) and T_s (max) are 20.87% and 19.73% lower than those of KAN-DQN, respectively. Compared with these different strategies, RAEP-KAN-DQN achieves a more balanced distribution of fuel consumption, terminal SOC deviation, and battery

thermal loading. The observed temperature values also indicate that the current risk-weighted decision mechanism reduces the relative preference for thermally unfavorable actions without strictly excluding all actions associated with temperature reference exceedance.

Figure 11b further explains the origin of this behavior. Under the HWFET driving cycle, RAEP-KAN-DQN no longer uses the scattered and short-duration engine interventions observed under the urban driving cycle. Instead, it forms a longer-duration engine output band with a moderate power level. This pattern indicates that the strategy allows the engine to meet the continuous power demand during high-speed cruising, rather than relying on frequent engine switching or large battery compensation. Its average number of engine starts is only 25.70, which is 51.74%, 33.93%, and 40.99% lower than those of KAN-DQN, RAEP-DQN, and DQN, respectively. This further indicates that its performance improvement is not achieved by increasing the engine switching frequency.

In summary, RAEP-KAN-DQN adapts its engine power allocation to different driving conditions. Under ChinaCity, it uses low-intensity power scheduling with fewer engine starts. Under HWFET, it shifts to more continuous medium-power output.

Table 5. HWFET high-speed condition.

Strategy	Cost	SOC Deviation	SOC Pen.	Tc (max)	Ts (max)	Reward
RAEP-KAN-DQN (ours)	3.439	0.111	173.908	44.180	42.807	−529.407
KAN-DQN	4.066	0.032	57.248	55.833	53.329	−526.791
RAEP-DQN	2.844	0.195	348.773	36.915	35.976	−598.603
DQN	2.909	0.243	382.571	47.787	45.881	−633.342

4.4. Sensitivity Analysis of Risk Weight

Table 6 and Figure 12 present the sensitivity analysis of RAEP-KAN-DQN under different risk weights λ. Overall, λ does not affect fuel consumption and terminal SOC in a monotonic manner. Instead, it changes the relative contribution of the long-term Q-value and one-step risk in action evaluation, thereby adjusting the power allocation between the engine and the battery.

In terms of fuel consumption, as shown in Figure 12 and Table 6, Fuel changes only slightly in the low-to-medium risk-weight range. When λ increases from 1 to 16, Fuel remains around 0.084 kg, with only 2.4% decrease at $\lambda = 4$. This indicates that moderate risk intervention does not noticeably weaken fuel economy. When $\lambda = 32$, Fuel decreases to 0.078 kg, which is the lowest value among all settings. It is about 4.9% and 7.1% lower than those at $\lambda = 4$ and $\lambda = 16$, respectively. However, this does not mean that $\lambda = 32$ provides the best overall performance, because its SOC penalty also increases to 11.335, which is about 118.1% higher than that at $\lambda = 16$. This suggests that the lowest Fuel is mainly accompanied by greater battery energy consumption, rather than being solely caused by improved engine operating efficiency.

The terminal SOC results further reveal this trade-off. When $\lambda = 1$, the SOC penalty is the lowest, at 3.713, and the terminal SOC is 0.599, indicating the best terminal-state maintenance. When $\lambda = 4$ and $\lambda = 16$, the terminal SOC values are 0.598 and 0.601, respectively. Both remain close to the target value, suggesting that this range provides a better balance between fuel economy and charge-sustaining performance. In contrast, although $\lambda = 32$ achieves the lowest Fuel, its SOC penalty increases markedly. When $\lambda = 64$, the SOC penalty decreases to 8.283, but the terminal SOC does not recover substantially. Therefore, increasing λ does not consistently improve SOC constraint satisfaction. Excessive risk intervention may instead disturb the coordination between long-term return and one-step risk.

The control behavior indicators explain this phenomenon. Across different λ values, the maximum difference in Starts is only 3.4, and the On-ratio varies only from 0.261 to 0.290. This indicates that the performance differences are not mainly caused by changes in engine start-stop frequency. In contrast, the change in average power allocation is more important. When λ = 16, the average engine power is 0.908 kW, higher than 0.872 kW at λ = 4. The average battery power decreases from 0.564 kW to 0.523 kW. This shows that the strategy tends to rely more on the engine to support the load, thereby reducing the battery burden and maintaining a higher SOC. Conversely, when λ = 32 and λ = 64, the average engine power decreases to 0.847 kW and 0.842 kW, while the average battery power increases to 0.572 kW and 0.575 kW. This change indicates that the lower Fuel values in these two settings rely more on battery power compensation, which explains the increase in SOC penalty.

Taken together, λ does not simply strengthen the constraint effect. It changes the decision preference among fuel economy, charge-sustaining performance, and power allocation. In this study, λ = 4 and λ = 16 form relatively stable balance range. Specifically, λ = 4 achieves the best total reward of −87.406, making it suitable as the default risk weight in the following experiments. In comparison, λ = 16 shows better performance in terminal SOC maintenance and battery-load reduction. Although λ = 32 and λ = 64 further reduce Fuel, they mainly depend on higher battery participation and lead to larger terminal SOC deviations. Therefore, they are not selected as the optimal settings for comprehensive performance.

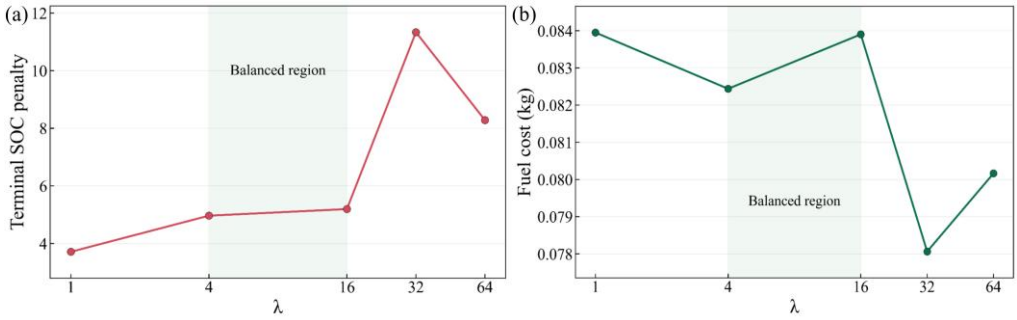


Figure 12. Key metrics under different λ settings. (**a**) Terminal SOC penalty. (**b**) Fuel consumption.

Table 6. Sensitivity analysis of RAEP-KAN-DQN under different risk weights.

λ	SOC	SOC Pen.	Starts	On-Ratio	Mean Eng. Power	Mean Bat. Power
1	0.599	3.713	47.7	0.276	0.887	0.548
4	0.598	4.967	47.3	0.290	0.872	0.564
16	0.601	5.196	47.0	0.284	0.908	0.523
32	0.596	11.335	48.8	0.274	0.847	0.572
64	0.596	8.283	45.4	0.261	0.842	0.575

4.5. Computational Cost Analysis

Table 7 presents the computational cost of the four strategies. The average decision times of DQN, KAN-DQN, RAEP-DQN, and RAEP-KAN-DQN are 0.108, 0.753, 6.175, and 7.476 ms per step, respectively. Although RAEP-KAN-DQN requires more computation than the other methods, its average decision time remains below 8 ms, with a P95 latency of 14.778 ms and an observed maximum of 24.699 ms. This indicates that the additional computational overhead introduced by executable-action mapping and one-step electro-thermal risk evaluation remains within a relatively manageable range under the tested computational environment. Nevertheless, further validation on automotive-grade hardware is required to confirm its compatibility with practical vehicle control cycles.

Table 7. Computational cost comparison of different strategies (unit: ms/step).

Strategy	Network Inference	Decision Time	P95	Max
DQN	0.085	0.108	0.185	0.714
KAN-DQN	0.718	0.753	1.056	2.980
RAEP-DQN	0.184	6.175	10.750	23.175
RAEP-KAN-DQN (ours)	0.849	7.476	14.778	24.699

5. Conclusions

To address the limitation of discrete-action DRL-based EMSs, in which action selection is mainly governed by long-term Q-values while battery SOC and thermal constraints are handled mostly through ex-post penalty terms, this study proposes RAEP-KAN-DQN and develops a dual-evaluation collaborative decision-making strategy. The proposed method combines short-term electro-thermal consequence evaluation, the KAN network, and dueling value decomposition. By evaluating candidate actions using both long-term return and short-term risk, battery SOC and thermal constraints can directly participate in the decision-making process before action execution.

RAEP-KAN-DQN shows multi-objective trade-off performance under different test conditions. Under the standard condition, RAEP-KAN-DQN reduces the fuel cost by 33.58% compared with DQN. It also decreases the peak battery core and surface temperatures by 11.73% and 10.99%, respectively. Although RAEP-KAN-DQN does not outperform RAEP-DQN and KAN-DQN in every single metric, it achieves a more balanced trade-off among fuel cost, terminal SOC, battery temperature level, and engine start-stop frequency.

To evaluate robustness to temperature variation, the initial temperature was set to 40 °C. Under this condition, the fuel cost of DQN increases by 22.05%, whereas that of RAEP-KAN-DQN decreases by 13.61%. T_c and T_s rises in RAEP-KAN-DQN are 2.673 °C and 2.474 °C, respectively, which are 62.13% and 62.66% lower than those of DQN. These results demonstrate the improved capability of RAEP-KAN-DQN to mitigate battery temperature rise under high-temperature simulation conditions while maintaining a reasonable fuel cost level.

To assess adaptability across different driving cycles, the HWFET driving cycle was introduced as supplementary validation. RAEP-KAN-DQN achieves fuel cost of 3.439 L/100 km, which is 15.42% lower than that of KAN-DQN. Although RAEP-KAN-DQN does not achieve the lowest fuel cost under HWFET, it provides better stability in terminal SOC and battery thermal safety. Overall, these results indicate that short-term electro-thermal consequence evaluation enables RAEP-KAN-DQN to account for potential thermal risks before action execution, thereby achieving better overall trade-off performance than conventional DQN under the simulated test conditions.

Future work will focus on two aspects. (1) The current study is mainly based on simulation. Hardware-in-the-loop tests or vehicle experiments will be conducted to further verify robustness under real execution delays, measurement noise, and model mismatch. (2) The KAN network structure and training process will be further optimized to reduce computational complexity and support deployment in resource-limited environments.

Author Contributions: Conceptualization, C.W. and J.L.; methodology, C.W., J.L. and Q.Z.; software, Q.Z. and H.J.; validation, H.J. and H.C.; formal analysis, Q.Z. and H.J.; investigation, H.C.; resources, C.W. and J.L.; data curation, Q.Z.; writing—original draft preparation, Q.Z.; writing—review and editing, C.W., J.L., H.J. and H.C.; visualization, Q.Z. and H.J.; supervision, C.W. and J.L.; project administration, C.W. and J.L.; funding acquisition, C.W. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Grant No. 52677269, 52407246); Sichuan Provincial Natural Science Foundation (Youth Science Fund, Category B) (Grant No. 2026NSFSC1314); Yibin Municipal Science and Technology Program (Grant No. 2025MZ012); Guang-dong Basic and Applied Basic Research Foundation (Grant No. 2025A1515010354) and Graduate Innovation Foundation of Sichuan University of Science & Engineering (Grant No. Y2025034). The systemic experiments are performed at the Advanced Energy Storage and Application (AESA) Group, Beijing Institute of Technology.

Data Availability Statement: The original data can be obtained from the corresponding author.

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

Abbreviations

ACC	Adaptive cruise control
ATSAC	Adaptive twin soft actor-critic
Bat.	Battery
BSFC	Brake specific fuel consumption
DDPG	Deep deterministic policy gradient
DQL	Deep Q-learning
DQL-H	Hierarchical deep Q-learning
DQN	Deep Q-network
DRL	Deep reinforcement learning
EMS	Energy management system
Eng.	Engine
HEV	Hybrid electric vehicle
HWFET	Highway fuel economy test
KAN	Kolmogorov–Arnold network
MPC	Model predictive control
MLP	Multi-layer perceptron
NMC	Nickel manganese cobalt
Pen.	Penalty
RAEP-KAN-DQN	Risk-Aware Electrothermal Prediction- KAN-DQN
RC	Resistor-capacitor
SEI	Solid electrolyte interphase
SOC	State of charge
TD	Temporal difference
TD3	Twin delayed deep deterministic policy gradient

References

1. Wang, Y.; Wu, J.; He, H.; Wei, Z.; Sun, F. Data-driven energy management for electric vehicles using offline reinforcement learning. *Nat. Commun.* **2025**, *16*, 2835. https://doi.org/10.1038/s41467-025-58192-9.
2. Li, W.; Li, S.; Du, C.; Yan, F. Cross-type transfer of fuel cell hybrid vehicles energy management strategies based on deep reinforcement learning. *Energy Convers. Manag.* **2026**, *356*, 121371. https://doi.org/10.1016/j.enconman.2026.121371.
3. Huang, H.; Lin, X.; Huang, Q. Enhancing adaptive health-conscious energy management control strategy with self-learning power correction for fuel cell hybrid electric vehicles. *Appl. Energy* **2026**, *402*, 127024. https://doi.org/10.1016/j.apenergy.2025.127024.
4. Dong, P.; Zhao, J.; Liu, X.; Wu, J.; Xu, X.; Liu, Y.; Wang, S.; Guo, W. Practical application of energy management strategy for hybrid electric vehicles based on intelligent and connected technologies: Development stages, challenges, and future trends. *Renew. Sustain. Energy Rev.* **2022**, *170*, 112947. https://doi.org/10.1016/j.rser.2022.112947.
5. Milojević, S.; Skrucany, T.; Milošević, H.; Stanojević, D.; Pantić, M.; Stojanović, B. Alternative drive systems and environmentally friendly public passengers transport. *Appl. Eng. Lett.* **2018**, *3*, 105–113. https://doi.org/10.18485/aeletters.2018.3.3.4.
6. Munsi, M.S.; Chaoui, H. Energy Management Systems for Electric Vehicles: A Comprehensive Review of Technologies and Trends. *IEEE Access* **2024**, *12*, 60385–60403. https://doi.org/10.1109/ACCESS.2024.3371483.

7. Ghani, A.A.; Abdullah, N.; B, K.; Saidur, R.; Pandey, A.K. Assessing technological-driven challenges and policies associated with electric vehicle (EV) adoption. *Transp. Policy* **2025**, *169*, 167–177. https://doi.org/10.1016/j.tranpol.2025.04.030.
8. Saiteja, P.; Ashok, B. Critical review on structural architecture, energy control strategies and development process towards optimal energy management in hybrid vehicles. *Renew. Sustain. Energy Rev.* **2022**, *157*, 112038. https://doi.org/10.1016/j.rser.2021.112038.
9. Ghosh, S.; Mukhopadhyay, S. Adaptive ECMS for trip level energy management of HEVs considering vehicle and route parameter variations. *Appl. Energy* **2025**, *399*, 126374. https://doi.org/10.1016/j.apenergy.2025.126374.
10. Chen, J.; Kuboyama, T.; Shen, T. Collective behavior information-based design approach to energy management strategy for large-scale population of HEVs. *Appl. Energy* **2025**, *377*, 124530. https://doi.org/10.1016/j.apenergy.2024.124530.
11. Shi, D.; Pisu, P.; Chen, L.; Wang, S.; Wang, R. Control design and fuel economy investigation of power split HEV with energy regeneration of suspension. *Appl. Energy* **2016**, *182*, 576–589. https://doi.org/10.1016/j.apenergy.2016.08.034.
12. Khosravi, N.; Dowlatabadi, M.; Abdelghany, M.B.; Tostado-Véliz, M.; Jurado, F. Enhancing battery management for HEVs and EVs: A hybrid approach for parameter identification and voltage estimation in lithium-ion battery models. *Appl. Energy* **2024**, *356*, 122364. https://doi.org/10.1016/j.apenergy.2023.122364.
13. Wei, Z.; Xu, J.; Halim, D. HEV power management control strategy for urban driving. *Appl. Energy* **2017**, *194*, 705–714. https://doi.org/10.1016/j.apenergy.2016.10.023.
14. Wang, H.; He, H.; Bai, Y.; Yue, H. Parameterized deep Q-network based energy management with balanced energy economy and battery life for hybrid electric vehicles. *Appl. Energy* **2022**, *320*, 119270. https://doi.org/10.1016/j.apenergy.2022.119270.
15. Jui, J.J.; Ahmad, M.A.; Molla, M.M.I.; Rashid, M.I.M. Optimal energy management strategies for hybrid electric vehicles: A recent survey of machine learning approaches. *J. Eng. Res.* **2024**, *12*, 454–467. https://doi.org/10.1016/j.jer.2024.01.016.
16. Zhu, Y.; Li, X.; Liu, Q.; Li, S.; Xu, Y. A comprehensive review of energy management strategies for hybrid electric vehicles. *Mech. Sci.* **2022**, *13*, 147–188. https://doi.org/10.5194/ms-13-147-2022.
17. Gan, J.; Li, S.; Wei, C.; Deng, L.; Tang, X. Intelligent Learning Algorithm and Intelligent Transportation-Based Energy Management Strategies for Hybrid Electric Vehicles: A Review. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 10345–10361. https://doi.org/10.1109/TITS.2023.3283010.
18. Wang, Z.; He, H.; Peng, J.; Chen, W.; Wu, C.; Fan, Y.; Zhou, J. A comparative study of deep reinforcement learning based energy management strategy for hybrid electric vehicle. *Energy Convers. Manag.* **2023**, *293*, 117442. https://doi.org/10.1016/j.enconman.2023.117442.
19. Huang, R.; He, H.; Su, Q.; Härtl, M.; Jaensch, M. Enabling cross-type full-knowledge transferable energy management for hybrid electric vehicles via deep transfer reinforcement learning. *Energy* **2024**, *305*, 132394. https://doi.org/10.1016/j.energy.2024.132394.
20. Ruan, S.; Ma, Y.; Yang, N.; Yan, Q.; Xiang, C. Multiobjective optimization of longitudinal dynamics and energy management for HEVs based on nash bargaining game. *Energy* **2023**, *262*, 125422. https://doi.org/10.1016/j.energy.2022.125422.
21. Estrada, P.M.; de Lima, D.; Bauer, P.H.; Mammetti, M.; Bruno, J.C. Deep learning in the development of energy Management strategies of hybrid electric Vehicles: A hybrid modeling approach. *Appl. Energy* **2023**, *329*, 120231. https://doi.org/10.1016/j.apenergy.2022.120231.
22. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. https://doi.org/10.1038/nature14236.
23. Qi, C.; Zhu, Y.; Song, C.; Yan, G.; Xiao, F.; Zhang, X.; Cao, J.; Song, S. Hierarchical reinforcement learning based energy management strategy for hybrid electric vehicle. *Energy* **2022**, *238*, 121703. https://doi.org/10.1016/j.energy.2021.121703.
24. Zhang, H.; Peng, J.; Dong, H.; Tan, H.; Ding, F. Hierarchical reinforcement learning based energy management strategy of plug-in hybrid electric vehicle for ecological car-following process. *Appl. Energy* **2023**, *333*, 120599. https://doi.org/10.1016/j.apenergy.2022.120599.
25. Lian, R.; Peng, J.; Wu, Y.; Tan, H.; Zhang, H. Rule-interposing deep reinforcement learning based energy management strategy for power-split hybrid electric vehicle. *Energy* **2020**, *197*, 117297. https://doi.org/10.1016/j.energy.2020.117297.
26. Liu, R.; Wang, C.; Tang, A.; Zhang, Y.; Yu, Q. A twin delayed deep deterministic policy gradient-based energy management strategy for a battery-ultracapacitor electric vehicle considering driving condition recognition with learning vector quantization neural network. *J. Energy Storage* **2023**, *71*, 108147. https://doi.org/10.1016/j.est.2023.108147.
27. Ma, X.; Liu, H.; Han, L.; Yang, N.; Li, M. An real-time intelligent energy management based on deep reinforcement learning and model predictive control for hybrid electric vehicles considering battery life. *Energy* **2025**, *324*, 135931. https://doi.org/10.1016/j.energy.2025.135931.

28. Liu, Z.E.; Li, Y.; Zhou, Q.; Li, Y.; Shuai, B.; Xu, H.; Hua, M.; Xu, L.; Tan, G. Real-time energy management for HEV combining naturalistic driving data and deep reinforcement learning with high generalization. *Appl. Energy* **2025**, *377*, 124350. https://doi.org/10.1016/j.apenergy.2024.124350.
29. Sun, H.; Li, J.; Cheng, C.; Shi, S.; Wang, J.; Lin, J.; Liu, Y. Health- and behavior-aware energy management strategy for fuel cell hybrid electric vehicles based on parallel deep deterministic policy gradient learning. *Eng. Appl. Artif. Intell.* **2025**, *158*, 111311. https://doi.org/10.1016/j.engappai.2025.111311.
30. Biswas, A.; Acquarone, M.; Wang, H.; Miretti, F.; Misul, D.A.; Emadi, A. Safe reinforcement learning for energy management of electrified vehicle with novel physics-informed exploration strategy. *IEEE Trans. Transp. Electrif.* **2024**, *10*, 9814–9828. https://doi.org/10.1109/TTE.2024.3361462.
31. Dong, H.; Hu, Q.; Li, D.; Li, Z.; Song, Z. Predictive Battery Thermal and Energy Management for Connected and Automated Electric Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 2144–2156. https://doi.org/10.1109/TITS.2024.3494734.
32. Xu, J.; Li, Z.; Gao, L.; Ma, J.; Liu, Q.; Zhao, Y. A Comparative Study of Deep Reinforcement Learning-based Transferable Energy Management Strategies for Hybrid Electric Vehicles. In *Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV)*; IEEE: New York, NY, USA, 2022; pp. 470–477. https://doi.org/10.1109/IV51971.2022.9827042.
33. Zeng, Q.; Wang, C.; Lu, J. Adaptive entropy soft actor-critic for hybrid electric vehicle energy management with battery degradation and thermal considerations. *J. Energy Storage* **2026**, *154*, 121148. https://doi.org/10.1016/j.est.2026.121148.
34. Wang, C.; Liu, R.; Tang, A. Energy management strategy of hybrid energy storage system for electric vehicles based on genetic algorithm optimization and temperature effect. *J. Energy Storage* **2022**, *51*, 104314. https://doi.org/10.1016/j.est.2022.104314.
35. Wang, H.; Ye, Y.; Zhang, J.; Xu, B. A comparative study of 13 deep reinforcement learning based energy management methods for a hybrid electric vehicle. *Energy* **2023**, *266*, 126497. https://doi.org/10.1016/j.energy.2022.126497.
36. Tang, S.; Liang, Y.; Zhong, C.; Peng, Y.; Hu, Y.; Hu, W.; Liao, Y.; Lin, J.; Yang, X.; Zhang, H.; et al. Revisiting the overdischarge process as a novel accelerated aging method for $LiFePO_4$/Graphite batteries through the unveiling of SEI evolution mechanism. *Energy Storage Mater.* **2025**, *74*, 103916. https://doi.org/10.1016/j.ensm.2024.103916.
37. Wang, J.; Zhou, J.; Zhao, W. Deep reinforcement learning based energy management strategy for fuel cell/battery/supercapacitor powered electric vehicle. *Green Energy Intell. Transp.* **2022**, *1*, 100028. https://doi.org/10.1016/j.geits.2022.100028.
38. Wang, M.; Jia, P.; Wei, W.; Xie, Z.; Chen, J.; Dong, H. Electro-thermal coupling modeling of energy storage station considering battery physical characteristics. *Front. Energy Res.* **2024**, *12*, 1433797. https://doi.org/10.3389/fenrg.2024.1433797.
39. Fan, X.; Guo, L.; Hong, J.; Wang, Z.; Chen, H. Constrained Hierarchical Hybrid Q-Network for Energy Management of HEVs. *IEEE Trans. Transp. Electrif.* **2024**, *10*, 9579–9590. https://doi.org/10.1109/TTE.2024.3353765.
40. Sun, H.; Tao, F.; Fu, Z.; Gao, A.; Jiao, L. Driving-Behavior-Aware Optimal Energy Management Strategy for Multi-Source Fuel Cell Hybrid Electric Vehicles Based on Adaptive Soft Deep-Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 4127–4146. https://doi.org/10.1109/TITS.2022.3233564.
41. Guo, X.; Yan, X.; Chen, Z.; Meng, Z. Research on energy management strategy of heavy-duty fuel cell hybrid vehicles based on dueling-double-deep Q-network. *Energy* **2022**, *260*, 125095. https://doi.org/10.1016/j.energy.2022.125095.
42. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljacic, M.; Hou, T.; Tegmark, M. KAN: Kolmogorov-Arnold Networks. *arXiv* **2024**, arXiv:2404.19756.
43. Ju, F.; Murgovski, N.; Zhuang, W.; Hu, X.; Song, Z.; Wang, L. Predictive energy management with engine switching control for hybrid electric vehicle via ADMM. *Energy* **2023**, *263*, 125971. https://doi.org/10.1016/j.energy.2022.125971.
44. Wang, T.; Tseng, K.J.; Zhao, J.; Wei, Z. Thermal investigation of lithium-ion battery module with different cell arrangement structures and forced air-cooling strategies. *Appl. Energy* **2014**, *134*, 229–238. https://doi.org/10.1016/j.apenergy.2014.08.013.

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.