drones

MDPI

Article

Soft Actor-Critic-Based Power Optimization Method for UAV Wireless Charging Systems

Zhuoyue Dai [1,2], Yongmin Yang [1,2,*], Yanting Luo [1,2,*], Zhilong Lin [1,2] and Guanpeng Yang [1,2]

[1] National Key Laboratory of Equipment State Sensing and Smart Support, National University of Defense Technology, Changsha 410073, China
[2] College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China
* Correspondence: yangyongmin@163.com (Y.Y.); luoyanting0717@126.com (Y.L.)

Highlights

What are the main findings?

- A reinforcement learning-based power optimization framework is proposed for UAV wireless charging systems to mitigate power degradation caused by landing position variations.
- The learned SAC controller maintains strong performance in both the training and expanded evaluation regions and remains effective under measurement noise and model mismatch, with its practical feasibility supported by hardware validation.

What are the implications of the main findings?

- Measurement-based current observations can support power optimization without explicit online mutual-inductance identification.
- The proposed framework provides a practical data-driven alternative to model-dependent WPT optimization in position-uncertain UAV charging scenarios.

Abstract

Maintaining high power delivery under uncertain landing positions is a key challenge for wireless charging of unmanned aerial vehicles (UAVs). This paper presents a data-driven power optimization method based on the Soft Actor-Critic algorithm for multi-transmitter single-receiver wireless power transfer (MTSR-WPT) systems. To support effective learning without explicit online parameter identification, a physics-informed dual-current state representation is constructed from measurable current responses, combining a zero-phase current with the current response under the applied phase command. The agent is trained using a reward defined directly from normalized load power, and the transmitter voltage phases serve as the control actions. In simulations of a five-transmitter system, the learned policy achieves about 97% of the theoretical maximum power in the training region and about 96% in the expanded evaluation region. Additional robustness studies show strong performance under moderate measurement noise and substantial recovery under model mismatch after short fine-tuning. Experimental validation on a physical prototype confirms the effectiveness of the method, yielding an average power improvement of 188% from a zero-phase baseline and reaching 87% of the maximum power measured on the hardware platform. These results support the proposed method as a practical data-driven alternative to model-dependent MTSR-WPT power optimization for UAV wireless charging.

Keywords: UAV wireless charging; soft actor-critic; dual-current state; power optimization; multi-transmitter single-receiver

Check for updates

Academic Editor: Rosdiadee Nordin

Received: 9 February 2026
Revised: 7 March 2026
Accepted: 17 March 2026
Published: 19 March 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

Addressing the limited endurance of unmanned aerial vehicles (UAVs) is a critical challenge, for which wireless charging presents a key solution. Magnetically coupled resonant wireless power transfer (WPT) technology, which enables reliable, contactless energy transmission, presents a viable solution and has found applications in diverse areas including consumer electronics [1–3], implantable medical devices [4,5], and electric vehicles [6–9]. Related low-power and energy-harvesting-assisted approaches have also been explored to improve power sustainability in embedded systems [10]. For UAV applications, a fundamental challenge lies in ensuring consistently high power delivery despite the positional uncertainty inherent in UAV landing.

In WPT systems, the transfer power is a core performance metric. For systems with a single transmitter, various methods for power optimization have been developed, such as coupling coefficient identification combined with impedance matching to achieve maximum power transfer tracking under varying loads [11], the use of LCL-N compensation topology on the transmitter side to enable a compact, uncompensated receiver while maintaining high-power delivery under strong coupling [12], and a hybrid MCR/PT operation scheme that extends the dynamic power range and improves robustness against coupling variations through mode-switching and current-mode modulation [13]. However, the spatial control over the magnetic field offered by a single transmitter is inherently limited, resulting in significant transfer power degradation due to coil misalignment. To overcome this, multi-transmitter single-receiver (MTSR) WPT systems have attracted significant interest [14–17]. By coordinating multiple sources, MTSR-WPT systems can actively shape the magnetic field, a capability that is particularly advantageous for creating larger effective charging areas. This feature makes them a promising candidate for applications with variable receiver positions, such as UAV wireless charging.

However, the multi-transmitter structure increases system complexity due to enhanced inter-coil couplings, which in turn raises the difficulty of designing effective optimization methods. Existing research on power optimization for MTSR-WPT systems often focuses on tuning the amplitudes of the transmitter currents. A typical strategy involves matching these amplitudes to the ratios of the mutual inductances between the transmitters and the receiver. Representative examples include maximum-efficiency point tracking [18,19] and magnetic-field-editing-based maximum-power tracking methods [20]. A critical prerequisite for this typical strategy is the accurate identification of these position-dependent mutual inductances. As the number of transmitters increases, the mutual inductance identification process becomes computationally complex, and its accuracy is highly sensitive to model imperfections and measurement noise. This creates a significant practical limitation for scenarios where the receiver position changes frequently and is not known precisely in advance. UAV wireless charging, characterized by unpredictable landing positions, exemplifies such a scenario and highlights the need for a control strategy that is robust to these uncertainties without relying on precise and complex system identification or an explicit mechanistic model.

Reinforcement learning (RL), a data-driven control paradigm, offers a promising path to transcend the limitations of model-based methods. RL has demonstrated considerable potential in managing complex systems, including in the domain of power electronics [21–26] and UAV-related decision problems [27].

This paper presents a Soft Actor-Critic (SAC)-based power optimization method for MTSR-WPT systems. Under the practical constraint of bounded voltage source amplitudes, the method addresses the power optimization problem under the UAV landing position uncertainty by directly controlling the phases of the transmitter voltages. Unlike methods that optimize current amplitudes and thus require an additional conversion step from

voltage sources, this method directly adjusts the source voltages, which are the native inputs to the practical WPT system. The SAC agent learns a control policy that maps real-time observations from the system directly to actions for power optimization.

The main contributions of this paper are as follows:

- A data-driven, SAC-based control framework is proposed for MTSR-WPT systems that achieves power optimization under uncertain UAV landing positions, formulating the problem as bounded-amplitude, direct phase control without explicit online parameter identification.
- A physics-informed dual-current state representation is constructed from measurable current responses, so that the controller can exploit position-related coupling information without requiring explicit mutual-inductance estimation.
- A deployment-oriented validation chain is provided, including nominal regional evaluation, robustness to measurement noise and model mismatch, and hardware verification on a physical prototype.

The remainder of this paper is organized as follows. Section 2 models the MTSR-WPT system and analyzes the challenges of traditional methods. Section 3 details the proposed SAC-based method, including problem formulation, the interaction framework, and algorithm design. Section 4 presents comprehensive simulation results and analysis. Section 5 describes the physical experimental validation. Section 6 discusses the results in relation to existing work, and Section 7 provides the concluding remarks.

2. Modeling and Analysis of the MTSR-WPT System

Figure 1 illustrates a target application scenario where a UAV lands on a planar charging structure equipped with a multi-transmitter coil array. This configuration aims to leverage the field-shaping capability of the array to compensate for variations in energy transfer level caused by landing inaccuracies.

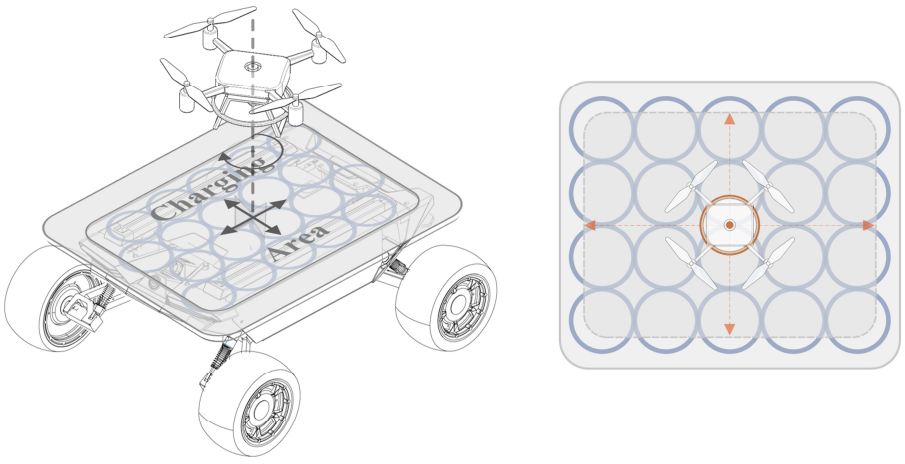


Figure 1. The application scenario using an array of transmitter coils for UAV charging. The blue and orange circles denote the transmitter and receiver coils, respectively, and the dashed arrows indicate the spatial freedom of movement.

Figure 2 shows the equivalent circuit of an n-transmitter, single-receiver MTSR-WPT system with series compensation. Taking Transmitter 1 (Tx1) as an example, V_1 and $R_{S,1}$ represent the voltage source and its internal resistance, respectively. $L_{T,1}$, $C_{T,1}$, and $R_{T,1}$ denote the inductance, resonant capacitance, and equivalent resistance of the transmitter coil. The receiver mounted on the UAV comprises a coil (L_R, R_R), a resonant capacitor C_R, and a load resistance R_L. The mutual inductance between the i-th and j-th transmitters is denoted by $M_{i,j}$, while $M_{R,k}$ represents the mutual inductance between the k-th transmitter

and the receiver. Crucially, $M_{R,k}$ varies with the receiver position relative to the transmitter array, corresponding to the UAV's landing spot.

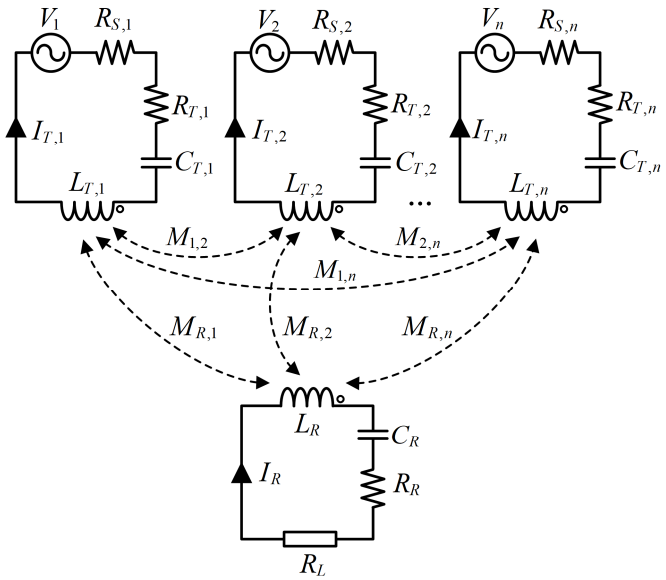


Figure 2. The equivalent circuit model of the MTSR-WPT system.

Assuming the power sources operate at an angular frequency ω, the Kirchhoff voltage law (KVL) equations for the MTSR-WPT system are written as follows:

$$\begin{bmatrix} Z_{T,1} & -j\omega M_{1,2} & \cdots & -j\omega M_{1,n} & -j\omega M_{R,1} \\ -j\omega M_{2,1} & Z_{T,2} & \cdots & -j\omega M_{2,n} & -j\omega M_{R,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -j\omega M_{n,1} & -j\omega M_{n,2} & \cdots & Z_{T,n} & -j\omega M_{R,n} \\ -j\omega M_{R,1} & -j\omega M_{R,2} & \cdots & -j\omega M_{R,n} & Z_R \end{bmatrix} \begin{bmatrix} I_{T,1} \\ I_{T,2} \\ \vdots \\ I_{T,n} \\ I_R \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \\ 0 \end{bmatrix} \tag{1}$$

where the impedances are defined as:

$$\begin{cases} Z_{T,k} = R_{S,k} + R_{T,k} + j(\omega L_{T,k} - 1/(\omega C_{T,k})) \\ Z_R = R_L + R_R + j(\omega L_R - 1/(\omega C_R)) \end{cases} \tag{2}$$

In this paper, all source voltages and coil currents are represented as RMS phasors. Under this convention, the load power is defined as

$$P_L = |I_R|^2 R_L \tag{3}$$

For clarity, the power transfer efficiency (PTE) is defined separately as

$$\eta = \frac{P_L}{P_L + P_{\text{loss}}} \tag{4}$$

where P_{loss} denotes the total ohmic loss in the resistive branches of the system. In the remainder of this paper, P_L is the optimization objective, whereas PTE is used only as an auxiliary performance metric. Alternative formulations of maximum-efficiency analysis in multi-port WPT systems, including S-parameter-based approaches, have also been reported [28]. However, the present work focuses on load-power maximization under phase control rather than direct PTE maximization.

Traditionally, model-based approaches for optimizing the load power P_L in such a system typically follow a two-step procedure. First, the uncertain transmitter-receiver mutual inductances, which depend on the receiver position, must be identified, often through dedicated excitation and measurement cycles. Second, with the estimated values $\hat{M}_{R,k}$, the optimal voltages are calculated by solving an optimization problem derived from the model in Equation (1).

For systems with a small number of transmitters, analytical solutions can be derived, but their complexity escalates rapidly. With n = 1, 2, for example, the estimated mutual inductance and receiver current are as follows:

$$\begin{aligned}
n=1,\quad & \hat{M}_{R,1}=\frac{Z_R^{(1/2)}(V_1-Z_{T,1}I_{T,1})^{(1/2)}}{I_{T,1}(1/2)}\\
& I_R=\frac{j\omega\hat{M}_{R,1}V_1}{\omega^2\hat{M}_{R,1}^2+Z_{T,1}Z_R}\\
n=2,\quad & \hat{M}_{R,1}=\frac{(-Z_R/\alpha)^{(1/2)}}{\omega\gamma/\beta},\hat{M}_{R,2}=\frac{(-Z_R/\alpha)^{(1/2)}}{\omega/\beta}\\
& \alpha=I_{T,1}V_1-I_{T,2}^2Z_{T,2}-I_{T,1}^2Z_{T,2}+I_{T,2}V_2+2jI_{T,1}I_{T,2}M_{1,2}+j\omega I_{T,1}I_{T,2}M_{1,2}\\
& \beta=2I_{T,1}M_{1,2}+jI_{T,2}Z_{T,2}-jV_2\\
& \gamma=(V_2-I_{T,2}Z_{T,2}+2jI_{T,1}M_{1,2})/(V_1-I_{T,1}Z_{T,1}+j\omega I_{T,2}M_{1,2})\\
& I_R=\frac{\omega\left(2M_{1,2}\hat{M}_{R,1}V_1-j\hat{M}_{R,1}V_1Z_{T,1}+\omega M_{1,2}\hat{M}_{R,1}V_2\right)}{Z_RZ_{T,1}Z_{T,2}+2\omega M_{1,2}^2Z_R+\left(2jM_{1,2}\hat{M}_{R,1}\hat{M}_{R,2}+\hat{M}_{R,1}^2Z_{T,2}+\hat{M}_{R,2}^2Z_{T,1}\right)+j\omega^3M_{1,2}\hat{M}_{R,1}\hat{M}_{R,2}}
\end{aligned} \tag{5}$$

These expressions illustrate a fundamental challenge: the complexity of the model-based solution grows rapidly with the number of transmitters. More critically, the accuracy of the final solution depends entirely on the precision of the circuit model and the identified parameters. In practical systems, unmodeled cross-couplings, parasitic components, and measurement noise can introduce significant errors in both steps. For a UAV application, where the landing position and consequently all $M_{R,k}$ values change unpredictably, the process of parameter identification and model-based optimization must be executed rapidly upon each landing. This requirement places a substantial computational burden and, more importantly, results in optimization performance that is highly sensitive to the accuracy of the model and parameter identification.

The analysis above underscores the limitations of model-dependent methods. Therefore, this paper proposes a data-driven control strategy. This strategy does not rely on a precise parametric model and is designed to adapt to the core challenge of UAV wireless charging: real-time power optimization under landing position uncertainty.

3. SAC-Based Power Optimization Method for MTSR-WPT Systems

To achieve power transfer objectives while circumventing the dependence on precise models, this section details a data-driven power optimization method based on the SAC reinforcement learning algorithm.

3.1. Problem Formulation

The phase control problem for the MTSR-WPT system in a UAV charging context is formally defined as follows. For any UAV landing position (x_R, y_R) within the operational plane, the objective is to adjust the input voltage phase vector $\mathbf{\Phi}=[\phi_1,\phi_2,\ldots,\phi_n]^T$ ($\phi_1\equiv 0^\circ$ as the reference) in real time to maximize the power delivered to the load P_L:

$$\max_{\mathbf{\Phi}} P_L(x_R, y_R, \mathbf{\Phi}) \tag{6}$$

Theorem 1. *For a fixed receiver position p in the MTSR-WPT system described by Equation (1), consider the source voltage vector.*

$$\mathbf{V} = \left[A_1 e^{j\phi_1}, A_2 e^{j\phi_2}, \ldots, A_n e^{j\phi_n}\right]^T \tag{7}$$

where the source amplitudes satisfy $0 \le A_i \le A_{i,\max}$ and the phases ϕ_i are freely adjustable. Then the maximum load power P_L is attained when

$$A_i = A_{i,\max},\ i = 1, 2, \ldots, n. \tag{8}$$

Proof of Theorem 1. Rewrite Equation (1) in block form. Let $\mathbf{Z}_T$ denote the $n \times n$ transmitter impedance matrix. $\mathbf{M}_R(p)$ denotes the transmitter–receiver mutual-inductance vector. $\mathbf{I}_T$ denote the transmitter current vector. Then the receiver current can be expressed as

$$I_R = \frac{-j\omega \mathbf{M}_R^T(p)\mathbf{Z}_T^{-1}\mathbf{V}}{Z_R + \omega^2 \mathbf{M}_R^T(p)\mathbf{Z}_T^{-1}\mathbf{M}_R(p)} = \mathbf{c}^T\mathbf{V} \tag{9}$$

where $\mathbf{c}^T$ is independent of $\mathbf{V}$ for a fixed p. Since the load power is $P_L = |I_R|^2 R_L$, maximizing P_L is equivalent to maximizing $\left|\mathbf{c}^T\mathbf{V}\right|$. Let $c_i = |c_i| e^{j\angle c_i}$, then

$$|\mathbf{c}^T\mathbf{V}| = \left|\sum_i c_i V_i\right| = \left|\sum_i |c_i| A_i e^{j\phi_i + \angle c_i}\right| \le \sum_i |c_i| A_i \tag{10}$$

and the upper bound is achieved when

$$\phi_i = -\angle c_i + \phi_0 \tag{11}$$

where ϕ_0 is an arbitrary common phase offset. Therefore, the original problem reduces to maximizing the linear function $\sum_i |c_i| A_i$ subject to the box constraints $0 \le A_i \le A_{i,\max}$, whose optimum is trivially attained at $A_i = A_{i,\max},\ \forall i$. □

Therefore, under the adopted load-power objective, this work fixes all source voltage amplitudes at their allowable upper limits and treats only the input voltage phases as the direct control variables. The specific phase configuration that yields the maximum load power still varies with the receiver position.

This problem is characterized by several challenges rooted in the system's physics. First, the action space is periodic and discontinuous at the $\pm 180^\circ$ boundaries due to the equivalence $\phi_i \equiv \phi_i + 360^\circ$, posing difficulties for standard optimization. Second, the relationship $P_L(\mathbf{\Phi})$ is highly nonlinear and non-convex, riddled with numerous local maxima. Third, the system model involves uncertainties from parasitic components, cross-coupling, and hardware imperfections. Crucially, the receiver position itself is unpredictable for each UAV landing event.

To address these challenges, the problem is cast as a sequential decision-making process within a Markov Decision Process framework. At each time step t, the agent observes the environment state $\mathbf{s}_t$ and outputs a phase adjustment action $\mathbf{a}_t$. The environment executes the action, transitions to a new state $\mathbf{s}_{t+1}$, and provides a scalar reward signal r_t. The reward at t is defined as

$$r_t = A \cdot \frac{P_L(t)}{P_{\text{ref}}} + B \tag{12}$$

where $P_L(t)$ is the instantaneous load power, P_{ref} is the reference load power under the single-transmitter single-receiver aligned condition, and A and B are the reward scaling and

bias coefficients, respectively. Under this definition, the quantity $-B/A$ can be interpreted as a positive-reward threshold with respect to the normalized load power P_L/P_{ref}. A higher threshold makes early-stage learning more likely to remain in the negative-reward regime, while an overly small scaling factor weakens the effective learning signal. Therefore, A and B influence the learning dynamics, but they do not change the physical optimization objective, which remains load-power maximization.

3.2. SAC-WPT Interaction Framework

The proposed interaction framework between the SAC agent and the MTSR-WPT system is depicted in Figure 3. The SAC algorithm is chosen for its effectiveness in continuous action spaces and its maximum entropy framework, which encourages exploration, a beneficial property for navigating the multi-modal optimization landscape of phase control.

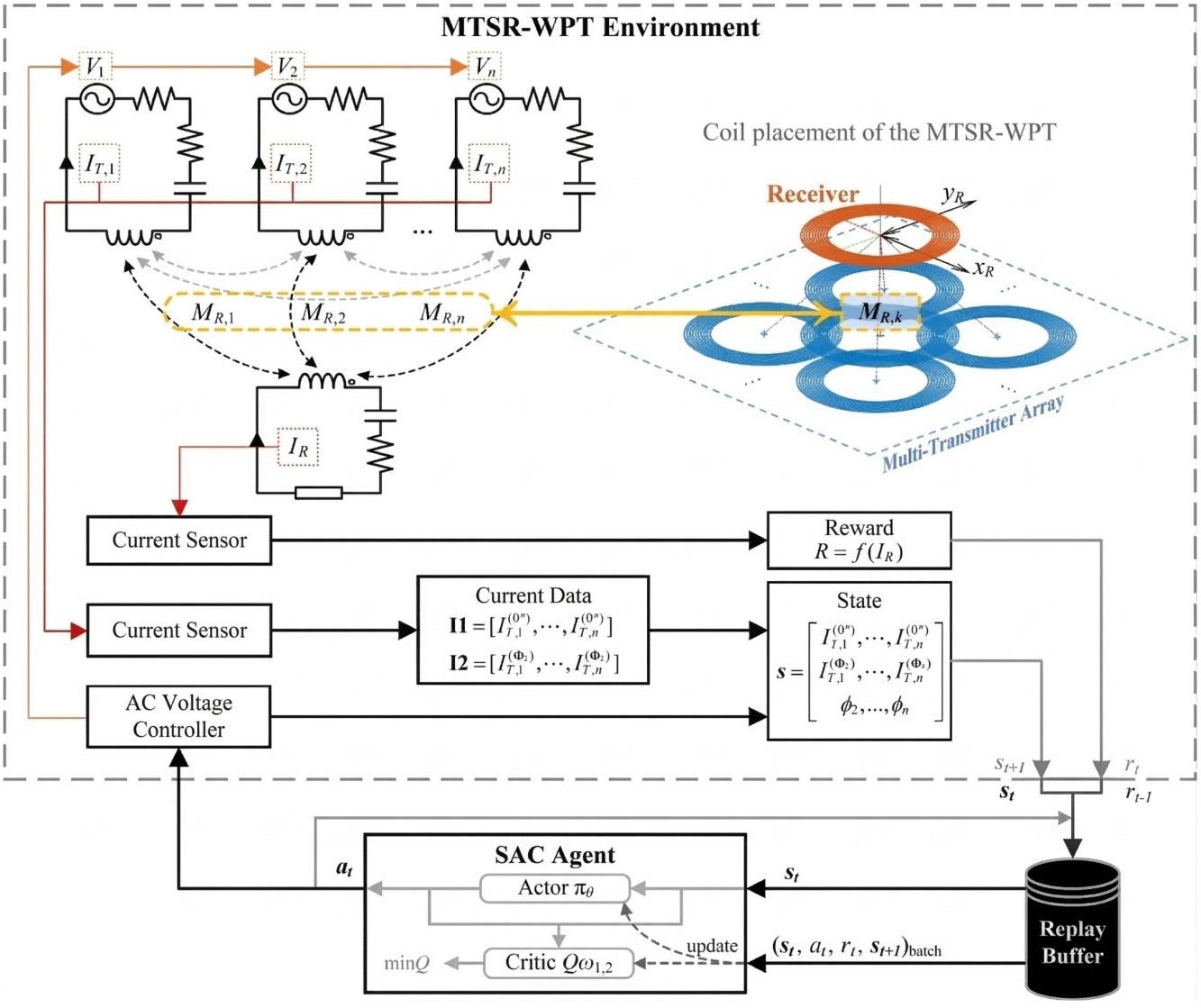


Figure 3. MTSR-WPT Phase Control Interaction Framework.

The core interaction loop is as follows:

(1) **State Observation**: The agent's observation $\mathbf{s}_t$ is a dual-current state vector $\mathbf{s}_t = \left[\mathbf{I1}^{(0^\circ)}, \mathbf{I2}^{(\mathbf{\Phi}_t)}, \mathbf{\Phi}_t\right]^T$ composed of three parts: $\mathbf{I1}^{(0^\circ)}$ (the transmitter current features measured under the fixed zero-phase reference excitation, implicitly reflecting the receiver's position-dependent coupling state), $\mathbf{I2}^{(\mathbf{\Phi}_t)}$ (the current features

measured under the current phase configuration $\mathbf{\Phi}_t$, reflecting the immediate effect of the action), and the voltage-source phase values $\mathbf{\Phi}_t$ themselves. This design decouples the "landing position effect" from the "action effect," providing structured physical prior knowledge to the agent.

(2) **Agent Decision**: The SAC actor network (policy π_θ) takes the state $\mathbf{s}_t$ and outputs a phase adjustment $\Delta\mathbf{\Phi}_t$.

(3) **Action Execution and Learning**: The phase controller applies the adjustment, yielding the new phase vector $\mathbf{\Phi}_{t+1} = \text{wrap}(\mathbf{\Phi}_t + \Delta\mathbf{\Phi}_t)$, which is fed to the system. The environment evolves, producing a new state $\mathbf{s}_{t+1}$ and a reward r_t. The experience tuple $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ is stored in a replay buffer for updating the agent's neural network parameters.

Here, **I1** is not treated as the receiver position itself. Instead, under the fixed zero-phase reference excitation, the transmitter current response is determined by the receiver-position-dependent coupling relationship. Therefore, **I1** can be regarded as a directly measurable proxy of the receiver's position-dependent coupling state. In contrast, **I2** represents the current response under the current phase command and thus provides complementary information on how the present control action affects the system. Accordingly, the dual-current state is adopted as a physics-informed and practically measurable state description, rather than being claimed as the unique minimal state representation.

3.3. SAC Algorithm Design

This section details the key design aspects of the SAC algorithm tailored for power optimization in MTSR-WPT systems for UAV wireless charging.

3.3.1. Maximum Entropy Objective and Exploration

The SAC agent aims to maximize the expected cumulative reward while also maximizing the entropy of its policy. Its objective function is:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t)\sim\rho_\pi} \left[\gamma^t (r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\cdot|\mathbf{s}_t)))\right] \tag{13}$$

Here, α is a temperature parameter controlling the trade-off between reward and entropy, and $\mathcal{H}$ represents policy entropy. The entropy term encourages stochastic exploration, which is crucial for escaping local optima in the phase space. Although this mechanism helps reduce the risk of premature convergence, it does not provide a theoretical guarantee of global optimality.

3.3.2. Handling Periodic Phase Actions

In this paper, bold $\mathbf{\Phi}$ denotes the voltage source phase vector, ϕ_i denotes its scalar components, and $\mathbf{\Phi}^*$ refers to phase information requiring periodic encoding (including both voltage source phases and current phases). To address the periodic nature of phases ($\phi_i \equiv \phi_i + 360^\circ$), specialized handling is incorporated into the agent network design:

(1) **Action Output and Decoding**: The policy network π_θ outputs a raw action $\mathbf{o}_t$ defined within the range $[-1, 1]^{n-1}$ (corresponding to $n-1$ adjustable phases). The actual phase adjustment is obtained through scaling and periodic wrapping:

$$\Delta\mathbf{\Phi}_t = C \cdot \mathbf{o}_t,\ \mathbf{\Phi}_{t+1} = \text{wrap}_{[-180^\circ, 180^\circ]}(\mathbf{\Phi}_t + \Delta\mathbf{\Phi}_t),\ C \in [0, 180] \tag{14}$$

(2) **Sine-Cosine Encoding of Phases**: Two types of phase information are involved in state construction:

- Transmitter Coil Current Phases: Contained within **I1** and **I2**.
- Voltage Source Phases: $\mathbf{\Phi}_t$ at the end of the state vector.

To avoid discontinuities in the value function at the $\pm 180^\circ$ boundaries, this paper employs a unified sine-cosine encoding for all phase information $\mathbf{\Phi}^*$:

$$\mathbf{\Phi}^*_{\text{encoded}} = [\sin(\mathbf{\Phi}^*), \cos(\mathbf{\Phi}^*)] \tag{15}$$

(3) **State Dimension Specification**: Based on the above unified encoding scheme, the total dimension of the state vector $\mathbf{s}$ is $8n - 2$.

3.3.3. Value and Policy Learning with the Dual-Current State

The SAC agent comprises a stochastic policy network (Actor) π_θ and two soft Q-function networks (Critics Q_{ω_1}, Q_{ω_2}) and their corresponding target networks ($Q_{\overline{\omega}_1}$, $Q_{\overline{\omega}_2}$). The policy is modeled as a Gaussian with mean and covariance output by the actor network. The learning process involves updating these networks using experience sampled from the replay buffer D.

(1) **Critic Update**: The critics are updated by minimizing the mean-squared Bellman error. The target value y_t for a transition $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}, d)$ is computed as:

$$y_t = r_t + \gamma(1-d)\left(\min_{j=1,2} Q_{\overline{\omega}_j}(s_{t+1}, \widetilde{a}_{t+1}) - \alpha \log \pi_\theta(\widetilde{a}_{t+1}|s_{t+1})\right), \widetilde{a}_{t+1} \sim \pi_\theta(\cdot|s_{t+1}) \tag{16}$$

The loss for critic i is then:

$$\mathcal{L}_Q(\omega_i) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t)\sim\mathcal{D}}\left[\left(Q_{\omega_i}(\mathbf{s}_t, \mathbf{a}_t) - y_t\right)^2\right],\ i = 1, 2 \tag{17}$$

(2) **Actor Update**: The actor is updated to maximize the expected Q-value and policy entropy:

$$\mathcal{L}_\pi(\theta) = \mathbb{E}_{\mathbf{s}_t\sim\mathcal{D}}\left[\alpha \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) - \min_{j=1,2} Q_{\omega_j}(\mathbf{s}_t, \mathbf{a}_t)\right] \tag{18}$$

(3) **Adaptive Temperature Adjustment**: The temperature parameter α, which controls the exploration-exploitation trade-off, can be learned automatically by minimizing:

$$\mathcal{L}(\alpha) = \mathbb{E}_{\mathbf{s}_t\sim\mathcal{D}}\left[-\alpha\left(\log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) + \overline{\mathcal{H}}\right)\right] \tag{19}$$

where $\overline{\mathcal{H}}$ is the target entropy, typically set to $-\dim(\mathcal{A})$.

3.4. Algorithm Implementation

Algorithm 1 provides the complete training procedure for the SAC agent in the MTSR-WPT phase control task.

Key Implementation Details:

(1) **Application of Dual-Current State**: The state $\mathbf{s}$ constructed incorporates both **I1** and **I2**, enabling the critic and actor networks to learn and make decisions based on decoupled position and action effects.
(2) **Hybrid Experience Replay**: Line 13 employs a sampling strategy that mixes data from the current episode with historical data from the buffer, balancing rapid adaptation to new positions with training stability.
(3) **Periodic Wrapping**: The wrap function in line 8 is crucial for maintaining phase values within the valid $[-180^\circ, 180^\circ]$ range.
(4) **Unified State Encoding**: The phase information in the state is included through sine-cosine encoding. The current-phase information is contained within **I1** and **I2**, while the voltage source phase information is appended as encoded values.

The actor and critic networks in the SAC agent are implemented as fully connected neural networks. The off-policy nature of SAC combined with experience replay ensures high sample efficiency during training. The proposed framework is not restricted to the coil configuration used in this paper. As the number of transmitters increases, the action dimension grows linearly, and the state dimension also increases accordingly, which raises training cost and data demand but does not change the formulation itself. In comparison with model-based online optimization, most of the computational burden of the present method is shifted to offline training, whereas online deployment only requires state acquisition and a forward pass of the policy network.

Algorithm 1: SAC-Based Phase Control for MTSR-WPT Systems

Input: MTSR-WPT environment Env, SAC agent with Actor πθ, Critics Qω1, Qω2, replay buffer D
Hyperparameters: discount γ, target update rate τ, target entropy $\overline{H}$
Output: Trained SAC agent model {πθ, Qω1, Qω2}

Initialize actor network θ, critic networks ω1, ω2, and target parameters $\overline{\omega}1 \leftarrow \omega1$, $\overline{\omega}2 \leftarrow \omega2$
for episode = 1 to M do
    Sample receiver position (xr, yr) within working area
    Reset Env at (xr, yr), obtain initial state s0 = [I1(0°), I2(Φ0), Φ0_encoded]^T
    for t = 0 to T − 1 do
        Select action at ∼πθ(· | st)
        Execute action at, observe reward rt and next state st + 1, done flag dt
        Wrap phase components in Φt + 1 to [−180°, 180°]
        Construct dual-current state st + 1 = [I1(0°), I2(Φt + 1), Φt + 1_encoded]^T
        Store transition (st, at, rt, st + 1, dt) in D.
        # Agent Update
        if | D | > batch_size then
            Batch ← sample_hybrid_batch(D, current_episode_ratio = 0.5)
            # Update Critic networks (Equations (16) and (17))
            y = r + γ(1 − d) * (min_j Qω-j(s′, $\tilde{a}'$) − α log πθ($\tilde{a}'$ | s′))
            Update ω1, ω2 by ∇ωi LQ(ωi) for i = 1, 2
            # Update Actor network (Equation (18))
            Update θ by ∇θ Lπ(θ)
            # Update temperature parameter α (Equation (19))
            Update α by ∇α L(α) if auto_alpha is True
            # Soft update target networks
            $\overline{\omega}_i \leftarrow \tau\omega_i + (1-\tau)\overline{\omega}_i$, for i = 1, 2
        end if
    end for
end for

4. Simulation Results and Analysis

4.1. Simulation Setup

To validate the proposed method, a 5-transmitter 1-receiver (5T1R) WPT system model is constructed. The system operates at 1.45 MHz. The key electrical and geometric parameters of the transmitter and receiver coils are listed in Table 1.

As shown in Figure 4, the UAV receiver moves within a two-dimensional plane above the transmitter array. The inner square region, $[0, D] \times [0, D]$, is defined as the training region, where $D = 200$ mm denotes the side length of the nominal service area. To evaluate how the learned policy handles unseen landing positions outside the training distribution, an expanded evaluation region is further introduced by extending each side of the training region outward by $[-\kappa D, (1+\kappa)D] \times [-\kappa D, (1+\kappa)D]$. Therefore, the evaluation region

fully contains the training region while additionally covering surrounding out-of-training positions near the array boundary.

Table 1. Parameter values of the simulated MTSR-WPT system.

Parameter Type	Parameter Value
Coil inductances (μH) $L_{T,k}/L_R$	56.0/56.0
Capacitances (pF) $C_{T,k}/C_R$	220/220
Resistances (Ω) $R_{S,k}/R_{T,k}/R_R/R_L$	50/3.5/3.5/50
Height of the receiver plane (mm) z_R	100
Voltage amplitude (V) V_k	8
Frequencies (MHz) f	1.45

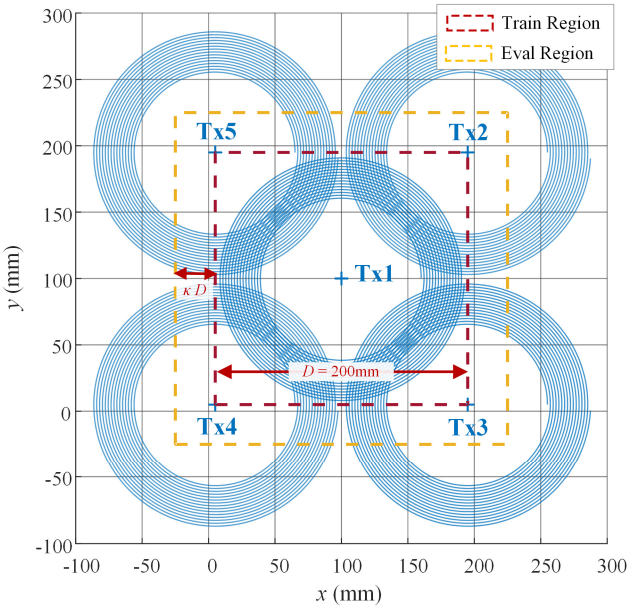


Figure 4. Geometry of the 5T1R transmitter array, the training region, and the expanded evaluation region used in simulation.

The RL environment and agent are configured as follows:

(1) **State Space**: The proposed dual-current state representation is employed. For the 5T1R system, the state vector contains: the zero-phase current response $\mathbf{I1}^{(0^\circ)}$ (15 dimensions), the current-phase response $\mathbf{I2}^{(\mathbf{\Phi}_t)}$ (15 dimensions), and the voltage source phase $\mathbf{\Phi}_t$ (8 dimensions), resulting in a total of 38 dimensions.

(2) **Phase Space**: The system controls four independent transmitter voltage phases $(\phi_2, \phi_3, \phi_4, \phi_5)$, each within the continuous range $[-180^\circ, 180^\circ]$.

(3) **Reward Function**: The reward function follows Equation (12), directly incentivizing an increase in the received charging power P_L.

(4) **Algorithm Hyperparameters**: The main SAC hyperparameters are summarized in Table 2.

During training, each episode randomly samples a receiver position within the training region for a maximum of 50 control steps, totaling 500 training episodes. For evaluation, dense grid sampling is performed on both regions. To establish a performance benchmark, the theoretical maximum power for each test point is obtained via a two-stage global search: an initial coarse grid search followed by multi-scale random exploration in promising regions.

Table 2. SAC Algorithm Hyperparameters Settings.

Hyperparameter	Setting
State/Action dimension	38/4
* Reward coefficient A/B	$10/-5$
Initial learning rate actor/critic/α	$1 \times 10^{-4}/1 \times 10^{-4}/1 \times 10^{-4}$
Hidden layers structure of actor/critic	[256, 256]
Discount factor γ	0.98
Target entropy $\overline{\mathcal{H}}$	-4
Initial temperature α	0.15
Soft update coefficient τ	0.005
Capacity of replay buffer	5000
Batch size	64
Total episodes	500
Extending coefficient κ	0.15
Sampling accuracy of training region	0.1 mm
Optimizer of actor/critic	Adaptive Moment Estimation (Adam)
Key phase for grid search	$[-135, -45, 45, 135]$
Noise scales for multi-scale perturbation	[60, 30, 15, 7.5]

* The values listed here correspond to the default setting used in the main experiments. Additional reward coefficient settings (10, −15) and (1, −5) are only used in the sensitivity study of Section 4.2.1.

4.2. SAC Agent Training and Performance

This section first examines the sensitivity of the training behavior to the reward coefficients. Then, the nominal performance of the learned SAC agent is evaluated in both the training region and the expanded evaluation region. Finally, the robustness of the method under practical imperfections is investigated, including measurement noise and model mismatch.

4.2.1. Sensitivity to Reward Coefficients

To examine how the reward coefficients affect the training behavior, three representative (A, B) settings are compared: (10, −5), (10, −15), and (1, −5), while all other training hyperparameters are kept unchanged. In particular, the target entropy, learning rates, replay-buffer settings, and environment configuration are fixed so that the comparison only reflects the influence of the reward coefficients. To better compare the differences in training dynamics among different A/B settings, the normalized return of each episode is used. Let T denote the number of steps in one episode. Then, the episode return is defined as

$$R = \sum_{t=1}^{T} r_t = A \sum_{t=1}^{T} (P_L / P_{\text{ref}}) + B \cdot T \tag{20}$$

Accordingly, the normalized return can be written as

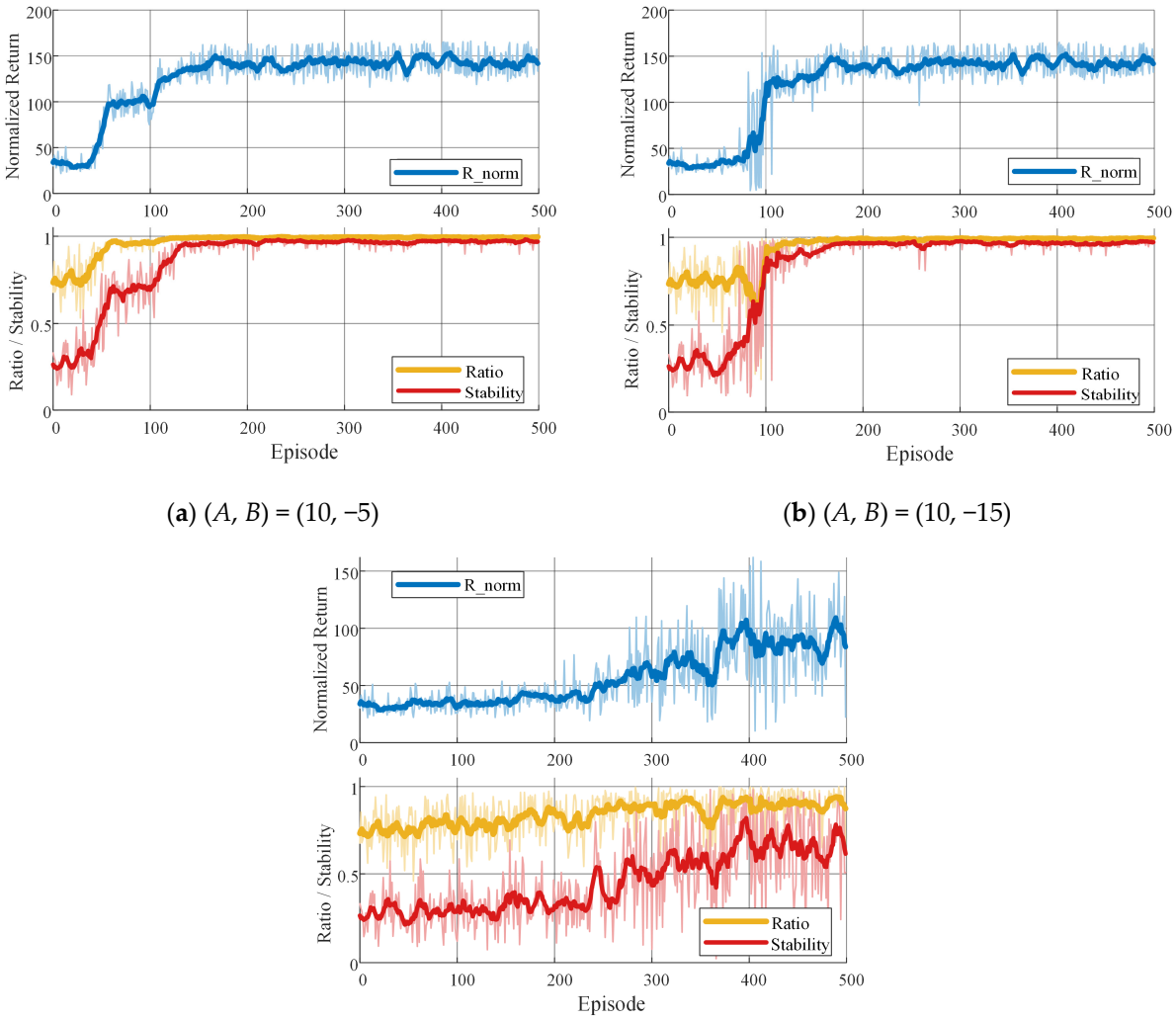
$$R_{\text{norm}} = \frac{R - B \cdot T}{A} = \sum_{t=1}^{T} (P_L / P_{\text{ref}}) \tag{21}$$

The normalized return removes the additive shift introduced by B and the scaling effect introduced by A, thereby enabling a direct comparison of training curves under different (A, B) settings. In addition, two complementary metrics are introduced.

Ratio is defined as $P_{\text{best}}/P_{\text{max}}$ within each episode, where P_{best} denotes the highest power achieved during RL training and P_{max} is the theoretical optimum. A value closer to 1 indicates closer agreement with the theoretical optimum.

Stability is defined as the average power over the last K steps (K = 5) divided by P_{best} within each episode. A value closer to 1 indicates that the learned policy can remain closer to its best-achieved operating point during the final stage of an episode.

Figure 5 compares the normalized return together with the Ratio and Stability curves under the three different (A, B) settings. The results show that the setting (1, −5) yields markedly slower and less stable learning, which is consistent with its substantially higher positive-reward threshold ($-B/A = 5$). By contrast, both (10, −15) and (10, −5) converge to high-performance solutions. Considering both training stability and consistency with the subsequent robustness analysis, (10, −5) is selected as the default reward setting throughout the following experiments.



(**a**) (A, B) = (10, −5)

(**b**) (A, B) = (10, −15)

(**c**) (A, B) = (1, −5)

Figure 5. Training curves under different reward coefficient settings.

The trained SAC agent is then evaluated on the dense grid of points within both the training and expanded evaluation regions. During testing, the agent operates autonomously based solely on the real-time state $\mathbf{s}_t$ without receiving the reward signal r_t.

4.2.2. Performance in the Training and Expanded Evaluation Regions

To evaluate the nominal performance of the trained SAC agent, the policy is first tested in the training region and then in the expanded evaluation region. In both regions, the zero-phase initial power P_0, the theoretical maximum power P_{max}, the peak power

achieved during RL interaction P_{best}, and the final stable power P_{last5} are compared. In addition, the power-transfer efficiency before and after optimization is summarized to show that the power improvement does not come at the expense of reduced efficiency.

(1) Performance in the Training Region

Figure 6 presents the three-dimensional power comparison in the training region. The SAC agent substantially improves the load power over the entire nominal service region. Statistically, the mean transfer power increases from 0.15 ± 0.08 W (P_0) to 0.44 ± 0.04 W (P_{best}), while P_{last5} remains at 0.43 ± 0.04 W. This corresponds to a +187.2% increase in the regional mean power relative to the zero-phase case. Compared with the theoretical maximum power 0.46 ± 0.04 W, P_{best} reaches 96.9% of P_{max} on average, and P_{last5} still maintains 95.9% of P_{max}. Moreover, the agent reaches 95% of P_{max} within only 2 steps at the p90 level. In terms of efficiency, the mean PTE increases from 89.4% to 93.4% after optimization.

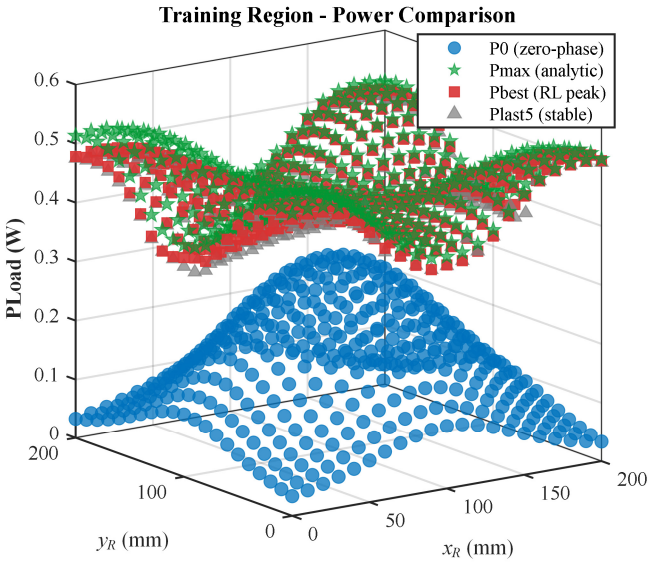


Figure 6. Power comparison in the training region.

(2) Performance in the Expanded Evaluation Region

Figure 7 further evaluates the trained policy in the expanded evaluation region, which includes positions outside the nominal training/service region. Although the receiver positions in this region are not all explicitly covered by the training distribution, the learned policy still preserves strong control performance. The mean transfer power rises from 0.11 ± 0.08 W (P_0) to 0.43 ± 0.05 W (P_{best}), while P_{last5} is 0.42 ± 0.06 W. Compared with the theoretical maximum power 0.45 ± 0.05 W, P_{best} reaches 95.6% of P_{max} on average, and P_{last5} remains at 94.4% of P_{max}. The agent still reaches 95% of P_{max} within 2 steps at the p90 level. Meanwhile, the mean PTE increases from 83.5% to 93.0%, indicating that the optimized phase control improves the load power while also maintaining a favorable efficiency level.

The quantitative statistics in Table 3 summarize the above observations. Overall, the proposed SAC-based phase controller achieves high nominal performance in both the training and expanded evaluation regions, with clear power improvement, near-optimal achieved power, fast convergence, and consistent efficiency enhancement.

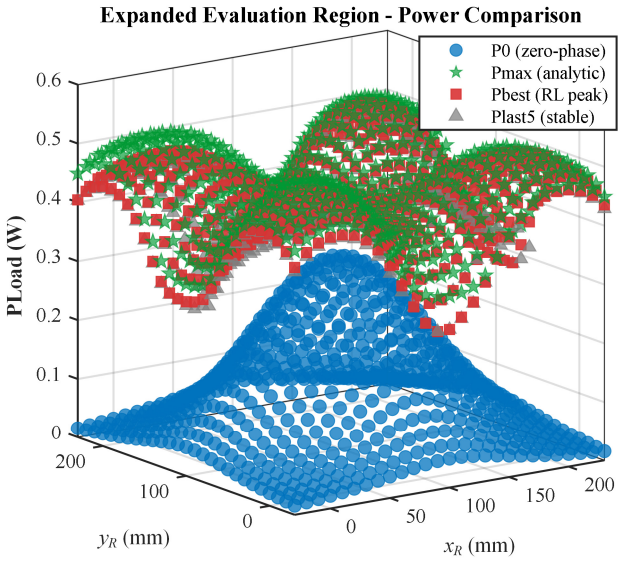


Figure 7. Power comparison in the expanded evaluation region.

Table 3. Statistical summary of performance in the training and expanded evaluation regions.

Metric	Training Region	Evaluation Region
P_0 (mean $\pm$ std)	0.15 ± 0.08 W	0.11 ± 0.08 W
P_{best} (mean $\pm$ std)	0.44 ± 0.04 W	0.43 ± 0.05 W
P_{last5} (mean $\pm$ std)	0.43 ± 0.04 W	0.42 ± 0.06 W
P_{max} (mean $\pm$ std)	0.46 ± 0.04 W	0.45 ± 0.05 W
Mean Gain $\overline{P_{\text{best}}}/\overline{P_0}$	2.87/+187.2%	3.92/+291.6%
Mean achieved ratio $\overline{P_{\text{best}}/P_{\text{max}}}$	96.9%	95.6%
Mean stable ratio $\overline{P_{\text{last5}}/P_{\text{max}}}$	95.9%	94.4%
Steps to 95%Pmax (p90)	2.0	2.0
PTE (zero-phase)	89.4%	83.5%
PTE (after optimization)	93.4%	93.0%

"p90" denotes the 90th percentile over all sampled positions in the region.

4.2.3. Robustness Under Practical Imperfections

Although the results in Section 4.2.2 demonstrate that the proposed SAC controller can achieve high transfer power and good generalization, practical deployment inevitably involves non-ideal factors such as measurement noise and model mismatch. Therefore, the robustness of the learned policy is further examined from two aspects: measurement noise in the observations and systematic mismatch between the nominal simulation model and the shifted deployment model. Unless otherwise stated, the following robustness statistics are reported over the expanded evaluation region to provide a stricter assessment of practical performance.

(1) Robustness to Measurement Noise

To evaluate robustness to noisy measurements, Gaussian noise is injected into the complex phasor observations used to construct the SAC state. More specifically, zero-mean noise is added to the real and imaginary parts of the measured voltage and current phasors, while the underlying physical circuit model and the corresponding power calculations remain noise-free. Therefore, this test specifically examines whether observation noise can mislead the learned controller, rather than whether the physical WPT system itself changes.

Table 4 summarizes the results in the expanded evaluation region under different signal-to-noise ratio (SNR) levels, including the clean baseline, 40 dB, 30 dB, and 20 dB. SNR_VI denotes the noisy phasor observations in the state construction. The results show

that when SNR_VI = 40 dB and 30 dB, the learned policy remains almost unchanged relative to the noise-free baseline. In particular, the mean ratio stays at 0.966 and 0.964, while the mean stability remains at 0.98 and 0.98, respectively. Meanwhile, steps to 95%Pmax (p90) stay around 2 steps, indicating that the controller still converges rapidly under moderate measurement noise.

Table 4. Statistical summary of performance under measurement noise.

SNR_VI	Ratio (Mean/p10)	Stability (Mean/p10)	Steps to 95%Pmax (p90)
clean	0.97/0.92	0.98/0.95	2.0
40 dB	0.97/0.92	0.98/0.95	2.0
30 dB	0.96/0.92	0.98/0.95	2.0
20 dB	0.95/0.90	0.97/0.93	2.9

"p10" and "p90" denote the 10th and 90th percentiles over all sampled positions in the region, respectively.

When the noise level is further increased to 20 dB, a noticeable degradation appears, but the controller does not collapse. Even in this stronger-noise case, the mean ratio remains at 0.95 and the mean stability remains at 0.97, indicating that most positions still achieve near-optimal power and acceptable end-of-rollout stability.

Overall, these results indicate that the proposed SAC controller is robust to moderate measurement noise, while under stronger noise it still remains functional but shows degraded performance at some extreme positions.

(2) Robustness to Model Mismatch and Fine-Tuning

Besides measurement noise, practical deployment may also suffer from model mismatch, since the true system parameters may deviate from the nominal model used during training. To examine this issue, a shifted WPT model is further constructed by introducing controlled perturbations into the circuit and coupling parameters. Three cases are then compared in the expanded evaluation region: nominal policy on the nominal model, nominal policy on the shifted model, and fine-tuned policy on the shifted model.

Table 5 shows that directly applying the nominally trained SAC policy to the shifted model causes a substantial performance drop. Specifically, the mean ratio decreases from 0.97 under the nominal policy on the nominal model to 0.54 under the nominal policy on the shifted model, and the mean stability decreases from 0.98 to 0.72. This indicates that model mismatch can significantly weaken both the optimality and the end-of-rollout stability of the learned policy.

Table 5. Statistical summary of performance under model mismatch and fine-tuning.

Case	Ratio (Mean/p10)	Stability (Mean/p10)	Steps to 95%Pmax (p90)
Nominal policy on nominal model	0.97/0.92	0.98/0.95	2.0
Nominal policy on shifted model	0.54/0.20	0.72/0.14	4.3
Fine-tuned policy on shifted model	0.93/0.84	0.92/0.80	3.0

The shifted model is constructed by applying the following perturbation ranges: $R/L/C/M \in \pm 5\%$, sampled with a fixed random seed. One realization is sampled once and then kept constant for all evaluation points. Fine-tuning is performed for 100 episodes on the shifted model, starting from the nominally trained agent.

Nevertheless, the degraded performance can be largely recovered by a short fine-tuning stage on the shifted model. After 100 fine-tuning episodes, the mean ratio rises to 0.93 and the mean stability rises to 0.92. Meanwhile, the steps to 95%Pmax (p90) are reduced from 4.3 to 3.0 after fine-tuning. These results indicate that, although model mismatch can noticeably degrade direct-transfer performance, the proposed SAC controller

still exhibits good adaptability, since most of the lost performance can be recovered through short fine-tuning on the shifted model.

Overall, the simulation results in Section 4.2 demonstrate four important properties of the proposed SAC-based phase-control method. First, with a properly selected reward structure, the training process remains stable and converges efficiently. Second, under the nominal model, the learned policy achieves substantial power enhancement in both the training region and the expanded evaluation region, while staying close to the theoretical power upper bound. Third, the controller shows strong robustness to moderate measurement noise. Finally, although model mismatch can noticeably degrade direct-transfer performance, most of the lost performance can be recovered through a short fine-tuning stage. Taken together, these results verify the effectiveness, spatial generalization capability, robustness, and practical adaptability of the proposed SAC controller for UAV wireless charging.

5. Experimental Validation

5.1. Experimental Platform Overview

To validate the practical feasibility of the proposed method, a physical five-transmitter single-receiver (5T1R) WPT hardware platform is constructed, emulating a UAV wireless charging setup. The platform, shown in Figure 8, consists of five transmitter coils arranged in a planar array, each driven by an individual channel of a multi-channel signal generator (FY7000 series, Feeltech, Zhengzhou, China) paired with a power amplifier (FPA301, Feeltech, Zhengzhou, China). A single receiver coil is connected to a load resistor, and its current is measured via a current probe (CP503B, Micsig, Shenzhen, China) to calculate the real-time transfer power $P_L(t)$. The currents of all transmitter coils are synchronously sampled to construct the state vector for the SAC agent.

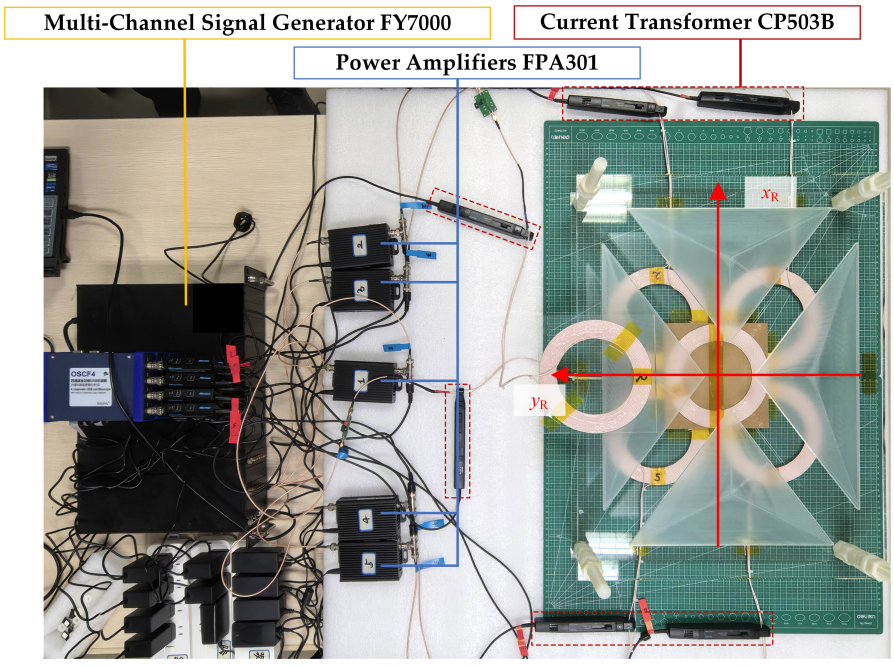


Figure 8. The 5T1R WPT experimental platform.

Compared to the ideal simulation model, this hardware platform introduces several practical non-ideal factors. These include nonlinearity and harmonic distortion from the power amplifiers, sensor noise and finite bandwidth, uncertainties in coil parasitic parameters, and quantization error in phase control commands. These factors collectively

create a challenging testbed to evaluate the effectiveness of the proposed SAC-based power optimization method. The key parameter values of the experimental MTSR-WPT system are given in Table 6. The fixed 8V voltage amplitude used in the hardware validation is a platform-specific safe operating setting rather than a universal assumption for arbitrary hardware. Practical constraints such as thermal dissipation, loop-current limits, and phase-command resolution remain relevant, while the same phase-control framework can also be used under a lower allowable amplitude bound if required.

Table 6. Parameter values of the experimental MTSR-WPT system.

Parameter Type	Parameter Value
Coil inductances (μH) $L_{T,1}/L_{T,2}/L_{T,3}/L_{T,4}/L_{T,5}/L_R$	57.2/57.1/57.7/57.5/57.6/56.8
Capacitances (pF) $C_{T,1}/C_{T,2}/C_{T,3}/C_{T,4}/C_{T,5}/C_R$	227.3/226.4/230.1/231.4/223.7/229.3
Resistances (Ω) $R_{T,1}/R_{T,2}/R_{T,3}/R_{T,4}/R_{T,5}/R_R/R_L$	3.7/3.6/4.7/3.8/3.8/3.5/51.5
Height of the receiver plane (mm) z_R	100
Voltage amplitude (V) of V_k	8
Frequencies (MHz) f	1.45

5.2. Experimental Procedure: Static Point Validation

A static point testing scheme is employed to efficiently evaluate the algorithm's performance across a spatial area within a limited time frame. A 5×5 grid of 25 test points is defined within the platform's effective working area, as illustrated in Figure 9, with a spacing of 25 mm. This grid represents a set of discrete potential landing spots for a UAV on the charging platform.

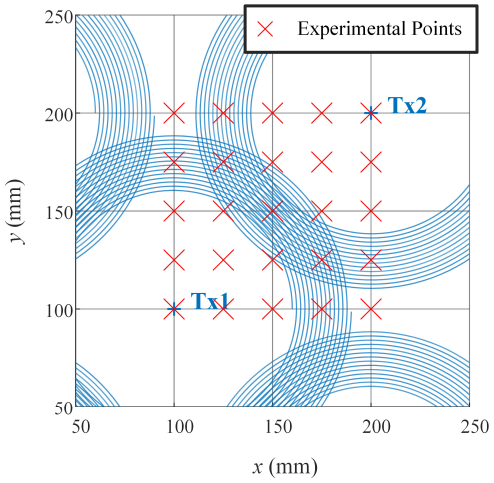


Figure 9. Spatial distribution of the static test points.

At each test point (x_R, y_R) representing a possible UAV resting position, the following standardized test procedure is executed:

(1) **Baseline Measurement**: Set all transmitter phases to zero ($\mathbf{\Phi} = \mathbf{0}$), measure and record the initial load power P_{init}, and simultaneously record **I1** as the position feature.
(2) **Phase Control**: Load the trained SAC agent for online phase control. Based on the real-time acquired dual-current state $\mathbf{s}_t$, the agent outputs a phase adjustment action $\Delta\mathbf{\Phi}_t$ once per control cycle.
(3) **Data Recording**: The transfer power $P_L(t)$, current state $\mathbf{s}_t$, and action $\mathbf{a}_t$ for each step during the control process are recorded.

In the present hardware implementation, one control step corresponds to one complete hardware cycle, including current acquisition, policy inference, phase update, and a short settling interval. Therefore, the step-based convergence reported in this paper

can be interpreted as convergence within the corresponding number of hardware control cycles. On the used experimental platform, the measured single-step cycle time is below 500 ms. The hardware validation is conducted under fixed-amplitude static-point operation rather than aggressive continuous phase sweeping; in stricter deployment scenarios, practical safeguards such as phase-step limiting and current/temperature protection can be incorporated within the same control framework.

5.3. Experimental Results and Analysis

Figure 10 summarizes the spatial distribution of the experimental results, comparing the initial zero-phase power P_0 (blue dots) and the maximum power achieved under SAC control P_{best} (red dots) at all 25 test points. The results visually demonstrate that P_{best} is significantly higher than P_0 at all test positions, confirming the global effectiveness of the algorithm across the physical charging area. The relative improvement is particularly pronounced in edge regions, which is consistent with the spatial variation characteristics of electromagnetic coupling and highlights the practical benefit for UAVs landing near the periphery of a charging pad.

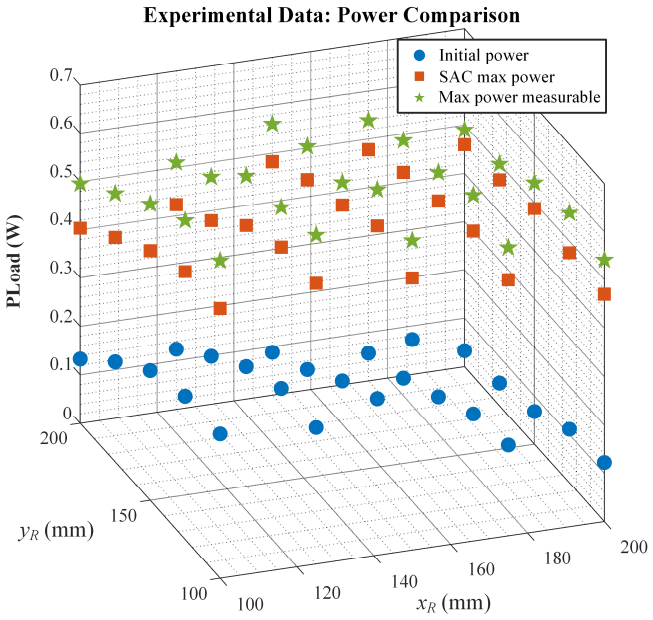


Figure 10. Power comparison of experimental outcomes.

Quantitatively, the proposed method delivers substantial and consistent performance gains. Compared to the zero-phase baseline, the SAC control policy increases the average load power from 0.18 ± 0.10 W to 0.51 ± 0.05 W. This represents an average relative improvement of 188.5%. Furthermore, the power level achieved by the SAC policy reaches 87% of the maximum power attainable through a guided, extensive phase search on the physical hardware.

The experimental results confirm the efficacy and robustness of the SAC-based phase control strategy in a real hardware environment. The observed performance gap between 87% of the maximum power attainable on the hardware and the more than 97% achieved in near-ideal simulation is primarily attributable to practical non-idealities inherent to the physical platform. These include nonlinearity and harmonic distortion from the power amplifiers, finite resolution and quantization errors in phase control, unmodeled parasitic couplings among transmitter coils, and sensor noise, all of which were not accounted for in the idealized simulation model. Crucially, despite the compounded effects of these hardware imperfections, the proposed method consistently achieves a high fraction of the

empirical optimum power, underscoring its practical robustness. This demonstrates that the dual-current state representation remains effective in capturing essential system state information from noisy real-world measurements, which enables the agent to make sound decisions. The experiment successfully validates the transfer of the learned policy from simulation to reality.

The actual power consumption of the computing platform is strongly deployment-dependent and is therefore not explicitly benchmarked in this work; however, the online computational requirement of the proposed method is limited to lightweight policy inference.

6. Discussion

Representative MTSR-WPT optimization studies have mainly followed model-based routes. In particular, refs. [18,19] target maximum-efficiency operation by explicitly exploiting transmitter-current relationships, load tuning, or magnetic-field shaping based on coupling-dependent models, whereas [20] addresses maximum-power tracking through magnetic-field editing and current synthesis. By comparison, the present work focuses on load-power maximization under receiver-position uncertainty through direct transmitter-phase control. The distinction is not that existing methods are less valid, but that they typically rely on more explicit coupling models or current-design relationships. Here, the control policy is learned from directly measurable electrical responses, so explicit online mutual-inductance identification is avoided while phase commands are generated directly from the observed state.

The role of SAC in this paper should also be interpreted in the context of recent SAC-based power-electronics control studies such as [21,23,25]. These studies show that SAC can be effective for converter predictive control, voltage regulation, or controller-parameter adaptation, but they do not address the spatially varying coupling problem of MTSR-WPT under uncertain receiver positions. Therefore, the contribution here is not merely the use of SAC itself. Rather, it lies in combining a physics-informed dual-current state, direct phase control, and a deployment-oriented evaluation protocol that includes nominal performance, expanded-region generalization, measurement-noise robustness, model-mismatch recovery, and hardware validation.

The present study focuses on a quasi-static receiver position and a five-transmitter single-receiver prototype. Future work will investigate dynamic charging scenarios, more severe out-of-distribution landing positions, and extensions to larger transmitter arrays or multiple receivers. Hybrid schemes that combine model-based initialization with RL-based online adaptation are also of interest.

7. Conclusions

This paper presents a SAC-based phase-control method for improving transferred power in multi-transmitter wireless charging systems under uncertain receiver positions. In simulation on a five-transmitter system, the learned policy achieves about 97% of the theoretical maximum power in the training region and about 96% in the expanded evaluation region, while also improving the post-optimization power-transfer efficiency. The robustness studies further show that high performance is retained under 30–40 dB measurement noise, that degradation at 20 dB is gradual rather than catastrophic, and that most performance loss caused by model mismatch can be recovered through short fine-tuning in the shifted model.

Hardware validation on the physical prototype further confirms the practical feasibility of the method, yielding an 188% average power improvement and reaching about 87% of the maximum power measured on the platform. Taken together, these results support the

proposed framework as a practical data-driven alternative to model-dependent MTSR-WPT optimization when explicit online parameter identification is undesirable.

Author Contributions: Conceptualization, Z.D. and Y.Y.; methodology, Z.D.; software, Z.D., Z.L. and G.Y.; validation, Z.D. and Y.L.; investigation, Z.D. and Y.L.; data curation, Z.D. and G.Y.; writing—original draft preparation, Z.D. and Y.L.; writing—review and editing, Y.Y. and Z.D.; visualization, Z.D. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Laboratory Foundation (no. WDZC20255290304).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

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

Abbreviations

The following abbreviations are used in this manuscript:

UAV	unmanned aerial vehicle
WPT	magnetically coupled resonant wireless power transfer
MTSR	multi-transmitter single-receiver
RL	reinforcement learning
SAC	Soft Actor-Critic
KVL	Kirchhoff's voltage law
PTE	power transfer efficiency
SNR	signal-to-noise ratio
5T1R	five-transmitter single-receiver

References

1. Jung, H.; Lee, B. Wireless Power and Bidirectional Data Transfer System for IoT and Mobile Devices. *IEEE Trans. Ind. Electron.* **2022**, *69*, 11832–11836. [CrossRef]
2. Liu, Z.; Wang, J.; Lim, E.G.; Leach, M.; Wang, Z.; Pei, R.; Jiang, Z.; Zhang, W.; Huang, Y. Efficiency-Enhanced Wireless Power Transfer System Featuring a Pattern-Reconfigurable Antenna for Mobile Charging. *IEEE Antennas Wirel. Propag. Lett.* **2025**, *24*, 4242–4246. [CrossRef]
3. Lee, C. Topology Optimization of the Transmitter Ferrite and Receiver Coil for Minimizing the Weight of the Wireless-Charging Portable Devices. *IEEE Trans. Ind. Electron.* **2024**, *71*, 12192–12201. [CrossRef]
4. Zhu, H.; Wu, X.; Xiong, F.; Tahir, M. Robust Wireless Power and Data Transmission System against Misalignment for Implantable Medical Devices. *IEEE Trans. Power Electron.* **2025**, *40*, 14169–14180. [CrossRef]
5. Ma, Y.; Sun, Y.; Cui, K.; Fan, X. A 6.78-MHz Digital Rectifier-Based Single-Stage Wireless Charger Using Digital-Controlled CC–CV Technique for Implantable Biomedical Devices. *IEEE Trans. Power Electron.* **2023**, *38*, 101–106. [CrossRef]
6. Wu, Y.; Pan, W.; Xu, W.; Xie, R.; Zhuang, Y.; Mao, X.; Zhang, Y. An Integrated Charger of Wireless Power Transfer, Onboard Charger, and Auxiliary Power Module for Electric Vehicles. *IEEE Trans. Power Electron.* **2025**, *40*, 6334–6344. [CrossRef]
7. Meira Gomes, Z.; Prado, E.D.O.; Le Gall, Y.; Damm, G.; Ripoll, C.; Pinheiro, J.R. Design, Model, and Control of a Dynamic Wireless Power Transfer System for a 30-kW Electric Vehicle Charger Application. *IEEE J. Emerg. Sel. Top. Power Electron.* **2025**, *13*, 3882–3894. [CrossRef]
8. Chu, S.Y.; Cui, X.; Zan, X.; Avestruz, A. Transfer-Power Measurement Using a Non-Contact Method for Fair and Accurate Metering of Wireless Power Transfer in Electric Vehicles. *IEEE Trans. Power Electron.* **2022**, *37*, 1244–1271. [CrossRef]
9. Gu, Y.; Wang, J.; Liang, Z.; Zhang, Z. Flexible Constant-Power Range Extension of Self-Oscillating System for Wireless In-Flight Charging of Drones. *IEEE Trans. Power Electron.* **2024**, *39*, 15342–15355. [CrossRef]
10. Citroni, R.; Mangini, F.; Frezza, F. Efficient Integration of Ultra-low Power Techniques and Energy Harvesting in Self-Sufficient Devices: A Comprehensive Overview of Current Progress and Future Directions. *Sensors* **2024**, *24*, 4471. [CrossRef]
11. Dai, X.; Li, X.; Li, Y.; Deng, P.; Tang, C. A Maximum Power Transfer Tracking Method for WPT Systems with Coupling Coefficient Identification Considering Two-Value Problem. *Energies* **2017**, *10*, 1665. [CrossRef]
12. Zhang, Y.; Yan, Z.; Liang, Z.; Li, S.; Mi, C.C. A High-Power Wireless Charging System Using *LCL*-N Topology to Achieve a Compact and Low-Cost Receiver. *IEEE Trans. Power Electron.* **2020**, *35*, 131–137. [CrossRef]

13. Li, J.; Liu, X.; Leung, K.N. A 24-to-240 W 95.6%-Efficiency <300-µs-Settling-Time Hybrid MCR/PT Wireless Power Transfer System. *IEEE Trans. Power Electron.* **2024**, *39*, 8928–8946. [CrossRef]
14. Li, Y.; Zhang, B.; Zhai, Y.; Wang, H.; Yuan, B.; Lou, Z. A Novel Type of 3-D Transmitter for Omnidirectional Wireless Power Transfer. *IEEE Trans. Power Electron.* **2024**, *39*, 6537–6548. [CrossRef]
15. Feng, J.; Li, Q.; Lee, F.C.; Fu, M. Transmitter Coils Design for Free-Positioning Omnidirectional Wireless Power Transfer System. *IEEE Trans. Ind. Inform.* **2019**, *15*, 4656–4664. [CrossRef]
16. Pahlavan, S.; Shooshtari, M.; Jafarabadi Ashtiani, S. Star-Shaped Coils in the Transmitter Array for Receiver Rotation Tolerance in Free-Moving Wireless Power Transfer Applications. *Energies* **2022**, *15*, 8643. [CrossRef]
17. Zhu, Q.; Su, M.; Sun, Y.; Tang, W.; Hu, A.P. Field Orientation Based on Current Amplitude and Phase Angle Control for Wireless Power Transfer. *IEEE Trans. Ind. Electron.* **2018**, *65*, 4758–4770. [CrossRef]
18. Kim, D.; Ahn, D. Maximum Efficiency Point Tracking for Multiple-Transmitter Wireless Power Transfer. *IEEE Trans. Power Electron.* **2020**, *35*, 11391–11400. [CrossRef]
19. Zhu, Z.; Yuan, H.; Liang, C.; Wang, C.; Lv, S.; Yang, A.; Chu, J.; Rong, M.; Wang, X.; Hu, A.P. Maximum Efficiency Tracking of a Wireless Power Transfer System With 3-D Coupling Capability Using a Planar Transmitter Coil Configuration. *IEEE Trans. Power Electron.* **2024**, *39*, 10594–10604. [CrossRef]
20. Tian, X.; Chau, K.T.; Liu, W.; Pang, H.; Lee, C.H.T. Maximum Power Tracking for Magnetic Field Editing-Based Omnidirectional Wireless Power Transfer. *IEEE Trans. Power Electron.* **2022**, *37*, 12901–12912. [CrossRef]
21. Liu, C.; Ma, J.; Liu, X.; Qiu, L.; Wu, W.; Fang, Y. A Predictive Control Method Based on Neural Predictor and Soft Actor–Critic for Power Converters. *IEEE Trans. Ind. Electron.* **2025**, *72*, 4556–4566. [CrossRef]
22. Yang, H.; Chen, Q.; Shi, X.; Xu, Y.; Zhang, X. Fast Charging Management of a Lithium-Ion Battery and Cooling System: A Stackelberg Game-Based Soft Actor Critic−Deep Reinforcement Learning Method. *IEEE Trans. Ind. Electron.* **2025**, *72*, 11347–11359. [CrossRef]
23. Ye, J.; Zhao, D.; Pan, X.; Li, S.; Wang, B.; Zhang, X.; Iu, H.H.C. Improving Voltage Regulation of Interleaved DC–DC Boost Converter via Soft Actor–Critic Algorithm-Based Reinforcement Learning Controller. *IEEE J. Emerg. Sel. Top. Power Electron.* **2025**, *13*, 5958–5969. [CrossRef]
24. Zeng, Y.; Pou, J.; Sun, C.; Maswood, A.I.; Dong, J.; Mukherjee, S.; Gupta, A.K. Multiagent Deep Reinforcement Learning-Aided Output Current Sharing Control for Input-Series Output-Parallel Dual Active Bridge Converter. *IEEE Trans. Power Electron.* **2022**, *37*, 12955–12961. [CrossRef]
25. Zeng, Y.; Liang, G.; Liu, Q.; Rodriguez, E.; Pou, J.; Jie, H.; Liu, X.; Zhang, X.; Kotturu, J.; Gupta, A. Multi-Agent Soft Actor-Critic Aided Active Disturbance Rejection Control of DC Solid-State Transformer. *IEEE Trans. Ind. Electron.* **2025**, *72*, 492–503. [CrossRef]
26. Yong, H.; Seo, J.; Kim, J.; Kim, M.; Choi, J. Suspension Control Strategies Using Switched Soft Actor-Critic Models for Real Roads. *IEEE Trans. Ind. Electron.* **2023**, *70*, 824–832. [CrossRef]
27. Xu, C.; Wang, J.; Ding, Y.; Zheng, C. UAV power line inspection strategy based on SAC algorithm. *Electr. Power Syst. Res.* **2025**, *248*, 111925. [CrossRef]
28. Yuan, Q. S-Parameters for Calculating the Maximum Efficiency of a MIMO-WPT System: Applicable to Near/Far Field Coupling, Capacitive/Magnetic Coupling. *IEEE Microw. Mag.* **2023**, *24*, 40–48. [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.