aerospace

MDPI

Article

Robust Reinforcement Learning-Based Guidance Strategy Against Cyber-Attacks

Yuting Shang and Yuanli Cai *

The School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China; yutingshang@stu.xjtu.edu.cn
* Correspondence: ylicai@mail.xjtu.edu.cn

Abstract

As electronic countermeasure techniques become increasingly sophisticated, missiles that rely on ground-based systems to generate and transmit guidance commands face the risk of cyber-attacks. Most existing guidance strategies do not account for such attacks, resulting in significant degradation of guidance performance once attacks occur. To address this issue, this paper proposes a robust reinforcement learning-based guidance strategy that incorporates cyber-attack factors. First, the kinematic and relative-motion models for the missile-maneuvering target engagement are established, and typical cyber-attack scenarios are analyzed. Based on these models, a robust Markov decision process incorporating cyber-attack effects is formulated, providing a theoretical framework for applying robust reinforcement learning to guidance problems. Within this framework, a robust reinforcement learning-based guidance strategy with a robust Actor–robust Critic architecture is developed. Specifically, the robust Critic updates its parameters by minimizing the mean-squared robust temporal-difference error. Meanwhile, the robust Actor updates the policy using the clipped objective of robust Proximal Policy Optimization. Together, these updates improve the robustness of the guidance policy under cyber-attacks. Extensive simulation results demonstrate that the proposed guidance strategy successfully intercepts maneuvering targets under diverse attack conditions, validating its robustness and guidance effectiveness under cyber-attacks.

Keywords: missile guidance; maneuvering target; cyber-attack; robust reinforcement learning

1. Introduction

With the continuous advancement of aerospace technology, missile guidance has become one of the major research topics in the aerospace field [1,2]. Traditional missile guidance methods have primarily focused on the interception of non-maneuvering targets [3,4]. However, with the rapid development of penetration technologies, target maneuverability has been significantly enhanced, while the guidance environment has become increasingly complex and unpredictable. These factors substantially increase the difficulty of missile guidance and degrade interception performance. Consequently, the missile guidance problem for maneuvering target interception has attracted growing attention in recent years [5,6]. To address this challenge, a variety of control methods have been proposed and successfully applied, including, but not limited to, proportional navigation [7,8], sliding mode control [9,10], differential game theory [11,12], and reinforcement learning [13,14].

Most existing guidance methods are developed under an idealized assumption that cyber-attacks do not occur during the guidance process [1,3,4]. However, with the rapid

Academic Editors: Vladimir S. Aslanov and Changzhu Wei

Received: 11 June 2026
Revised: 11 August 2026
Accepted: 17 August 2026
Published: 19 August 2026

advancement of electronic warfare technologies, this assumption is no longer valid in combat scenarios that rely on data-link communications. In particular, under the Track-Via-Missile (TVM) guidance mode illustrated in Figure 1, the guidance process relies heavily on wireless data links. This guidance mode has been widely adopted in representative air defense and missile defense systems, such as the U.S. Patriot-2 missile system. In the TVM guidance mode, sensing information is transmitted from the missile to ground-based systems through the downlink, while guidance commands are sent from ground-based systems to the in-flight missile through the uplink [15]. Due to its strong dependence on wireless communication links, the reliability and accuracy of critical guidance information may be compromised once the communication link is subjected to cyber-attacks. This can impair the missile's decision-making and guidance performance and potentially lead to mission failure. Under such circumstances, conventional guidance strategies, which do not explicitly account for cyber-attacks, often suffer significant performance degradation and cannot guarantee the desired interception effectiveness. Therefore, it is necessary to redesign or enhance guidance strategies to improve the robustness and reliability of guidance systems operating under cyber-attacks.

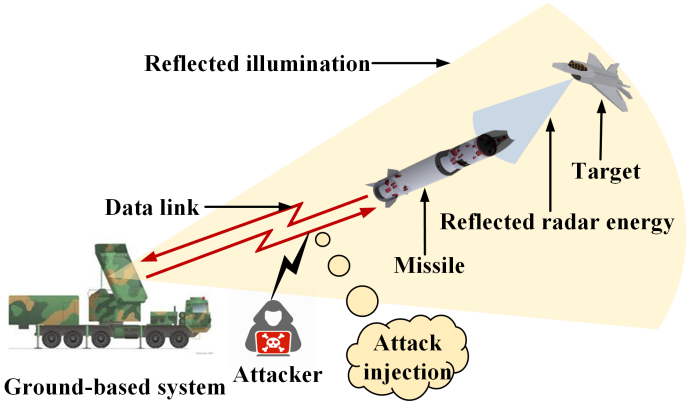


Figure 1. TVM guidance.

Currently, studies on missile guidance under cyber-attacks remain relatively limited [16–18]. For example, Wang et al. [16] modeled false data injection attacks on the magnitude of guidance commands as additive disturbances and developed a secure guidance strategy by estimating the attack signals using a super-twisting observer. Li et al. [17] proposed a leader-follower cooperative guidance scheme based on distributed observers to counter false data injection attacks, enabling multiple vehicles to achieve synchronized target interception in the presence of false data injection. Although these studies considered the complexities introduced by cyber-attacks and estimated the attack signals using observers, the estimated values were incorporated as compensation terms into the control inputs to construct cyber-resilient guidance strategies based on conventional control theory. However, most conventional control methods generally rely on accurate system models [19,20]. Owing to the strong nonlinearity and strong coupling inherent in guidance systems, it is extremely difficult to establish highly accurate mathematical models.

As a data-driven intelligent learning paradigm, reinforcement learning is capable of autonomously learning guidance policies through continuous interactions with the guidance environment [13,21]. This characteristic enables reinforcement learning to exhibit significant advantages in complex decision-making problems where accurate system models are difficult to obtain. However, the cyber-attack estimation process inevitably introduces estimation errors, which are subsequently propagated into the guidance decision-making process. As a result, the adverse effects of cyber-attacks cannot be effectively mitigated, leading to degraded guidance performance. Moreover, conventional reinforcement learn-

ing methods generally lack explicit robustness mechanisms for handling model uncertainties and external disturbances. Consequently, when biases or inaccuracies exist in cyber-attack estimation, the learned policies often struggle to maintain stable and reliable guidance performance.

As a branch of reinforcement learning, robust reinforcement learning aims to learn stable and reliable decision-making policies in the presence of model uncertainties, environmental disturbances, or adversarial factors [22,23]. Unlike conventional reinforcement learning, which primarily seeks to optimize the expected average performance, robust reinforcement learning places greater emphasis on maintaining satisfactory performance and safety under worst-case conditions. In recent years, robust reinforcement learning has achieved remarkable progress in the aerospace field [24–26]. For example, Choi et al. [24] proposed an adaptive robust Markov decision process framework for intelligence, surveillance, and reconnaissance missions of collaborative combat aircraft, enabling more efficient and safer decision-making while ensuring robustness. Deshpande et al. [25] employed a robust Markov decision process framework to train control policies for quadrotor unmanned aerial vehicles, improving the robustness and generalization capability of the learned policies across different environments through pessimistic optimization. Nevertheless, the application of robust reinforcement learning to missile guidance for intercepting maneuvering targets under cyber-attacks remains largely unexplored.

In recent years, significant progress has been made in the theory of robust reinforcement learning [23,27,28]. For example, Neufeld et al. [27] proposed a robust Q-learning algorithm based on Wasserstein-ball uncertainty modeling, which is capable of maintaining optimal control performance and theoretical convergence guarantees in the presence of transition probability estimation errors. Wang et al. [28] developed a robust policy gradient method with global optimality guarantees and complexity analysis under model mismatch conditions, and further extended it to a model-free robust Actor-critic framework for stable and efficient policy learning. However, most existing robust reinforcement learning methods have been studied in tabular settings, where the algorithmic design relies on the assumption of finite and low-dimensional state spaces. Consequently, these methods are difficult to apply directly to complex decision-making problems with large state spaces, such as missile guidance. Specifically, robust reinforcement learning typically requires the estimation of the robust Bellman operator, which involves solving an inner optimization problem over an uncertainty set. When commonly used uncertainty sets are adopted, such as f-divergence-based uncertainty sets [29], R-contamination models [28], and l_p norm-based uncertainty sets [30], a separate inner optimization problem must generally be solved for each state using samples drawn from the nominal model to obtain an unbiased estimate of the robust Bellman operator. As the state space grows, this procedure incurs a substantial computational burden. This may limit the practical applicability of existing robust reinforcement learning methods to large-scale and complex decision-making problems.

Motivated by the above discussion, a robust reinforcement learning-based guidance strategy against cyber-attacks is proposed to effectively mitigate the adverse effects of cyber-attacks on guidance performance. The main contributions of this work are summarized as follows:

(1) The missile guidance problem under cyber-attacks is formulated as a robust Markov decision process. By constructing state representations and reward mechanisms that explicitly capture the effects of cyber-attacks, the proposed robust reinforcement learning framework effectively suppresses attack-induced disturbances, thereby enhancing the missile's capability to intercept targets in a stable and reliable manner.

(2) Compared with existing studies [16–18], a robust reinforcement learning-based guidance strategy is developed based on a robust Actor–robust Critic learning architecture to effectively enhance guidance performance and improve interception success rates in cyber-attack environments.

(3) Compared with existing studies [23,27,28], a double-sampling uncertainty set is introduced to effectively characterize uncertainties in state transitions. The proposed uncertainty set enables an efficient and unbiased estimation of the robust Bellman operator using only samples collected from the nominal model, thereby facilitating policy learning in missile guidance problems with large state spaces.

(4) Compared with existing studies [23,27,28], the robust Critic updates its parameters by minimizing the mean-squared robust temporal-difference error, while the robust Actor performs policy updates using a robust Proximal Policy Optimization (RPPO) clipping objective. This learning mechanism improves policy optimization efficiency while maintaining training stability, thereby further enhancing robustness and guidance performance under cyber-attacks.

The remainder of this paper is organized as follows. Section 2 describes the missile guidance problem under cyber-attacks. Section 3 introduces the fundamentals of robust reinforcement learning. Section 4 presents the proposed robust reinforcement learning-based guidance strategy. Section 5 discusses the training and testing results. Finally, Section 6 summarizes the conclusions.

2. Guidance Problem Formulation Under Cyber-Attacks

This section introduces the missile guidance problem under cyber-attacks, in which a missile subjected to cyber-attack disturbances attempts to intercept a target moving according to a predefined maneuvering pattern. In this problem, cyber-attacks degrade the guidance performance of the missile, thereby increasing the difficulty and uncertainty of the interception mission.

2.1. Missile-Target Engagement Scenario

During the guidance process, the relative motion between the missile and the target can be decomposed into longitudinal plane motion and lateral plane motion. To focus on the design of the guidance strategy, only a two-dimensional engagement scenario is considered. Furthermore, to simplify the problem analysis, the following commonly adopted assumptions are introduced [31]:

Assumption 1. *The Earth is regarded as an ideal sphere, and the effect of its rotational angular velocity is neglected.*

Assumption 2. *Both the missile and the target are modeled as controllable point masses.*

Assumption 3. *The velocities of both the missile and the target are assumed to be constant.*

Figure 2 illustrates the schematic of the missile-target engagement scenario, where M and T denote the missile and the target, respectively. The velocities of the missile and the target are represented by v_M and v_T, respectively. The normal accelerations of the missile and the target are denoted by a_M and a_T, respectively. The flight-path angles of the missile and the target are represented by θ_M and θ_T, respectively. The relative distance between the missile and the target is denoted by r_{MT}, while the line-of-sight (LOS) angle between them is denoted by q_{MT}.

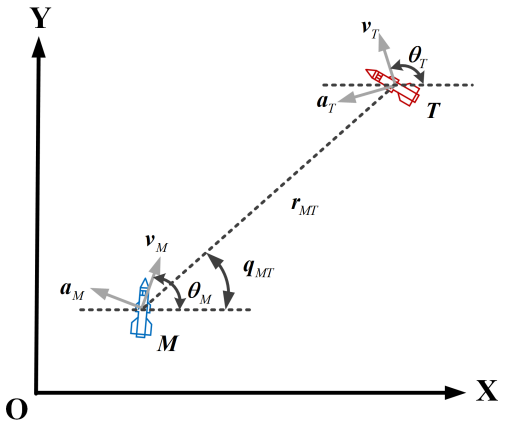


Figure 2. The schematic of the missile-target engagement scenario.

2.2. Relative Kinematic Equations

According to the missile-target engagement scenario shown in Figure 2, the kinematic equations of the missile and the target can be expressed as [32]:

$$\begin{cases} \dot{x}_i = v_i \cos\theta_i \\ \dot{y}_i = v_i \sin\theta_i \\ \dot{\theta}_i = a_i / v_i \end{cases} \quad i = M, T \tag{1}$$

where (x_M, y_M) and (x_T, y_T) denote the position coordinates of the missile and the target along the X- and Y-axes, respectively.

Based on the kinematic equations of the missile M and the target T, the relative kinematic equations between them can be expressed as follows:

$$\begin{aligned} \dot{r}_{MT} &= v_T \cos(\theta_T - q_{MT}) - v_M \cos(\theta_M - q_{MT}) \\ r_{MT}\dot{q}_{MT} &= v_T \sin(\theta_T - q_{MT}) - v_M \sin(\theta_M - q_{MT}) \end{aligned} \tag{2}$$

2.3. Cyber-Attack Model

In the missile-target engagement scenario shown in Figure 2, when the missile guidance commands generated by the ground system and transmitted through the wireless communication link are subjected to malicious attacks, the normal acceleration of the missile can be expressed as [17]

$$a_M = a_G + a_A \tag{3}$$

where a_G denotes the guidance command generated by the ground system, and a_A denotes the deceptive acceleration command.

2.4. Robust Markov Decision Process Framework

To characterize the missile guidance problem under the combined effects of cyber-attacks and target maneuvers, a robust Markov decision process framework is introduced on the basis of the Markov decision process to describe the influence of state transition uncertainty on decision-making.

A standard Markov decision process is represented by the tuple $(\boldsymbol{S}, \boldsymbol{A}, \kappa, R, \gamma)$ [33], where $\boldsymbol{S}$ is the state space, $\boldsymbol{A}$ is the action space, $\kappa = (\kappa_0, \kappa_1, \cdots)$ is a possibly non-stationary sequence of transition kernels with $\kappa_t : \boldsymbol{S} \times \boldsymbol{A} \to \Delta_{\boldsymbol{S}}$, $R : \boldsymbol{S} \times \boldsymbol{A} \to [0, 1]$ is the reward function, and $\gamma \in [0, 1]$ is the discount factor.

In the presence of model uncertainty, a single transition kernel is insufficient to accurately describe the system dynamics. Therefore, a robust Markov decision process is

introduced, which is represented by the tuple $(\boldsymbol{S}, \boldsymbol{A}, P, R, \gamma)$ [34]. Here, P is a set of transition kernels, referred to as the uncertainty set, which is used to capture perturbations around the nominal transition kernel $p^{\circ} : \boldsymbol{S} \times \boldsymbol{A} \to \Delta_{\boldsymbol{S}}$.

Under this framework, the missile guidance problem under cyber-attacks can be uniformly formulated as a modeling problem involving the state, action, reward function, and uncertainty set. The key lies in properly constructing these elements so as to accurately reflect the relative motion characteristics between the missile and the target, as well as environmental uncertainty.

2.4.1. State Definition

In the robust Markov decision process, to accurately characterize the interaction between the missile and the target, the defined state $\boldsymbol{s}$ contains the key information of their relative motion as well as disturbance information, which is specifically given as follows:

$$\boldsymbol{s} = [r_{MT}, \dot{r}_{MT}, q_{MT}, \dot{q}_{MT}, \hat{w}] \tag{4}$$

where $\hat{w}$ denotes the disturbance estimation associated with cyber-attacks and target maneuvers obtained through the existing observer (see Appendix A) [17].

To improve the training efficiency, the states are normalized as follows [35]:

$$\boldsymbol{s} = \left[\frac{r_{MT}}{r_{MT\max}}, \frac{\dot{r}_{MT}}{\dot{r}_{MT\max}}, \frac{q_{MT}}{q_{MT\max}}, \frac{\dot{q}_{MT}}{\dot{q}_{MT\max}}, \frac{\hat{w}}{\hat{w}_{\max}} \right] \tag{5}$$

where $r_{MT\max}$ denotes the maximum value of r_{MT}, $q_{MT\max}$ denotes the maximum value of q_{MT}, and $\hat{w}_{\max}$ denotes the maximum value of $\hat{w}$.

2.4.2. Action Selection

In the robust Markov decision process, to accurately reflect the control variable executable by the missile, the action $\boldsymbol{a}$ is defined as the missile guidance command generated by the ground system, which can be expressed as follows:

$$\boldsymbol{a} = [a_G] \tag{6}$$

To accelerate the training process, the action $\boldsymbol{a}$ is normalized as follows [35]:

$$\boldsymbol{a} = \left[\frac{a_G}{a_{G\max}} \right] \tag{7}$$

where $a_{G\max}$ denotes the maximum value of a_G.

2.4.3. Reward Design

In the robust Markov decision process, to ensure that the missile effectively intercepts the target, the reward function r is designed as follows:

$$r = r_{\text{immediate}} + r_{\text{terminal}} \tag{8}$$

where $r_{\text{immediate}}$ is the immediate reward obtained during the guidance process, and r_{terminal} is the terminal reward obtained at the end of the guidance process.

To improve the guidance performance, the immediate reward $r_{\text{immediate}}$ consists of two components. The first component is the distance reward, which characterizes the relative error between the missile and the target. Its value increases monotonically as the relative distance decreases, thereby continuously guiding the missile toward the target during policy optimization. The second component is the LOS angular-rate reward, which

represents the intensity of LOS direction variation. Its value increases as the LOS angular rate decreases, thereby suppressing rapid LOS rotation and enhancing the stability of the guidance process and the target tracking capability

$$r_{\text{immediate}} = k_r \exp\left(-\left(\frac{r_{MT}}{r_{MT\max}}\right)^2\right) + k_q \exp\left(-\left(\frac{\dot{q}_{MT}}{\dot{q}_{MT\max}}\right)^2\right) \tag{9}$$

where k_r and k_q denote the weighting coefficients associated with the relative distance and the LOS angular rate, respectively.

When the missile successfully intercepts the target, a positive reward is assigned; otherwise, no reward is provided. Therefore, the terminal reward r_{terminal} is defined as follows:

$$r_{\text{terminal}} = \begin{cases} k_t, if\ r_{MT} \le r_{MT\text{m}} \\ 0, if\ r_{MT} > r_{MT\text{m}} \end{cases} \tag{10}$$

where k_t denotes the weighting coefficient of the terminal reward, and $r_{MT\text{m}}$ denotes the maximum miss distance allowable for successful target interception by the missile.

2.4.4. Uncertainty Set Construction

In the robust Markov decision process, the uncertainty set is used to accurately characterize the uncertainty in state transition probabilities. To effectively solve decision-making problems with large state spaces, the uncertainty set is constructed as a double-sampling uncertainty set. Its specific definition and solution procedure will be presented in the subsequent sections.

Remark 1. *The cyber-attack acts on the missile dynamics in the form of an additional acceleration command, thereby altering the relative motion between the missile and the target. When the attack parameters are random or uncertain, the state at the next time step may have multiple possible outcomes, which is reflected as uncertainty in the state-transition probabilities of the robust Markov decision process.*

3. Robust Reinforcement Learning

Definition 1. *For an arbitrary set χ, let $|\chi|$ denote its cardinality, and let Δ_χ denote the $(|\chi|-1)$-dimensional probability simplex over χ. Let* $\text{Unif}(\chi)$ *denote the uniform distribution on χ. In addition, define $[m] := \{1, \cdots, m\}$.*

Under the framework of a Markov decision process, for a stationary policy $\pi : \boldsymbol{S} \to \Delta_{\boldsymbol{A}}$, the value function is defined as follows:

$$V_\kappa^\pi(\boldsymbol{s}) := \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t | \boldsymbol{s}_0 = \boldsymbol{s}\right] \tag{11}$$

where $r_t = R(\boldsymbol{s}_t, \boldsymbol{a}_t)$ denotes the reward received at time step t, $\boldsymbol{s}_t$ denotes the state at time step t, and $\boldsymbol{a}_t$ denotes the action taken at time step t. The expectation is taken over trajectories $(\boldsymbol{s}_0, \boldsymbol{a}_0, \boldsymbol{s}_1, \boldsymbol{a}_1, \ldots)$ satisfying $\boldsymbol{a}_t | \boldsymbol{s}_t \sim \pi(\boldsymbol{s}_t)$ and $\boldsymbol{s}_{t+1} | (\boldsymbol{s}_t, \boldsymbol{a}_t) \sim \kappa_t(\boldsymbol{s}_t, \boldsymbol{a}_t)$. Similarly, under the initial state distribution $\rho \in \Delta_S$, the discounted state visitation distribution is defined as follows:

$$d_\rho^{\pi,\kappa}(\boldsymbol{s}) := (1-\gamma)\mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}(\boldsymbol{s}_t = \boldsymbol{s}) | \boldsymbol{s}_0 \sim \rho\right] \tag{12}$$

The state–action value function (i.e., the Q-function) is defined as follows:

$$Q_{\kappa}^{\pi}(\boldsymbol{s},\boldsymbol{a}) := \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}r_{t}|\boldsymbol{s}_{0}=\boldsymbol{s},\boldsymbol{a}_{0}=\boldsymbol{a}\right] \tag{13}$$

The advantage function is defined as follows:

$$A_{\kappa}^{\pi}(\boldsymbol{s},\boldsymbol{a}) := Q_{\kappa}^{\pi}(\boldsymbol{s},\boldsymbol{a}) - V_{\kappa}^{\pi}(\boldsymbol{s}) \tag{14}$$

Under the framework of a robust Markov decision process, the robust value function is defined as the value under the corresponding worst-case transition dynamics

$$V_{\mathcal{P}}^{\pi}(\boldsymbol{s}) := \inf_{\kappa\in\otimes_{t\geq 0}\mathcal{P}} V_{\kappa}^{\pi}(\boldsymbol{s}) \tag{15}$$

The corresponding robust Bellman operator, denoted by $\mathcal{T}_{\mathcal{P}}^{\pi}$, is defined as follows:

$$(\mathcal{T}_{\mathcal{P}}^{\pi}V)(\boldsymbol{s}) = \mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s})}\left[r+\gamma\inf_{p\in\mathcal{P}_{\boldsymbol{s},\boldsymbol{a}}} p^{\top}V\right], \forall \boldsymbol{s}\in\boldsymbol{S}, V\in\mathbb{R}^{\boldsymbol{S}} \tag{16}$$

where p denotes the state transition probability distribution, $\mathcal{P}_{\boldsymbol{s},\boldsymbol{a}}$ denotes a ball $\{v\subseteq\Delta_{\boldsymbol{S}} : d\left(v, p_{\boldsymbol{s},\boldsymbol{a}}^{\circ}\right)\leq\delta\}$ centered at the nominal model p° of the training environment, $d(\cdot,\cdot)$ is a distance metric between probability distributions, and $\delta>0$ is a parameter that controls the degree of robustness. The corresponding Bellman equation is given by $V=\mathcal{T}_{\mathcal{P}}^{\pi}V$, where $V=V^{\pi}$ is the unique fixed point of the robust Bellman operator, whose existence and uniqueness are guaranteed by the Banach fixed-point theorem.

There exists an optimal stationary policy π^{*} that maximizes the robust value function uniformly over all admissible transition models [36], i.e.,

$$V_{\mathcal{P}}^{\pi^{*}}(\boldsymbol{s}) = \sup\{V_{\mathcal{P}}^{\pi}(\boldsymbol{s}) : \text{history dependent } \pi\}, \forall \boldsymbol{s} \tag{17}$$

Therefore, without loss of generality, the optimization can be restricted to the space of stationary policies. Moreover, for any stationary policy π, there always exists a stationary worst-case transition kernel κ_{π} such that $V_{\kappa_{\pi}}^{\pi}(\boldsymbol{s}) = V_{\mathcal{P}}^{\pi}(\boldsymbol{s}), \forall \boldsymbol{s}$. The robust Q-function and the robust advantage function are defined, respectively, as follows:

$$Q_{\mathcal{P}}^{\pi}(\boldsymbol{s},\boldsymbol{a}) := Q_{\kappa_{\pi}}^{\pi}(\boldsymbol{s},\boldsymbol{a}) \tag{18}$$

$$A_{\mathcal{P}}^{\pi}(\boldsymbol{s},\boldsymbol{a}) := A_{\kappa_{\pi}}^{\pi}(\boldsymbol{s},\boldsymbol{a}) \tag{19}$$

In summary, the motivation of the robust reinforcement learning framework is to learn an optimal robust policy by training solely on a simulator that implements an (unknown) nominal model p° [37]. Under this framework, the robust reinforcement learning algorithm has access only to data generated by p° and cannot access data from any other models contained in the uncertainty set $\mathcal{P}$.

4. Robust Reinforcement Learning-Based Guidance Strategy

4.1. Overview

Under the robust Markov decision process framework, this paper proposes a robust reinforcement learning-based guidance strategy to achieve effective target interception in the presence of cyber-attacks, as illustrated in Figure 3. The proposed framework consists of three key modules: the guidance simulation module, the sample collection module, and the policy training module. The guidance simulation module shown in Figure 3a is used to simulate the missile guidance process under cyber-attack disturbances and to generate a

large number of samples for subsequent training. During the simulation process, the missile selects an action $\boldsymbol{a}$ according to the current state $\boldsymbol{s}$. After executing action $\boldsymbol{a}$, the missile interacts with the nominal environment, transitions to the next state $\boldsymbol{s}'$, and receives a reward r. This interaction process is repeated until a sufficient number of samples have been collected. The sample collection module shown in Figure 3b is used to store the generated data. During this stage, all samples are stored in an experience replay buffer for subsequent training. Specifically, $\mathbf{S}$, $\mathbf{A}$, $\mathbf{R}$, and $\mathbf{S}'$ are used to store the states, actions, rewards, and next states, respectively. The policy training module shown in Figure 3c is responsible for optimizing the network parameters. During training, mini-batches of samples are randomly drawn from the experience replay buffer and fed into the network. The network parameters are then updated according to the specified learning algorithm. Through iterative optimization, the performance of the guidance strategy is progressively improved until the termination criterion is satisfied.

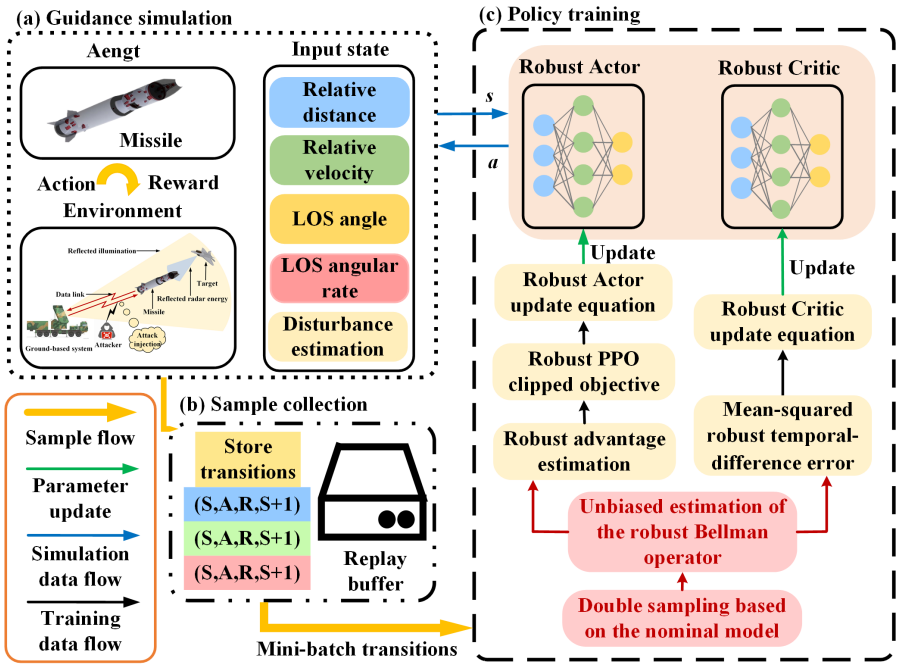


Figure 3. Robust reinforcement learning-based guidance strategy framework.

To enhance adaptability and robustness under cyber-attack disturbances, the proposed robust reinforcement learning-based guidance strategy employs two fully connected neural networks: a robust Actor and a robust Critic. The robust Actor determines the missile's action $\boldsymbol{a}$ based on the given state $\boldsymbol{s}$. The robust Critic evaluates the quality of the selected action by estimating the corresponding value function. During the training phase, a double-sampling uncertainty set is introduced to model the uncertainty in state transitions. This enables an efficient and unbiased estimation of the robust Bellman operator using only samples generated from the nominal model. Based on this estimation, the robust Critic updates its network parameters by minimizing the mean-squared robust temporal-difference error. The robust Actor updates its network parameters by maximizing the RPPO clipped objective. The detailed formulations of these two update procedures are provided in Sections 4.3 and 4.4.

4.2. Double-Sampling Uncertainty Set

When estimating the robust Bellman operator $\mathcal{T}_{\mathcal{P}}^{\pi}$, it is necessary to solve the optimization problem $\inf_{p\in\mathcal{P}_{s,a}} p^{\top}V$. For commonly used uncertainty sets, this optimization problem is generally intractable in large-scale state spaces. In addition, during practical guidance

applications, only data generated by the nominal model p° are accessible, which makes direct estimation of $\mathcal{T}_{\mathcal{P}}^{\pi}$ particularly challenging. For these reasons, existing uncertainty sets are often difficult to handle in robust Markov decision processes with large state spaces. To address this issue, inspired by the idea of constructing unbiased estimators through sampling [38], a double-sampling uncertainty set is introduced. This uncertainty set not only yields a tractable form of the robust Bellman operator, but also enables its unbiased estimation using samples generated solely from the nominal model, thereby facilitating efficient learning.

To construct a tractable robust Bellman operator, the following assumptions are introduced. First, it is assumed that the nominal transition probability model $p^\circ_{s,a}$ can be sampled through an environment simulator. Specifically, for any state–action pair (s, a), next-state samples following the distribution $p^\circ_{s,a}$ can be obtained. Second, it is assumed that the uncertainty set satisfies the state–action separable form $\mathcal{P} = \otimes_{s,a}\mathcal{P}_{s,a}$, indicating that the transition probability uncertainties associated with different state–action pairs can be modeled independently.

Let$\{s_1', s_2', \cdots, s_k'\}$ be k independent and identically distributed samples drawn from the nominal model $p^\circ_{s,a}$. For any given divergence measure $d(\cdot,\cdot)$ and radius $\delta > 0$, there exists an uncertainty set $\mathcal{P} = \otimes_{s,a}\mathcal{P}_{s,a}$ such that

$$\inf_{p\in\mathcal{P}_{s,a}} p^\top V = \mathbb{E}_{s'_{1:k} \overset{i.i.d.}{\sim} p^\circ_{s,a}}\left[\inf_{\alpha\in\Delta_{[k]}:d(\alpha,\inf([k]))\le\delta}\sum_{i=1}^{k}\alpha_i V(s_i')\right] \tag{20}$$

By optimizing the worst-case probability weights in (20) within the sample space, the original transition probability uncertainty optimization problem is transformed into a finite-dimensional optimization problem. This can avoid the direct solution of an infinite-dimensional optimization problem over probability distributions. Therefore, the empirical robust Bellman operator corresponding to (16), denoted by $\hat{\mathcal{T}}_{\mathcal{P}}^{\pi}$, is defined as [23]

$$\left(\hat{\mathcal{T}}_{\mathcal{P}}^{\pi}V\right)(s, a, s'_{1:k}) := r + \gamma \inf_{\alpha\in\Delta_{[k]}:d(\alpha,\mathrm{Unif}([k]))\le\delta}\sum_{i=1}^{k}\alpha_i V(s_i') \tag{21}$$

where the second term in (21) represents the worst-case expected value that the system may encounter under transition probability uncertainties, which is used to characterize the robustness of the policy against environmental perturbations. Based on (16) and (21), we have

$$(\mathcal{T}_{\mathcal{P}}^{\pi}V)(s) = \mathbb{E}_{a\sim\pi(\cdot|s),\, s'_{1:k} \overset{i.i.d.}{\sim} p^\circ_{s,a}}\left[(\hat{\mathcal{T}}_{\mathcal{P}}^{\pi}V)(s, a, s'_{1:k})\right] \tag{22}$$

(22) indicates that when the sampled states satisfy the independent and identically distributed condition the constructed estimator can serve as an unbiased estimate of the true robust Bellman operator.

When (22) holds, $\hat{\mathcal{T}}_{\mathcal{P}}^{\pi}$ provides an unbiased estimate. Since $\sum_{i=1}^{k}\alpha_i V(s_i')$ represents the expected value obtained by selecting a sample s_i' from $\{s_1', s_2', \cdots, s_k'\}$ and evaluating $V(s_i')$, this approach is referred to as double sampling. In this framework, when $\delta = 0$, $\alpha = \mathrm{Unif}([k])$, and $s_i' \sim p^\circ_{s,a}$, α can be interpreted as a perturbation of $\mathrm{Unif}([k])$. The uncertainty set associated with double sampling is defined implicitly by specifying k, $d(\cdot,\cdot)$, and δ. The key advantage of this construction is that all samples can be generated solely from the nominal model p°.

Double sampling requires generating multiple next states for each given state–action pair. This procedure can be readily implemented in a simulator that allows resetting the environment to any desired state. Since the computation of α is restricted to $\Delta_{[k]}$, the empirical robust Bellman operator remains computationally tractable when k is chosen to be mod-

erately sized. Under this setting, for most divergence measures $d(\cdot,\cdot)$, the corresponding solution can be derived explicitly as follows:

$$\left(\hat{\mathcal{T}}_{\mathcal{P}}^{\pi} V\right)\left(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}'_{1:k}\right) := r + \gamma\left(\frac{1}{k}\sum_{i=1}^{k} V\left(\boldsymbol{s}'_i\right) - \delta \max_{m,n\in[k]} \left|V\left(\boldsymbol{s}'_m\right) - V\left(\boldsymbol{s}'_n\right)\right|\right) \tag{23}$$

where the first term in (23) represents the average value estimation obtained from samples generated by the nominal transition model, while the second term denotes the robustness correction term introduced by transition probability uncertainty, which compensates for the potential impact of model mismatch on policy performance.

4.3. Robust Critic

To enable the robust Critic to estimate the robust value function as accurately as possible, the robust Critic is updated by minimizing the mean-squared robust temporal-difference error.

Following the classical temporal-difference learning algorithm and using the empirical robust Bellman operator [39,40], the parameters w of the robust Critic are updated by minimizing the mean-squared robust temporal-difference error (MSRTDE):

$$\mathrm{MSRTDE}^{\pi}(w_t) = \left\| V_{w_t} - \mathcal{T}_{\mathcal{P}}^{\pi} V_{w_{t-1}} \right\|^2 \tag{24}$$

where V_{w_t} denotes the value function estimate at iteration t, and $V_{w_{t-1}}$ denotes the value function estimate at iteration $t-1$.

Since samples can only be generated from the nominal model p°, the robust Bellman operator $\mathcal{T}_{\mathcal{P}}^{\pi}$ cannot be evaluated directly. Therefore, it must be approximated using sampled transitions. Based on (22) and (24), the mean-squared robust temporal-difference error is defined as

$$\mathrm{MSRTDE}^{\pi}(w_t) = \mathbb{E}_{\left(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}'_{1:k}\right)\sim O}\left[\left(V_{w_t}(\boldsymbol{s}) - \hat{\mathcal{T}}_{\mathcal{P}}^{\pi} V_{w_{t-1}}\left(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}'_{1:k}\right)\right)^2\right] \tag{25}$$

where O denotes the set of samples collected under the nominal transition model p°, and $\hat{\mathcal{T}}_{\mathcal{P}}^{\pi}$ denotes the empirical approximation of the robust Bellman operator.

In practice, this expectation is approximated using Monte Carlo estimation over the sample set O. Accordingly, the update rule for the robust Critic is given by

$$w_t = \arg\min_{w} \frac{1}{|O|} \sum_{\left(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}'_{1:k}\right)\in O} \left[V_{w_t}(\boldsymbol{s}) - \hat{\mathcal{T}}_{\mathcal{P}}^{\pi} V_{w_{t-1}}\left(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}'_{1:k}\right)\right]^2 \tag{26}$$

4.4. Robust Actor

To enable the robust Actor to generate effective guidance actions, it is updated by maximizing the RPPO clipped objective.

By employing the empirical robust Bellman operator [40], the robust advantage function is defined as

$$A_{t-1}(\boldsymbol{s}, \boldsymbol{a}) = \mathcal{T}_{\mathcal{P}}^{\pi} V_{w_{t-1}}\left(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}'_{1:k}\right) - V_{w_{t-1}}(\boldsymbol{s}) \tag{27}$$

Since samples can only be generated from the nominal model p°, the robust Bellman operator $\mathcal{T}_{\mathcal{P}}^{\pi}$ cannot be computed directly. Based on (24) and (27), the estimate of the robust advantage function is defined as

$$\hat{A}_{t-1}(\boldsymbol{s}, \boldsymbol{a}) = \hat{\mathcal{T}}_{\mathcal{P}}^{\pi} V_{w_{t-1}}\left(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}'_{1:k}\right) - V_{w_{t-1}}(\boldsymbol{s}) \tag{28}$$

Inspired by the classical Proximal Policy Optimization (PPO) algorithm [40] and based on (28), the RPPO clipped objective is defined as

$$L_{\mathrm{RPPO}}(\theta) = \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim D}\left[\min\left(r_{\theta_t}(\boldsymbol{s},\boldsymbol{a})\hat{A}_{t-1}(\boldsymbol{s},\boldsymbol{a}), \mathrm{clip}(r_{\theta_t}(\boldsymbol{s},\boldsymbol{a}), 1\pm\varepsilon)\hat{A}_{t-1}(\boldsymbol{s},\boldsymbol{a})\right)\right] \tag{29}$$

where $r_{\theta_t}(\boldsymbol{s},\boldsymbol{a}) = \frac{\pi_{\theta_t}(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\theta_{t-1}}(\boldsymbol{a}|\boldsymbol{s})}$ and ε is the clipping threshold.

Based on (29), the update rule for the robust Actor is given by

$$\theta_t = \arg\max_{\theta} \frac{1}{|D|} \sum_{(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}'_{1:k})\in D} \min\left(r_{\theta_t}(\boldsymbol{s},\boldsymbol{a})\hat{A}_{t-1}(\boldsymbol{s},\boldsymbol{a}), \mathrm{clip}(r_{\theta_t}(\boldsymbol{s},\boldsymbol{a}), 1\pm\varepsilon)\hat{A}_{t-1}(\boldsymbol{s},\boldsymbol{a})\right) \tag{30}$$

4.5. Algorithm Procedure

This section presents the complete algorithm, which is summarized in Algorithm 1.

Algorithm 1 Robust reinforcement learning-based guidance strategy

Input: training episodes K, total guidance steps M, maximum guidance steps N, training steps L, experience replay buffer D, mini-batch size B, discount factor γ, the learning rate α for the robust Critic, and the learning rate η for the robust Actor
Output: Actor parameters θ
Initialize: Actor parameters θ_0 and Critic parameters w_0
for $k = 0, 1, \ldots, K-1$ **do**
    Randomly initialize the motion states of the missile and the target.
    for $m = 0, 1, \ldots, M-1$ **do**
        Input the current state $\boldsymbol{s}$ into the current robust Actor to obtain the missile action $\boldsymbol{a}$.
        Execute the missile action $\boldsymbol{a}$ under the nominal model p° to obtain the next state $\boldsymbol{s}'$ and the reward r.
        Store the transition $(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}')$ in the experience replay buffer D.
        If $\dot{r}_{MT} > 0$ or m mod $N = 0$, randomly initialize the motion states of the missile and the target.
    end for
    for $l = 0, 1, \ldots, L-1$ **do**
        Randomly sample a mini-batch B from the experience replay buffer D.
        Under the double-sampling uncertainty set construction, perform double sampling for each transition in the mini-batch B using the nominal model, and obtain an unbiased estimate of the robust Bellman operator.
        Update the robust Critic by minimizing the mean-squared robust temporal-difference error according to (26) using the learning rate α.
        Compute the robust advantage estimate according to (28).
        Based on the stored log probabilities, the probability ratio between the new policy and the old policy for the same action is calculated.
        Update the robust Actor by maximizing the RPPO clipped objective according to (30) using the learning rate η.
    end for
end for

5. Experiments and Analysis

5.1. Experiment Setup

5.1.1. Environmental Parameters

To enhance the generalization capability of the guidance strategy across different scenarios, the initial environmental states are randomly sampled within predefined ranges. The corresponding environment parameter settings are listed in Table 1.

Table 1. The configurations of the environment parameters.

Parameter	Value
The coordinates of the missile (m)	0∼2000
The velocity of the missile (m/s)	350∼400
The flight-path angle of the missile (°)	50∼80
The coordinates of the target (m)	8000∼9000
The velocity of the target (m/s)	300∼350
The flight-path angle of the target (°)	190∼220
The maximum miss distance (m)	10
The simulation step when $r_{MT} > 50$ (s)	0.1
The simulation step when $r_{MT} \leq 50$ (s)	0.001

5.1.2. Hyperparameters

To ensure training stability while achieving satisfactory guidance performance, the hyperparameters were carefully configured. The detailed hyperparameter settings are listed in Table 2.

Table 2. The settings of the hyperparameters.

Parameter	Value	Parameter	Value
Training episodes K	7000	Robust Actor learning rate η	0.0003
Replay buffer size D	600,000	Optimizer	Adam
Total guidance steps M	600,000	Number of samples k	2
Maximum episode steps N	400	Perturbation radius δ	1/6
Training steps L	320	Clipping threshold ϵ	0.2
Mini-batch size B	2048	Reward coefficient k_r	0.3
Discount factor γ	0.99	Reward coefficient k_q	0.1
Robust Critic learning rate α	0.0003	Reward coefficient k_t	40

5.1.3. Network Parameters

To approximate the guidance strategy more accurately, the detailed network architecture is listed in Table 3.

Table 3. The configurations of the network parameters.

Parameter	Robust Actor		Robust Critic	
	Unit	Activation Function	Unit	Activation Function
Input Layer	5	None	5	None
Hidden Layer 1	64	ReLU	64	ReLU
Hidden Layer 2	64	ReLU	64	ReLU
Output Layer	1	Tanh	1	None

5.1.4. Performance Metrics

To evaluate the performance of the proposed guidance strategy, Monte Carlo tests with 2000 episodes were conducted for each scenario under different conditions, including

various cyber-attack types and guidance strategy configurations. The performance was quantitatively assessed using the following metrics:

Success Rate (SR): The proportion of episodes in which the missile successfully intercepts the maneuvering target. An episode is considered successful if the miss distance at the end of the engagement is smaller than the maximum allowable miss distance specified in Table 1; otherwise, the episode is regarded as a failure.

Average Miss Distance (AMD): The average miss distance over all successful interception episodes.

5.1.5. Comparison Guidance Strategies

To further evaluate the effectiveness of the proposed guidance strategy, its performance is compared with that of existing guidance strategies. The benchmark methods considered in this study are summarized as follows:

Reinforcement Learning-based Guidance Strategy (RL-GS) [41]: This guidance strategy employs the PPO algorithm to intercept the target while explicitly considering the effects of cyber-attacks.

Classical Theory-based Guidance Strategy (CT-GS) [17]: This guidance strategy achieves target interception under cyber-attack conditions by driving the state associated with the LOS angular rate to converge to zero within a finite time.

5.2. Training Results

During the training process, the missile is assumed to be subjected to cyber-attacks, while the target performs random maneuvers. The cyber-attack imposed on the missile is modeled as

$$a_A = 100\sin(0.5t) \tag{31}$$

The normal acceleration of the target is given by

$$a_T = n \cdot rand(-1, 1)g \tag{32}$$

where $g = 9.8\ \mathrm{m/s^2}$, $n = 6$. At the beginning of each training episode, the relative motion parameters of the missile and target are randomly initialized within the ranges specified in Table 1. Subsequently, the proposed robust reinforcement learning-based guidance strategy (RRL-GS) is trained for 7000 episodes. To evaluate its performance, 20 test episodes are conducted every 500 training episodes, and the average reward is calculated as the evaluation metric. In addition, for comparative analysis, RL-GS is trained and evaluated under the same environment and following the same evaluation procedure. It should be noted that CT-GS does not involve a training process. Therefore, it is not included in the comparison of average rewards.

Figure 4 illustrates the average reward curves of RRL-GS and RL-GS during the training process. In the early stage of training, the average reward of RRL-GS increases rapidly, indicating that it can effectively capture the underlying state–action mapping of the environment and achieve fast policy improvement with relatively limited interactions. Subsequently, the average reward exhibits a gradual upward trend over a prolonged training period and becomes stable after approximately 4000 episodes, demonstrating that the proposed method is capable of continuously refining its policy and ultimately achieving convergence. In contrast, the average reward of RL-GS increases at a noticeably slower rate. From the perspective of convergence characteristics, RRL-GS not only converges to a higher average reward but also reaches convergence more rapidly. These results indicate that RRL-GS achieves superior training efficiency and guidance performance. From a mechanistic perspective, RRL-GS explicitly accounts for the influence of disturbances during policy optimization through

the introduction of a robustness mechanism. In contrast, conventional RL-GS typically learns policies under a nominal environment and lacks explicit modeling of uncertainties. As a result, its guidance performance is more susceptible to degradation in the presence of disturbances. Overall, RRL-GS outperforms RL-GS in terms of both average reward and convergence speed. These results demonstrate that RRL-GS not only significantly improves training efficiency but also learns a more effective guidance strategy, thereby exhibiting greater potential for practical deployment under cyber-attack conditions.

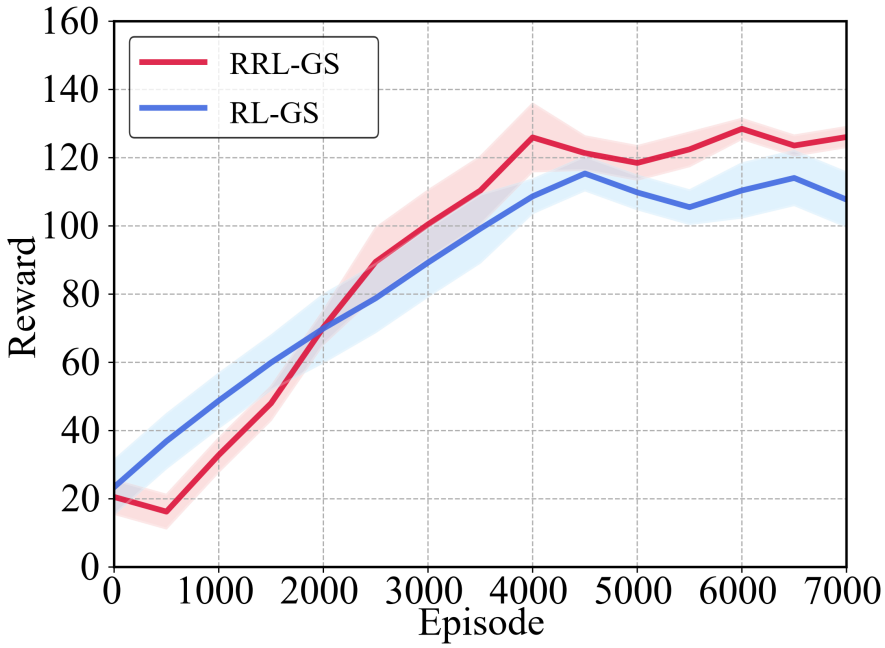


Figure 4. The average rewards of RRL-GS and RL-GS.

5.3. Testing Results

After being trained for 7000 episodes in the designated training environment, RRL-GS and RL-GS are evaluated together with CT-GS under four different test environments, namely Test Environment I, Test Environment II (i.e., the training environment), Test Environment III, and Test Environment IV. Specifically, in Test Environment I, the missile intercepts a randomly maneuvering target without cyber-attack interference, aiming to evaluate the fundamental effectiveness of the guidance strategy. In Test Environment II, the missile performs the same interception task under the cyber-attack described by (31), which is used to assess the interception performance of the guidance strategy under a previously learned attack pattern. In Test Environment III, the missile carries out the interception task under a different cyber-attack pattern, thereby further evaluating the robustness of RRL-GS against unseen attack scenarios. Through these three test environments, the effectiveness, adaptability, and robustness of the proposed guidance strategy can be comprehensively assessed under different cyber-attack conditions. In Test Environment IV, the missile intercepts the target under a different cyber-attack pattern.

5.3.1. Test Environment I

In Test Environment I, the target performs random maneuvers while the missile is free from cyber-attack interference. To comprehensively evaluate the generalization capability of the guidance strategy, the motion parameters are randomly initialized within the ranges specified in Table 1 at the beginning of each test episode. To assess the effectiveness of the guidance strategy in an attack-free environment, two performance metrics are adopted: SR and AMD. Table 4 presents the comparative results of RRL-GS, RL-GS, and CT-GS

obtained from 10 independent Monte Carlo experiments, with each experiment consisting of 2000 test episodes.

As shown in Table 4, RRL-GS achieves the highest success rate of 95.0% (95%CI: 94.1–95.9%), compared with 90.3% (95%CI: 89.3–91.3%) for RL-GS and 87.7% (95%CI: 86.5–88.9%) for CT-GS. This demonstrates the superior guidance performance of RRL-GS. Moreover, RRL-GS achieves an AMD of 1.26 m (95%CI: 1.05–1.47 m), which is lower than those of RL-GS, at 1.81 m (95%CI: 1.56–2.06 m), and CT-GS, at 2.29 m (95%CI: 2.15–2.43 m). This indicates that RRL-GS not only successfully intercepts the target but also achieves higher interception accuracy. Further analysis reveals that the higher success rate reflects stronger adaptability of the learned policy to varying initial conditions, while the lower average miss distance indicates more precise guidance capability. Together, these two metrics demonstrate the superiority of RRL-GS, showing that it can not only achieve reliable interceptions but also maintain high interception accuracy. Overall, RRL-GS outperforms the benchmark guidance strategies in terms of both success rate and average miss distance, indicating clear advantages in generalization capability and guidance accuracy. These results further validate the effectiveness of RRL-GS and highlight its potential for practical engineering applications.

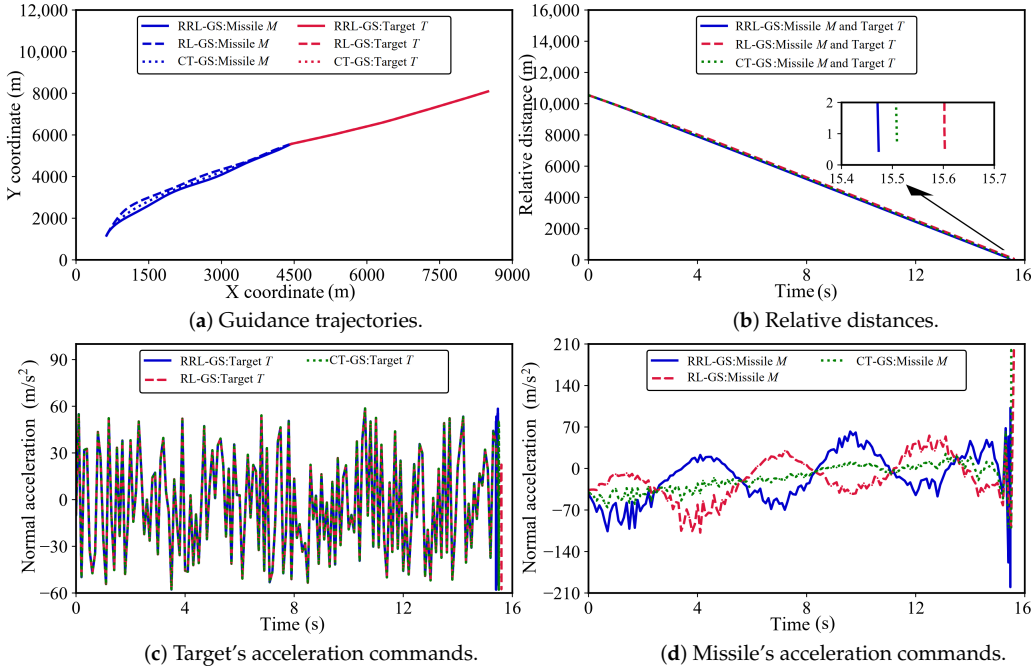
Table 4. Monte Carlo test results in Test Environment I.

Guidance Strategy	SR (%)	AMD (m)
RRL-GS	95.0 (95%CI: 94.1–95.9)	1.26 (95%CI: 1.05–1.47)
RL-GS	90.3 (95%CI: 89.3–91.3)	1.81 (95%CI: 1.56–2.06)
CT-GS	87.7 (95%CI: 86.5–88.9)	2.29 (95%CI: 2.15–2.43)

The simulation results of different guidance strategies in Test Environment I are shown in Figure 5. All three guidance strategies successfully guide the missile to intercept the target. As illustrated in Figure 5a, the missile trajectory generated by RRL-GS is smoother and does not exhibit significant abrupt maneuvering adjustments. This indicates superior guidance performance and a more desirable control behavior. As shown in Figure 5b, the relative distance decreases monotonically over time for all three guidance strategies. When RRL-GS is employed, the final miss distance is 0.45 m. In comparison, the miss distances achieved by RL-GS and CT-GS are 0.47 m and 0.71 m, respectively. Although the numerical difference between RRL-GS and RL-GS is relatively small, RRL-GS still achieves the best performance, while demonstrating a more substantial improvement over CT-GS. These results indicate that RRL-GS is capable of not only reliably accomplishing the interception task but also achieving higher interception accuracy. In summary, RRL-GS outperforms both RL-GS and CT-GS in terms of trajectory smoothness and guidance accuracy. These findings demonstrate that RRL-GS can achieve high-precision target interception while exhibiting superior dynamic response characteristics, thereby offering greater potential for practical engineering applications.

5.3.2. Test Environment II

In Test Environment II (i.e., the training environment), the target performs random maneuvers. The only difference from Test Environment I is that the missile is subjected to cyber-attack interference. All guidance strategies are evaluated through 10 independent Monte Carlo experiments, with each experiment consisting of 2000 test episodes, and the corresponding results are presented in Table 5.



(**a**) Guidance trajectories. (**b**) Relative distances.

(**c**) Target's acceleration commands. (**d**) Missile's acceleration commands.

Figure 5. Comparison case of RRL-GS, RL-GS, and CT-GS in Test Environment I.

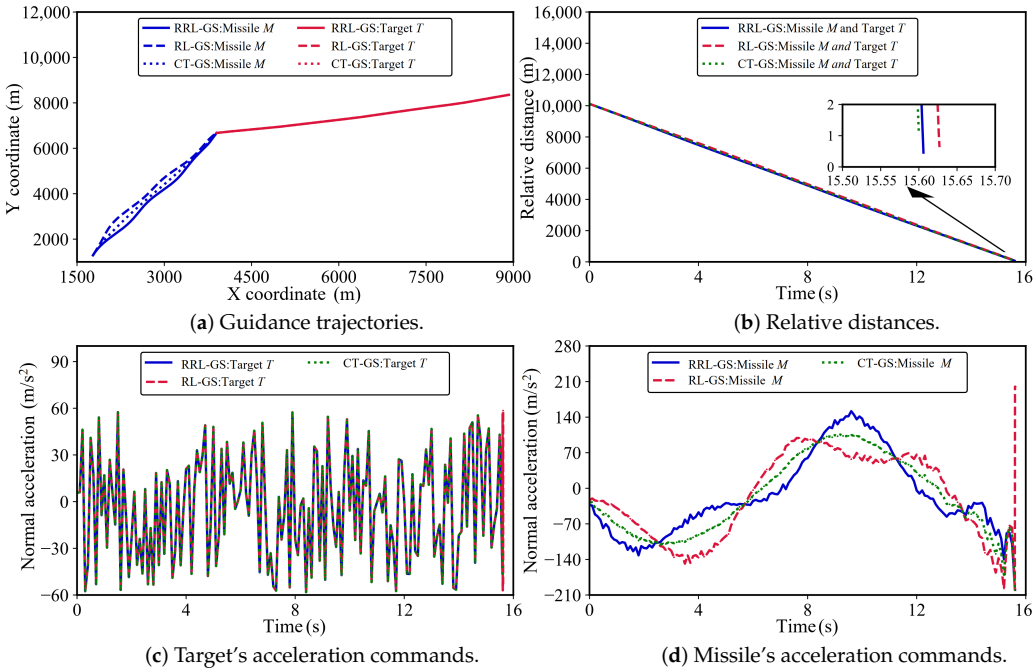
Table 5. Monte Carlo test results in Test Environment II.

Guidance Strategy	SR (%)	AMD (m)
RRL-GS	91.1 (95%CI: 90.4–91.8)	1.62 (95%CI: 1.45–1.79)
RL-GS	86.8 (95%CI: 85.4–88.2)	1.99 (95%CI: 1.81–2.17)
CT-GS	83.5 (95%CI: 82.6–84.4)	2.62 (95%CI: 2.38–2.86)

As shown in Table 5, the success rates of all three guidance strategies decrease to varying degrees under cyber-attack interference compared with the attack-free scenario. This indicates that cyber-attacks can directly affect the effectiveness of guidance commands. Nevertheless, RRL-GS achieves the highest success rate of 91.1% (95%CI: 90.4–91.8%), compared with 86.8% (95%CI: 85.4–88.2%) for RL-GS and 83.5% (95%CI: 82.6–84.4%) for CT-GS. This demonstrates that RRL-GS is capable of maintaining superior guidance performance even in the presence of cyber-attacks. From the perspective of guidance accuracy, RRL-GS achieves an AMD of 1.62 m (95%CI: 1.45–1.79 m), which is lower than those of RL-GS, at 1.99 m (95%CI: 1.81–2.17 m), and CT-GS, at 2.62 m (95%CI: 2.38–2.86 m). This result indicates that RRL-GS can still maintain a high level of interception accuracy under cyber-attack conditions. Furthermore, RRL-GS consistently outperforms the competing methods in both success rate and average miss distance. This suggests that its performance improvement is not limited to a single metric but instead reflects a balanced enhancement of both reliability and accuracy. From a methodological perspective, RRL-GS explicitly accounts for environmental uncertainties during policy learning through the incorporation of a robust optimization mechanism. As a result, it is able to preserve favorable guidance performance in the presence of disturbances and attacks. Overall, RRL-GS maintains clear advantages in both success rate and average miss distance under cyber-attack conditions, thereby providing strong evidence of its robustness and superior performance in adversarial environments.

The simulation results in Test Environment II are shown in Figure 6. As illustrated in Figure 6a, all three guidance strategies are still capable of guiding the missile toward the target under cyber-attack interference, indicating that each method possesses a certain degree of resilience against disturbances. As shown in Figure 6b, the relative distance

decreases gradually over time for all three guidance strategies, demonstrating that they retain target interception capability even in the presence of attacks. In terms of interception accuracy, RRL-GS achieves a miss distance of 0.46 m, which is significantly lower than those of RL-GS (0.63 m) and CT-GS (1.03 m). This result indicates that RRL-GS is able to maintain a high level of guidance accuracy despite the uncertainties introduced by cyber-attacks. Overall, under cyber-attack conditions, RRL-GS not only effectively mitigates the adverse effects of external disturbances but also maintains superior interception performance in complex and uncertain environments.



(**a**) Guidance trajectories. (**b**) Relative distances.

(**c**) Target's acceleration commands. (**d**) Missile's acceleration commands.

Figure 6. Comparison case of RRL-GS, RL-GS, and CT-GS in Test Environment II.

5.3.3. Test Environment III

In Test Environment III, the target performs random maneuvers while the missile is subjected to a different cyber-attack pattern:

$$a_A = \begin{cases} 0 & 0 \le t < 10 \\ 20\tanh(0.5t - 10) & t \ge 10 \end{cases} \tag{33}$$

The motion parameters of the missile and target are randomly initialized at the beginning of each simulation episode within the ranges specified in Table 1. Table 6 presents the test results of RRL-GS, RL-GS, and CT-GS obtained from 10 independent Monte Carlo experiments, with each experiment consisting of 2000 test episodes.

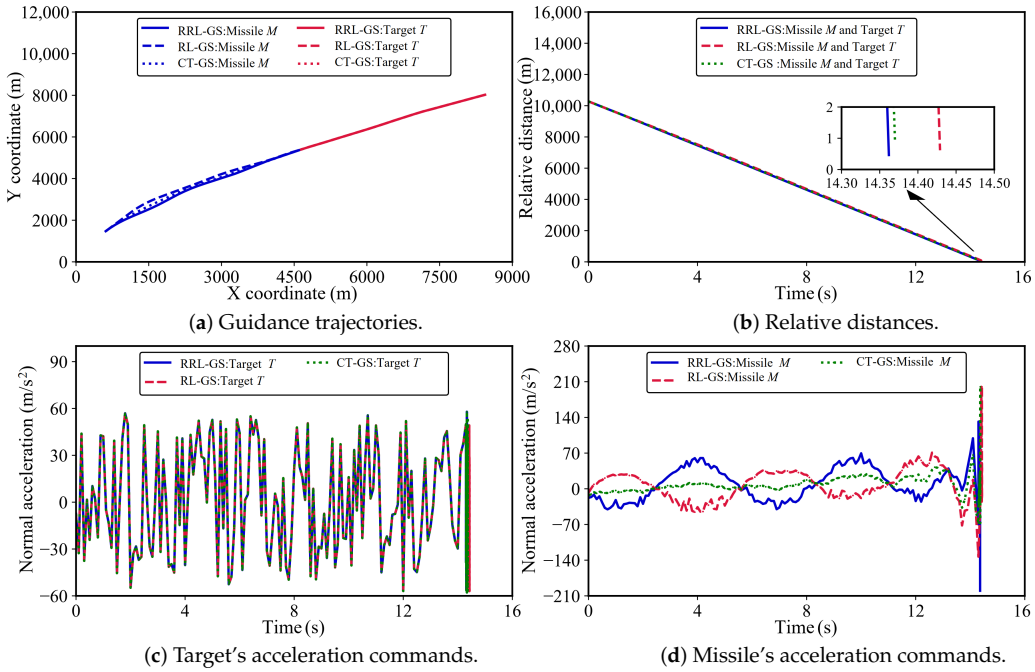
Table 6. Monte Carlo test results in Test Environment III.

Guidance Strategy	SR (%)	AMD (m)
RRL-GS	86.4 (95%CI: 85.2–87.6)	1.71 (95%CI: 1.52–1.90)
RL-GS	83.2 (95%CI: 81.7–84.7)	2.01 (95%CI: 1.89–2.13)
CT-GS	79.9 (95%CI: 78.9–80.9)	2.74 (95%CI: 2.55–2.93)

As shown in Table 6, the success rates of all three guidance strategies decrease under the unseen cyber-attack pattern compared with those obtained under the previously learned attack pattern. Nevertheless, RRL-GS achieves the highest success rate of 86.4% (95%CI: 85.2–87.6%), compared with 83.2% (95%CI: 81.7–84.7%) for RL-GS and 79.9% (95%CI:

78.9–80.9%) for CT-GS. This demonstrates that RRL-GS possesses stronger generalization capability and superior guidance performance. RRL-GS also achieves the lowest AMD of 1.71 m (95%CI: 1.52–1.90 m), whereas the AMDs of RL-GS and CT-GS are 2.01 m (95%CI: 1.89–2.13 m) and 2.74 m (95%CI: 2.55–2.93 m), respectively. These results indicate that RRL-GS maintains higher guidance accuracy than the competing methods. Overall, in the more challenging scenario represented by Test Environment III, RRL-GS continues to outperform the benchmark guidance strategies in both success rate and guidance performance, demonstrating stronger robustness and generalization capability. These results further validate the effectiveness of RRL-GS in complex and uncertain environments.

The simulation results in Test Environment III are shown in Figure 7. As illustrated in Figure 7a, all three guidance strategies are capable of guiding the missile toward the target under the unseen cyber-attack pattern, indicating their basic feasibility in this challenging scenario. As shown in Figure 7b, the relative distance decreases gradually over time for all three guidance strategies; however, noticeable differences can be observed in their convergence behaviors. RRL-GS reduces the relative distance more rapidly than the other methods, resulting in a shorter overall interception time. This suggests that RRL-GS is able to maintain high decision-making efficiency when confronted with previously unseen attack patterns, thereby reducing the time loss caused by ineffective maneuvers. In terms of interception accuracy, RRL-GS achieves a miss distance of 0.48 m, which is lower than those of RL-GS (0.61 m) and CT-GS (0.93 m). This further demonstrates the ability of RRL-GS to maintain a high level of guidance accuracy under unseen attack conditions. Overall, under the unseen cyber-attack pattern, RRL-GS outperforms both RL-GS and CT-GS in terms of interception efficiency and guidance accuracy. These results further confirm its robustness and superior performance in uncertain environments.



(**a**) Guidance trajectories. (**b**) Relative distances.

(**c**) Target's acceleration commands. (**d**) Missile's acceleration commands.

Figure 7. Comparison case of RRL-GS, RL-GS, and CT-GS in Test Environment III.

5.3.4. Test Environment IV

In Test Environment IV, the target performs random maneuvers, while the missile is subjected to another cyber-attack pattern:

$$a_A = \begin{cases} 0 & 0 \le t < 5 \\ 30rand(-1,1) & t \ge 5 \end{cases} \tag{34}$$

At the beginning of each simulation episode, the motion parameters of the missile and target are randomly initialized within the ranges specified in Table 1. Table 7 presents the test results of RRL-GS, RL-GS, and CT-GS, obtained from 10 independent Monte Carlo experiments, with each experiment consisting of 2000 test episodes.

As shown in Table 7, the success rates of all three guidance strategies decrease under the unseen cyber-attack pattern compared with those achieved under the previously encountered attack pattern. Nevertheless, RRL-GS still achieves the highest success rate of 84.6% (95%CI: 83.8–85.4%), outperforming RL-GS (81.5% (95%CI: 80.1–82.9%)) and CT-GS (78.1% (95%CI: 76.2–80.0%)). This result demonstrates that RRL-GS possesses superior generalization capability and maintains more effective guidance performance under unseen cyber-attack conditions. Furthermore, RRL-GS achieves the lowest AMD of 1.83 m (95%CI: 1.66–2.00 m), compared with 2.06 m (95%CI: 1.81–2.31 m) for RL-GS and 2.61 m (95%CI: 2.41–2.81 m) for CT-GS. These results indicate that RRL-GS can consistently achieve a higher interception success rate and a smaller average miss distance when facing diverse cyber-attack patterns, thereby further verifying the enhanced guidance accuracy, adaptability, and robustness of the proposed strategy in complex cyber-attack environments.

Table 7. Monte Carlo test results in Test Environment IV.

Guidance Strategy	SR (%)	AMD (m)
RRL-GS	84.6 (95%CI: 83.8–85.4)	1.83 (95%CI: 1.66–2.00)
RL-GS	81.5 (95%CI: 80.1–82.9)	2.06 (95%CI: 1.81–2.31)
CT-GS	78.1 (95%CI: 76.2–80.0)	2.61 (95%CI: 2.41–2.81)

5.3.5. Different Guidance Strategies

In the training environment, the guidance strategies based on Wasserstein robust Q-learning (WRQL-GS) [27], Twin Delayed Deep Deterministic Policy Gradient (TD3-GS) [42], and Deep Deterministic Policy Gradient (DDPG-GS) [43] were trained for 7000 episodes. After training, all terminal guidance strategies were evaluated through 2000 Monte Carlo simulations. The corresponding results are presented in Table 8.

Table 8. Comparative analysis of different guidance strategies.

Guidance Strategy	SR (%)	AMD (m)
RRL-GS	90.9	1.59
WRQL-GS	88.1	1.85
TD3-GS	85.7	2.03
DDPG-GS	84.2	2.31

As shown in Table 8, RRL-GS achieves an interception success rate of 90.9%, which is higher than those of WRQL-GS (88.1%), TD3-GS (85.7%), and DDPG-GS (84.2%). In terms of AMD, RRL-GS obtains the best performance with an AMD of 1.59 m, which is lower than those of WRQL-GS (1.85 m), TD3-GS (2.03 m), and DDPG-GS (2.31 m). These results demonstrate that, compared with the other benchmark strategies, RRL-GS can achieve a higher interception success rate and a lower AMD under complex cyber-attack environments, exhibiting superior guidance accuracy and robustness. This further verifies the effectiveness of the proposed method in improving the performance of guidance systems under adversarial cyber-attack conditions.

5.3.6. Sensitivity Analysis

The clipping threshold is an important hyperparameter that controls the update magnitude of the guidance strategy. Its primary purpose is to prevent training instability caused

by excessively large policy updates while avoiding reduced training efficiency resulting from overly conservative updates. To evaluate its impact on guidance performance, guidance strategies with different clipping thresholds were trained using five independent random seeds for 7000 episodes in the training environment. After training, each guidance strategy was evaluated in Test Environment II (i.e., the training environment) through 2000 Monte Carlo simulations. The results of the clipping-threshold sensitivity analysis are summarized in Figure 8 and Table 9. In addition, the key parameters of the double-sampling method were subjected to sensitivity analysis using the same procedure, and the corresponding results are summarized in Tables 10 and 11.

Figure 8 presents the average reward curves obtained under different clipping thresholds, where each curve is averaged over five independent training runs with different random seeds. It can be observed that, when ϵ = 0.2, the average reward of RRL-GS increases rapidly during the early stage of training. As the number of training episodes increases, the average reward continues to improve steadily and gradually converges after approximately 4000 episodes. This indicates that the selected clipping threshold effectively facilitates policy optimization while maintaining the stability of policy updates, enabling the agent to ultimately learn a high-performance guidance strategy. In contrast, when ϵ = 0.1 or ϵ = 0.3, the growth rate of the average reward is significantly reduced, and the final converged reward is lower than that achieved with ϵ = 0.2. These results demonstrate that ϵ = 0.2 provides a favorable balance between training stability and policy improvement efficiency, allowing RRL-GS to achieve superior guidance performance and stronger application potential in cyber-attack environments.

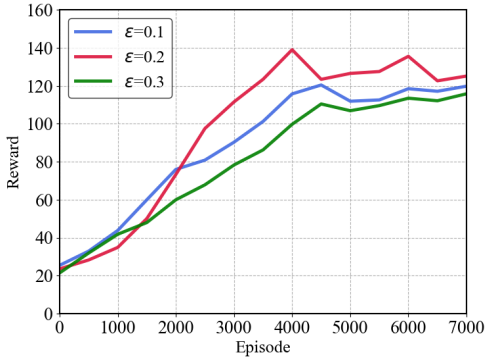


Figure 8. The average rewards of different clipping thresholds.

Table 9. Sensitivity analysis of the clipping threshold.

Clipping Threshold ϵ	SR (%)	AMD (m)
0.1	73.4 ± 2.8	3.21 ± 0.43
0.2	91.3 ± 2.2	1.68 ± 0.29
0.3	66.8 ± 3.0	3.81 ± 0.38

As shown in Table 9, the clipping threshold ϵ has a significant impact on guidance performance. When ϵ = 0.2, the proposed guidance strategy achieves the highest success rate of 91.3% ± 2.2% and the lowest average miss distance of 1.68 m ± 0.29 m. When ϵ = 0.1, the policy updates become overly conservative, resulting in degraded performance, with the success rate decreasing to 73.4% ± 2.8% and the average miss distance increasing to 3.21 m ± 0.43 m. In contrast, when ϵ = 0.3, the success rate drops to 66.8% ± 3.0%, while the average miss distance increases to 3.81 m ± 0.38 m. From a mechanistic perspective, the clipping threshold ϵ directly determines the allowable range of policy updates. When ϵ is relatively small (e.g., 0.1), policy updates are strictly constrained, leading to only minor improvements in each iteration. Although such a setting can enhance training

stability, it substantially slows down convergence and prevents the policy from approaching the optimum efficiently. This may lead to lower success rates and reduced guidance accuracy. Conversely, when ϵ is relatively large (e.g., 0.3), the constraints on policy updates become weaker. While this may accelerate short-term performance improvement, it can also introduce excessive policy fluctuations and potentially disrupt previously learned effective policy structures, leading to training instability and degraded overall performance. Further analysis indicates that $\epsilon = 0.2$ provides a favorable trade-off between exploration and stability. It allows sufficiently large policy updates to facilitate effective learning while simultaneously preventing instability caused by excessive updates. This balanced update mechanism enables the policy to continuously improve while maintaining stable convergence, resulting in superior performance in terms of both success rate and guidance accuracy. Overall, the results demonstrate that an appropriate choice of the clipping threshold can effectively improve the convergence behavior of the guidance strategy, thereby leading to enhanced guidance performance and a higher interception success rate.

As shown in Table 10, when $k = 2$, the proposed guidance strategy achieves the highest interception success rate (91.3% $\pm$ 2.2%) and the lowest average miss distance (1.68 m $\pm$ 0.29 m). When $k = 1$, the interception success rate decreases to 76.1% $\pm$ 2.6%, while the average miss distance increases to 2.97 m $\pm$ 0.40 m. In contrast, when $k = 3$, the interception success rate drops to 68.5% $\pm$ 2.9%, and the average miss distance increases to 3.65 m $\pm$ 0.34 m. These results indicate that the choice of k has a substantial impact on guidance performance. When $k = 2$, the double-sampling size is moderate, enabling the empirical robust Bellman operator to achieve a favorable balance between estimation accuracy and computational complexity. This balance improves training efficiency and leads to superior guidance performance.

Table 10. Sensitivity analysis of k value.

k	SR (%)	AMD (m)
1	76.1 $\pm$ 2.6	2.97 $\pm$ 0.40
2	91.3 $\pm$ 2.2	1.68 $\pm$ 0.29
3	68.5 $\pm$ 2.9	3.65 $\pm$ 0.34

Table 11. Sensitivity analysis of δ value.

δ	SR (%)	AMD (m)
1/9	70.6 $\pm$ 3.4	3.46 $\pm$ 0.46
1/6	91.3 $\pm$ 2.2	1.68 $\pm$ 0.29
1/3	62.7 $\pm$ 3.6	3.93 $\pm$ 0.41

As shown in Table 11, when $\delta = 1/6$, the proposed guidance strategy achieves the highest interception success rate (91.3% $\pm$ 2.2%) and the lowest average miss distance (1.68 m $\pm$ 0.29 m). When $\delta = 1/9$, the interception success rate decreases to 70.6% $\pm$ 3.4%, while the average miss distance increases to 3.46 m $\pm$ 0.46 m. In contrast, when $\delta = 1/3$, the interception success rate drops to 62.7% $\pm$ 3.6%, and the average miss distance increases to 3.93 m $\pm$ 0.41 m. These results indicate that the value of δ requires a trade-off among policy robustness, value-estimation accuracy, and training efficiency. When $\delta = 1/9$, the uncertainty set is relatively narrow and may fail to adequately characterize potential uncertainties, resulting in insufficient policy robustness. In contrast, when $\delta = 1/3$, the uncertainty set is excessively large, causing the value estimates to become overly conservative, potentially increasing estimation bias and reducing policy-update efficiency. Therefore, selecting an appropriate value of δ is essential for achieving a balance among policy robustness, estimation accuracy, and training efficiency.

6. Conclusions

This paper proposes a robust reinforcement learning-based guidance strategy for missile guidance under cyber-attacks. First, the missile guidance problem in the presence of cyber-attacks is formulated as a robust Markov decision process. Subsequently, a robust reinforcement learning-based guidance strategy with a robust Actor–robust Critic architecture is developed to improve guidance performance and interception success rate under cyber-attack disturbances. Finally, extensive simulations under various cyber-attack models are conducted to validate the effectiveness and feasibility of the proposed method. The simulation results demonstrate that the proposed guidance strategy can effectively guide missiles to intercept maneuvering targets under diverse attack scenarios. Compared with the deep reinforcement learning-based guidance strategy and the conventional theory-based guidance strategy, the proposed method exhibits superior robustness against cyber-attacks and achieves better terminal miss-distance performance.

In future work, we will investigate guidance strategy design under more complex cyber-attack scenarios, including delay attacks, replay attacks, packet loss, and random disturbances, to enhance the robustness and security of guidance systems. In addition, practical factors associated with missile systems, such as autopilot dynamics, overload constraints, actuator response delays, actuator faults, and aerodynamic parameter variations, will be incorporated to establish a high-fidelity guidance simulation environment that better reflects engineering requirements.

Author Contributions: Y.S.: investigation, conceptualization, methodology, validation, writing—original draft, writing—review and editing, software. Y.C.: conceptualization, project administration, supervision, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Innovation Special Field of China (grant no. 23-TQ01-04-ZT-01-011) and the APC was funded by the Innovation Special Field of China.

Data Availability Statement: No data was used for the research described in the article.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

In order to estimate the cyber-attacks and the target maneuvering information, the observer is defined as follows:

$$\dot{\hat{x}} = F(x) + Ba_G + \kappa_1 \frac{\tilde{x}}{\|\tilde{x}\|^{1/2}} + \hat{w}, \tag{A1}$$

$$\dot{\hat{w}} = \kappa_2 \frac{\tilde{x}}{\|\tilde{x}\|}. \tag{A2}$$

where $x = \left[r_{MT}\dot{q}_{MT} - c_1 r_{MT}\right]$, $F(x) = \left[-\dot{r}_{MT}\dot{q}_{MT} - \frac{c_1 \dot{r}_{MT}}{2\sqrt{r_{MT}}}\right]$, $B = \left[-\cos\theta_M\right]$, and $\tilde{x} = x - \hat{x}$. Here, $c_1 > 0$ is the designed constants, and $\kappa_1 > 0$ and $\kappa_2 > 2\kappa_1$ are the observer gains.

References

1. Yang, X.; Wu, Y.; Xia, W.; Liu, X.; Duan, H. Nonsingular field-of-view constrained cooperative guidance with improved time-to-go estimation. *Chin. J. Aeronaut.* **2025**, *38*, 103565. [CrossRef]
2. Li, G.; Tang, Q.; Zuo, Z.; Wu, Y.; Lü, J. Resilient cooperative guidance for leader-follower flight vehicles against maneuvering target. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 6310–6324. [CrossRef]

3. Wang, J.; Tao, X.; Dong, W.; Wang, C. Three-dimensional predefined-time impact angle control guidance law with field-of-view limit. *J. Frankl. Inst.* **2023**, *360*, 7621–7644. [CrossRef]
4. Yang, Z.; Liu, X.; Liu, H. Impact time control guidance law with time-varying velocity based on deep reinforcement learning. *Aerosp. Sci. Technol.* **2023**, *142*, 108603. [CrossRef]
5. Wang, W.; Zhang, H.; Huang, P.; Wang, Y. Adaptive fixed-time observer-based impact angle control guidance law for intercepting maneuvering target. *Aerosp. Sci. Technol.* **2025**, *158*, 109871. [CrossRef]
6. Wang, X.; Huang, X.; Ding, S. Terminal angle constraint finite-time guidance law with input saturation and autopilot dynamics. *J. Frankl. Inst.* **2022**, *359*, 8687–8712. [CrossRef]
7. Shin, H.S.; Li, K.B. An improvement in three-dimensional pure proportional navigation guidance. *IEEE Trans. Aerosp. Electron. Syst.* **2021**, *57*, 3004–3014. [CrossRef]
8. Wang, Y.; Wang, H.; Lin, D.; Wang, W. Nonlinear modified bias proportional navigation guidance law against maneuvering targets. *J. Frankl. Inst.* **2022**, *359*, 2949–2975. [CrossRef]
9. Chen, Y.; Wu, S.; Wang, X.; Zhang, D.; Jia, J.; Li, Q. Time and FOV constraint guidance applicable to maneuvering target via sliding mode control. *Aerosp. Sci. Technol.* **2023**, *133*, 108104. . [CrossRef]
10. Zheng, Z.; Li, J.; Feroskhan, M. Three-dimensional terminal angle constraint guidance law with class K∞ function-based adaptive sliding mode control. *Aerosp. Sci. Technol.* **2024**, *147*, 109005. [CrossRef]
11. Liu, S.; Yan, B.; Zhang, T.; Zhang, X.; Yan, J. Three-dimensional coverage-based cooperative guidance law with overload constraints to intercept a hypersonic vehicle. *Aerosp. Sci. Technol.* **2022**, *130*, 107908. [CrossRef]
12. Turetsky, V.; Weiss, M.; Shima, T. A combined linear–quadratic/bounded control differential game guidance law. *IEEE Trans. Aerosp. Electron. Syst.* **2021**, *57*, 3452–3462. [CrossRef]
13. Chen, W.; Gao, C.; Gao, W. Proximal policy optimization guidance algorithm for intercepting near-space maneuvering targets. *Aerosp. Sci. Technol.* **2023**, *132*, 108031. [CrossRef]
14. Gaudet, B.; Furfaro, R.; Linares, R. Reinforcement learning for angle-only intercept guidance of maneuvering targets. *Aerosp. Sci. Technol.* **2020**, *99*, 105746. [CrossRef]
15. Siouris, G. *Missile Guidance and Control Systems*; Springer Science & Business Media: New York, NY, USA, 2004. [CrossRef]
16. Wang, Y.; Wang, Z.; Wang, C.; Bai, L.; Qiao, J.; Guo, L. Multivariable secure guidance for interceptors against cyber-attacks in guidance commands. *Aerosp. Sci. Technol.* **2023**, *139*, 108378. [CrossRef]
17. Li, G.; Zuo, Z. Robust leader–follower cooperative guidance under false-data injection attacks. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4511–4524. [CrossRef]
18. Wang, Y.; Wang, C.; Zhu, B.; Liu, G.; Wang, E.; Guo, L. Secure three-dimensional guidance for target strikes under unknown direction attacks and multiple constraints. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 6951–6963. [CrossRef]
19. Celentano, L.; Basin, M.V.; Shi, P. Majorant-based control methodology for mechatronic and transportation processes. *IEEE Access* **2021**, *9*, 157916–157939. [CrossRef]
20. Celentano, L.; Basin, M.V.; Chadli, M. Robust tracking design for uncertain MIMO systems using proportional–integral controller of order v. *Asian J. Control* **2021**, *23*, 2042–2063. [CrossRef]
21. Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; Meger, D. Deep reinforcement learning that matters. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 3207–3214. [CrossRef]
22. Wang, Y.; Zou, S. Online robust reinforcement learning with model uncertainty. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 7193–7206.
23. Panaganti, K.; Xu, Z.; Kalathil, D.; Ghavamzadeh, M. Robust reinforcement learning using offline data. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 32211–32224. [CrossRef]
24. Choi, J.; Li, M.; Li, M.Z. Adaptive robust Markov decision process for wide-area surveillance with collaborative combat aircraft. In Proceedings of the AIAA SCITECH Forum, Orlando, FL, USA, 12–16 January 2026; AIAA 2026-2884. [CrossRef]
25. Deshpande, A.M.; Minai, A.A.; Kumar, M. Robust deep reinforcement learning for quadcopter control. *IFAC-PapersOnLine* **2021**, *54*, 90–95. [CrossRef]
26. Xue, J.; Liu, Z.; Liu, G.; Zhou, Z.; Zhang, K.; Tang, Y.; Wang, J. Robust wind-resistant hovering control of quadrotor UAVs using deep reinforcement learning. *IEEE Trans. Intell. Veh.* **2023**, 1–10. [CrossRef]
27. Neufeld, A.; Sester, J. Robust Q-learning algorithm for Markov decision processes under Wasserstein uncertainty. *Automatica* **2024**, *168*, 111825. [CrossRef]
28. Wang, Y.; Zou, S. Policy gradient method for robust reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2022; Volume 162, pp. 23484–23526.
29. Xu, Z.; Panaganti, K.; Kalathil, D. Improved sample complexity bounds for distributionally robust reinforcement learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*; PMLR: Cambridge, MA, USA, 2023; Volume 206, pp. 9728–9754.
30. Kumar, N.; Derman, E.; Geist, M.; Levy, K.Y.; Mannor, S. Policy gradient for rectangular robust Markov decision processes. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 59477–59501. [CrossRef]

31. Wang, N.; Wang, X.; Cui, N.; Li, Y.; Liu, B. Deep reinforcement learning-based impact time control guidance law with constraints on the field-of-view. *Aerosp. Sci. Technol.* **2022**, *128*, 107765. [CrossRef]
32. Shalumov, V. Cooperative online guide–launch–guide policy in a target–missile–defender engagement using deep reinforcement learning. *Aerosp. Sci. Technol.* **2020**, *104*, 105996. [CrossRef]
33. Garcia, F.; Rachelson, E. Markov decision processes. In *Markov Decision Processes in Artificial Intelligence*; John Wiley & Sons: Hoboken, NJ, USA, 2013; pp. 1–38. [CrossRef]
34. Iyengar, G.N. Robust dynamic programming. *Math. Oper. Res.* **2005**, *30*, 257–280. [CrossRef]
35. Xu, X.; Cai, P.; Ahmed, Z.; Yellapu, V.S.; Zhang, W. Path planning and dynamic collision avoidance algorithm under COLREGs via deep reinforcement learning. *Neurocomputing* **2022**, *468*, 181–197. [CrossRef]
36. Nilim, A.; El Ghaoui, L. Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.* **2005**, *53*, 780–798. [CrossRef]
37. Panaganti, K.; Kalathil, D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. *Proc. Int. Conf. Mach. Learn.* **2021**, *139*, 511–520.
38. Zhou, R.; Liu, T.; Cheng, M.; Kalathil, D.; Kumar, P.R.; Tian, C. Natural actor-critic for robust reinforcement learning with function approximation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 97–133. [CrossRef]
39. Sutton, R.S.; Maei, H.R.; Szepesvári, C. A convergent O(n) temporal-difference algorithm for off-policy learning with linear function approximation. *Adv. Neural Inf. Process. Syst.* **2008**, *21*, 1609–1616.
40. Kadurha, D.K.; Moutouo, D.J.L.; Gaba, Y.U. Bellman operator convergence enhancements in reinforcement learning algorithms. *arXiv* **2025**, arXiv:2505.14564. [CrossRef]
41. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347. [CrossRef]
42. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2018; Volume 80, pp. 1587–1596.
43. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.