Intelligent Penetration Guidance for Hypersonic Missiles via Reinforcement Learning and Optimal Control

Tianya Liu , Fengshuo Wang , Peng Li *

Posted Date: 20 August 2025

doi: 10.20944/preprints202508.0144.v2

Keywords: hypersonic missile penetration; optimal guidance law; deep reinforcement learning; imitation learning

 doi:10.20944/preprints202508.0144.v2

Article

Intelligent Penetration Guidance for Hypersonic Missiles via Reinforcement Learning and Optimal Control

Tianya Liu, Fengshuo Wang and Peng Li *

College of Intelligence Science, National University of Defense Technology, Changsha 410073, China
* Correspondence: lipeng_2010@163.com

Abstract

To enhance the penetration capability and strike accuracy of missiles in scenarios involving multiple interceptor missiles, this paper proposes a penetration guidance strategy based on deep reinforcement learning (DRL). First, the BANG-BANG optimal penetration strategy is derived as an expert policy by optimizing the maximum miss distance between the attacking missile and the interceptor missiles as a performance metric. Subsequently, a Markov Decision Process (MDP) model for missile penetration guidance is established, and we design a multi-objective reward function to integrate penetration success rate, miss distance, and energy consumption. Furthermore, a penetration strategy learning network based on Generative Adversarial Imitation Learning (GAIL) and Proximal Policy Optimization (PPO) is constructed and trained. Simulation results demonstrate that the proposed strategy exhibits high training efficiency and enables superior decision-making in complex adversarial scenarios.

Keywords: hypersonic missile penetration; optimal guidance law; deep reinforcement learning; imitation learning

1. Introduction

The precise strike capabilities of hypersonic missiles significantly affect modern warfare. In practice, the flight trajectory of the hypersonic missile can be divided into three phases: the boost phase, midcourse phase, and terminal phase. The terminal phase is crucial, as it directly determines the success of the mission and influences the missile's final flight path and strike accuracy. Recently, global military powers have been researching interception technologies for hypersonic missiles and missile defense systems, intensively [1–4]. The terminal phase dive, characterized by low altitude limited acceleration, and high detectability, has become a critical interception zone for defense systems [5,6]. Enhancing missile maneuverability during the terminal phase to improve penetration and achieve high-precision strikes has become a key area of missile guidance research [7–9].

Currently, missile maneuver penetration techniques can be classified into two main categories: programmatic maneuver penetration and game-theoretic maneuver penetration [10]. Programmatic maneuver penetration involves determining the timing and program of terminal maneuvers before the missile is launched, without accounting for potential interference from enemy interception systems. Consequently, when faced with high-precision interception systems, successful penetration becomes challenging. As a result, autonomous maneuvering has become a primary focus in the development of penetration guidance laws [11–14].

In contrast, game-theoretic maneuver penetration refers to a scenario where, upon detecting an incoming interceptor, the attack missile acquires the interceptor's flight parameters and, using its onboard computation module, calculates real-time maneuver commands to perform penetration maneuvers. This strategy allows the missile to select the optimal approach to penetration based on the interception method, thereby significantly improving the likelihood of successful penetration.

 doi:10.20944/preprints202508.0144.v2

Shinar applied a two-dimensional linearized kinematic method to analyze the penetration problem of interceptor missiles under proportional guidance and identify the key factors influencing the miss distance. A simple search algorithm was also used to determine the optimal timing and direction of the maneuver [15]. Ref.16 addresses the issue of strategy implementation for attacking missiles under limited observation by introducing a network adaptation feedback strategy and inverse game theory. It also selects strategies that meet consistency standards through optimization methods [16]. Ref.17 introduced a Linear-Quadratic (LQ) differential game approach to model the missile offense-defense interaction [17]. By combining the Hamilton-Jacobi adjoint vector with the conjugate method, the authors proposed a novel conjugate decision-making strategy and provided an analytical solution for the optimal parameters. Ref.18 presents an optimal guidance solution based on the linear quadratic differential game method and the numerical solution of the Riccati differential equation. It studies the interception problem of ballistic missiles with active defense capabilities and proposes an optimal guidance solution based on differential game strategies [18]. Ref.19 designs a cooperative guidance law by establishing a zero-sum two-player differential game framework, allowing the attack missile to intercept the target while evading the defender, and satisfying the constraint on the relative interception angle [19]. Ref.20 derived a penetration guidance law with a controllable miss distance, using optimal control theory to fit guidance parameters via a Back Propagation (BP) neural network and achieve optimal energy expenditure during the guidance process [20]. Compared with traditional pre-programmed maneuver strategies, maneuver penetration based on differential games possesses intelligent characteristics and provides real-time decision-making capabilities. However, it also presents challenges, including high computational complexity, difficulties in mathematical modeling, and the need for precise problem formulation.

With the integration of artificial intelligence technologies into differential pursuit-escape problems, novel approaches have emerged for missile terminal penetration. For the three-body pursuit-escape problem in a two-dimensional plane, Ref.21 utilized the Twin Delayed Deep Deterministic policy gradient (TD3) algorithm to train the attacker's agent, enabling it to learn a guiding strategy in response to the defender's actions, thus achieving successful target capture [21]. In the missile penetration scenario during an missile's dive phase, Ref.22 employed an enhanced Prioritized Experience Replay-Deep Deterministic Policy Gradient (PER-DDPG) algorithm, which emphasized learning from successful penetration experiences. This approach notably accelerated the convergence of the training process [22]. Ref.23 introduced a maneuvering game-based guidance law based on Deep Reinforcement Learning (DRL), which, in comparison to traditional programmatic maneuvering penetration, significantly enhanced the stability of the penetration [23]. Ref.24 proposes a hypersonic missile penetration strategy optimized using Reinforcement Meta Learning (RML), which increases the difficulty of interception through multiple random transitions [24]. Ref.25 treated the penetration process as a linear system and derived an analytical solution for the miss distance [25]. However, the penetration strategy obtained in this manner requires complete knowledge of the interceptor's guidance parameters, which is highly challenging to obtain in practical confrontation scenarios.

Most existing penetration guidance laws mainly focus on the confrontation between the attacker and defender, while neglecting the impact of penetration on strike accuracy. While penetration is critical, excessive maneuvering may cause the missile to miss the target despite successful evasion. Therefore, an integrated penetration guidance strategy is needed: one that accounts for both penetration and strike accuracy, while staying within the missile's acceleration and performance limits, and minimizing energy consumption during the penetration process. Ref. 26 innovatively designs a reward function incorporating an energy consumption factor to balance miss distance and energy efficiency. Additionally, a regression neural network is utilized to enhance the generalization capability of the penetration strategy and achieve evasion of interpret missile and precise strikes on the target [26]. Regarding the issue of missile penetration time, Ref.27 proposed an integrated guidance and strike penetration law based on optimal control, which ensures that the Line-of-sight (LOS) angular velocity

 doi:10.20944/preprints202508.0144.v2

between the attack missile and the defending interceptor reaches a specified value within a given time, thereby achieving penetration [27].

In summary, existing penetration strategies predominantly focus on the one-on-one adversarial scenario between the attacking missile and interceptor missiles, and heavily rely on the engagement context between them, often neglecting the subsequent guidance tasks. To address these issues, this paper explores the integration of optimal control and DRL, designing a guidance law that combines intelligent penetration and steering. The main contributions of this paper are as follows:

1. To address the attacking-multi interceptor-target adversarial scenario, a Markov Decision Process (MDP) model is constructed. This model takes the observable states of both sides as input and outputs the penetration acceleration commands for the attacking missile, enabling intelligent maneuvering penetration decisions in a continuous state space.
2. To tackle the coupling problem between penetration maneuvers and guidance tasks, a multi-objective reward function is designed. It maximizes the penetration success rate while constraining the maneuvering range through an energy consumption penalty term, ensuring terminal strike accuracy.
3. To overcome the training efficiency bottleneck caused by sparse rewards, a fusion of Generative Adversarial Imitation Learning (GAIL) and Proximal Policy Optimization (PPO) algorithms is proposed. Expert trajectory priors are utilized to guide exploration, significantly improving policy sampling efficiency and asymptotic performance.

The organization of this study is as follows: Section 2 establishes the mathematical model of the adversarial scenario and derives the optimal BANG-BANG penetration strategy. Section 3 constructs the MDP model for the penetration process and designs a GAIL-PPO-based hybrid training framework. Section 4 presents the training and testing experimental results. Finally, Section 5 summarizes the research conclusions.

2. Optimal BANG-BANG Penetration Strategy

In this section, we first establish a mathematical model of the engagement scenario, then derive a maximum miss distance BANG-BANG penetration strategy to provide expert experience for the GAIL-PPO training in Section 3.

2.1. Mathematical Model of Engagement Scenario

Figure 1 illustrates the planar penetration scenario. An attack missile targets a ground target, and when the target detects the incoming missile, a interceptor missiles is launched from the ground to intercept it. The interception is deemed successful when the distance R_{FD} between the attack missile and the interceptor missiles is falls below the lethal radius R_{D} of the interceptor missiles. Similarly, the attack missile is considered to have hit the target when the distance R_{FT} between them is falls below the lethal radius R_{F} of the attack missile. The initial launch position of the missiles is set at the origin of the x-axis, with the horizon corresponding to the origin of the y-axis, establishing an inertial coordinate system. In Figure 1, F, T, D1, and D2 represent the attack missile, the target, the interceptor missiles 1 and the interceptor missiles 2, respectively. The velocities of the attack missile and the interceptor missiles are denoted by $\boldsymbol{V_{\mathrm{F}}}$ and $\boldsymbol{V_{\mathrm{D}}}$. The flight path angle, denoted by $\boldsymbol{\varphi_{\mathrm{F}}}$ and $\boldsymbol{\varphi_{\mathrm{D}}}$, are measured by clockwise rotation around the x-axis. The vertical velocity components $\boldsymbol{a_{\mathrm{F}}}$ and $\boldsymbol{a_{\mathrm{D}}}$ represent the accelerations of the attack missile and interceptor missiles. Finally, the LOS angle of the attack missile and interceptor missiles are represented by $\boldsymbol{q_{\mathrm{FT}}}$ and $\boldsymbol{q_{\mathrm{FD}}}$.

 doi:10.20944/preprints202508.0144.v2

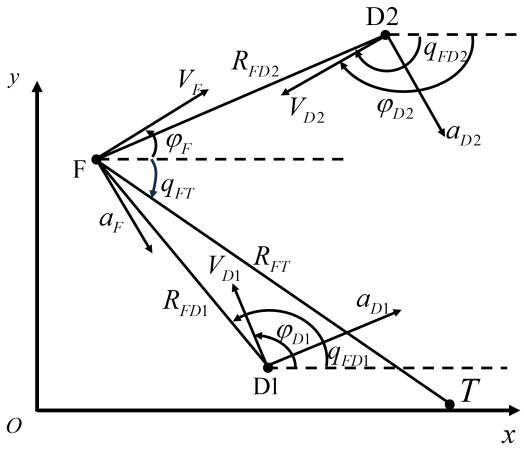


Figure 1. Missile Penetration Engagement Diagram.

Since the engagement scenario involves a hypersonic missile attacking a ground target, we approximate the target as stationary. Given that the attack missile has no propulsion during the terminal guidance phase and that air resistance is neglected, the velocity change of the attack missile in this phase is considered negligible. Similarly, we assume that the interceptor missiles has already reached its maximum speed during the penetration phase, so the velocity change of the interceptor missiles is also negligible. Therefore, both the velocities of the attack missile and the interceptor missiles can be treated as constant. From Figure 1, the relative motion equation of the attack missile and the target in the inertial coordinate system can be expressed as:

$$\begin{cases} \dot{R}_{\mathrm{FT}} = V_{\mathrm{T}}\cos(q_{\mathrm{FT}} - \varphi_{\mathrm{T}}) - V_{\mathrm{F}}\cos(q_{\mathrm{FT}} - \varphi_{\mathrm{F}}) \\ R_{\mathrm{FT}}\dot{q}_{\mathrm{FT}} = V_{\mathrm{T}}\cos(q_{\mathrm{FT}} - \varphi_{\mathrm{T}}) + V_{\mathrm{F}}\sin(q_{\mathrm{FT}} - \varphi_{\mathrm{F}}) \end{cases} \tag{1}$$

The relative motion equation between the attack missile and the interceptor missiles is given by:

$$\begin{cases} \dot{R}_{\mathrm{FD}} = V_{\mathrm{F}}\cos(q_{\mathrm{FD}} - \varphi_{\mathrm{F}}) - V_{\mathrm{D}}\cos(q_{\mathrm{FD}} - \varphi_{\mathrm{D}}) \\ R_{\mathrm{FD}}\dot{q}_{\mathrm{FD}} = V_{\mathrm{D}}\sin(q_{\mathrm{FD}} - \varphi_{\mathrm{D}}) - V_{\mathrm{F}}\sin(q_{\mathrm{FD}} - \varphi_{\mathrm{F}}) \end{cases} \tag{2}$$

At the same time, the motion models of the attack missile and the interceptor missiles can be derived as:

$$\begin{aligned} \dot{x}_{\mathrm{F}} &= V_{\mathrm{F}}\cos\varphi_{\mathrm{F}} \quad \dot{y}_{\mathrm{F}} = V_{\mathrm{F}}\sin\varphi_{\mathrm{F}} \quad \dot{\varphi}_{\mathrm{F}} = \tfrac{a_{\mathrm{F}}}{V_{\mathrm{F}}} + d \\ \dot{x}_{\mathrm{D}} &= V_{\mathrm{D}}\cos\varphi_{\mathrm{D}} \quad \dot{y}_{\mathrm{D}} = V_{\mathrm{D}}\sin\varphi_{\mathrm{D}} \quad \dot{\varphi}_{\mathrm{D}} = \tfrac{a_{\mathrm{D}}}{V_{\mathrm{D}}} \end{aligned} \tag{3}$$

where x_{F}, $y_{\mathrm{F}}$and x_{D}, y_{D} represent the current positions of the attack missile and the interceptor missiles, respectively.

During the engagement, we assume that the interceptor missiles follows a proportional guidance law to intercept the attack missile. Specifically, the guidance law is expressed as:

$$a_{\mathrm{D}} = N_{\mathrm{D}}V_{\mathrm{D}}\dot{q}_{\mathrm{FD}} \tag{4}$$

where N_{D} is the proportional gain coefficient.

 doi:10.20944/preprints202508.0144.v2

2.2. Derivation of the Optimal BANG-BANG Penetration Strategy

Differentiating both sides of the first Equation in (2) yields:

$$\begin{aligned}\ddot{R}_{\mathrm{FD}} = &- v_D \sin(q_{FD} - \varphi_D)\left(\dot{q}_{FD} - \frac{a_D}{v_D}\right) \\ &+ v_F \sin(q_{FD} - \varphi_F)\left(\dot{q}_{FD} - \frac{a_F}{v_F}\right)\end{aligned} \tag{5}$$

Let $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} R_{FD} \\ \dot{R}_{FD} \end{bmatrix}$, $A = \begin{bmatrix} 0 & 1 \\ \dot{q}_{FD}^2 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 & 0 \\ \sin(q_{FD} - \varphi_D) & \sin(q_{FD} - \varphi_F) \end{bmatrix}$, $u = \begin{bmatrix} a_D & a_F \end{bmatrix}$, we can establish the state equation:

$$\dot{x} = Ax + Bu \tag{6}$$

Let $t_{go}^{FD} = \frac{R_{FD}}{v_F - v_D}$ approximated the remaining Time-to-Go of the interceptor missiles and $t_{go}^{FD} = \frac{R_{FD}}{v_F - v_D}$ approximated the engagement time between the attacking missile and the interceptor missiles. To reduce the order of system (6), introduce the Zero-Effort Miss (ZEM) distance $z(t)$ as follows:

$$z(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \Phi\left(t_f^{FD}, t\right) x(t) \tag{7}$$

where $\Phi\left(t_f^{FD}, t\right)$ is the state transition matrix, whose expression is obtained by solving the homogeneous equation $\dot{x}(t) = Ax(t)$ as:

$$\Phi\left(t_f^{FD}, t\right) = e^{A\left(t_f^{FD} - t\right)} = \begin{bmatrix} \cosh\left(\dot{q} t_{go}^{FD}\right) & \frac{\sinh\left(\dot{q} t_{go}^{FD}\right)}{\dot{q}} \\ \dot{q} \sinh\left(\dot{q} t_{go}^{FD}\right) & \cosh\left(\dot{q} t_{go}^{FD}\right) \end{bmatrix} \tag{8}$$

where $\dot{q} = \sqrt{\dot{q}_{FD}^2} = |\dot{q}_{FD}|$.

From the properties of the state transition matrix, we obtain:

$$\dot{\Phi} = \Phi\left(t_f^{FD}, t\right) A \tag{9}$$

Differentiating Equation (7) and substituting Equation (9) yields the simplified expression:

$$\begin{aligned} z(t) &= \cosh\left(\dot{q} t_{go}^{FD}\right) R_{FD} + \frac{\sinh\left(\dot{q} t_{go}^{FD}\right)}{\dot{q}} \dot{R}_{FD} \\ \dot{z}(t) &= \frac{\sinh\left(\dot{q} t_{go}^{FD}\right)}{\dot{q}} Bu \end{aligned} \tag{10}$$

To maximize the penetration success rate as much as possible, we adopt the maximization of the miss distance between the attacking missile and the interceptor missiles as the performance metric:

$$J = \frac{1}{2}\left[z\left(t_f^{FD}\right)\right]^2 \tag{11}$$

The solution is derived using the maximum principle, establishing the Hamiltonian equation as:

$$H = \lambda \dot{z}(t) \tag{12}$$

the canonical equations are:

$$\dot{\lambda} = -\frac{\partial H}{\partial z} = 0 \tag{13}$$

From the cross-ratio condition:

$$\lambda\left(t_f^{FD}\right) = \frac{\partial J}{\partial z} = a \tag{14}$$

hence H is a linear function of a_F.

 doi:10.20944/preprints202508.0144.v2

Assuming the maximum acceleration of the attacking missile is $a_{F_{\max}}$, the penetration command of the attacking missile is derived as follows:

$$u^* = a_{F_{\max}} \cdot \text{sign}\left(\frac{\sinh\left(\dot{q}t_{go}^{FD}\right)}{\dot{q}}\sin(q_{FD} - \varphi_F)\right) \tag{15}$$

where sign denotes the sign function.

If the attacking missile employs an optimal guidance law with angle constraints to engage the target during the non-penetration phase:

$$a_{F_{guidanc}} = -\frac{2R}{t_{go}^{FT}}x_1 - 4Rx_2 \tag{16}$$

where $t_{go}^{FT} = \frac{R_{FD}}{v_F - v_D}$ approximated the remaining Time-to-Go of the attacking missile. The guidance law of the attacking missile throughout the entire engagement process is derived as follows:

$$a_F = \begin{cases} a_{F_{penetrate}} & if\ R_{FD} < R^* \\ a_{F_{giudance}} & else \end{cases} \tag{17}$$

where $a_{F_{penetrate}}$ is calculated based on Equation (15), and R^* is the penetration initiation distance.

2.3. Performance Evaluation of the BANG-BANG Penetration Strategy

To evaluate the effectiveness of the BANG-BANG penetration strategy, it is tested within the engagement scenario shown in Figure 1. The simulation parameters are summarized in Table 1 as follows:

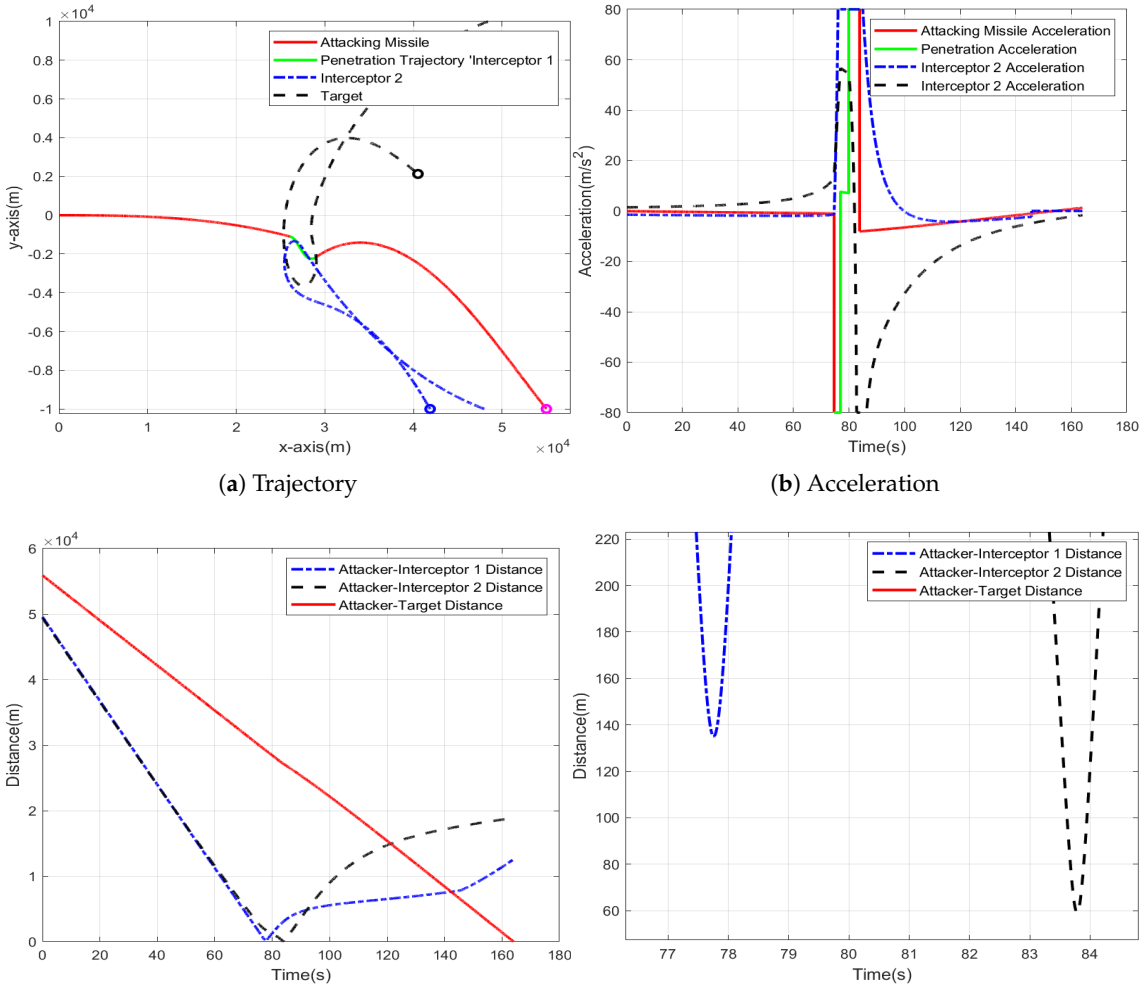
Table 1. Simulation Parameters

Parameters	Value
$a_{F_{max}}(\text{m/s}^2)$	80
$a_{D_{max}}(\text{m/s}^2)$	80
Interceptor 1 Initial Position(m)	(48000, -10000)
Interceptor 2 Initial Position(m)	(48000, 10000)
Target Position(m)	(55000, -10000)

Assuming the target is stationary, the attacking missile velocity is set to 800m/s to prevent secondary penetration. The interceptor missiles velocity is set slightly lower than that of the attacking missile at 780 m/s. The initial attack angles of the attacking missile and the two interceptor missiles are set to 0°, 170°, and 190°, respectively. Since the target is located on the ground, the desired attack angle of the attacking missile must be a positive value and neither too large nor too small. Here, it is set to $q_{FT_{end}} = \pi/3$. The lethal radius of both the attacking missile and the interceptor missiles is set to 20m. Based on experience, the penetration initiation distance is set to 2000m.

Figure 2 illustrates the test results of the BANG-BANG penetration strategy. When the distance between the attacking missile and the interceptor missiles falls below R^* , the attacking missile switches its acceleration to $a_{F_{giudance}}$. Due to the maneuverability of the attacking missile, the interceptor's acceleration rapidly saturates, allowing it to slightly overshoot the target and achieve successful penetration.

 doi:10.20944/preprints202508.0144.v2



(**a**) Trajectory (**b**) Acceleration

(**c**) Relative Distance (**d**) Relative Distance (part)

Figure 2. Simulation Results of the BANG-BANG Penetration Strategy

3. GAIL-PPO Penetration Strategy

In Section II, we derived the BANG-BANG penetration strategy that maximizes the miss distance. However, this derivation was conducted under a one-on-one engagement scenario. When facing multiple interceptor missiles simultaneously, penetration can only be achieved by switching targets, which does not guarantee a high success rate. Additionally, energy consumption during the penetration process was not considered. To address these issues, this section proposes an intelligent penetration strategy based on GAIL-PPO.

3.1. Construction of the MDP Model for the Penetration Process

In order to use DRL to solve the problem of generating penetration strategies, the penetration problem must be transformed into a DRL framework. First, a MDP model for missile penetration is constructed to define how the agent interacts with the environment to make decisions. The model primarily consists of elements such as S, A, P, R, and γ, where S represents the finite state space, with any state $s \in S$; A represents the finite action space, with any action $a \in A$; P is the state transition probability; R is the reward function; and γ is the discount factor and $\gamma \in [0, 1]$, used to calculate the accumulated reward. In the context of this missile penetration problem, the state transition probability $P = 1$ is defined, and the state space, action space, and reward function are outlined as follows:

 doi:10.20944/preprints202508.0144.v2

3.1.1. Definition of the State Space

The penetration process must consider subsequent guidance tasks, requiring the state space design to account for the states of the attack missile, intercept missile, and target. The penetration direction of the attack missile has a significant impact on its flight altitude after penetration. Different LOS angle require different flight altitudes after penetration. Smaller LOS angle prefers a lower flight altitude after penetration, while larger LOS angle prefer a higher flight altitude. Hence, we incorporate LOS-related terms into the state space to optimize the penetration direction. In order to enhance learning stability, accelerate convergence, and alleviate numerical issues, we have normalized the state space. The state space is therefore constructed as follows:

$$s = [R^*_{\mathrm{FD1}}, R^*_{\mathrm{FD2}}, R^*_{\mathrm{FT}}, \delta^*_{\mathrm{FD1}}, \delta^*_{\mathrm{FD2}}, \delta^*_{\mathrm{FT}}] \tag{18}$$

where $R^*_{\mathrm{FD1}} = R_{\mathrm{FD1}}/R_{\mathrm{FD1}}(0)$, $R^*_{\mathrm{FD2}} = R_{\mathrm{FD2}}/R_{\mathrm{FD2}}(0)$, $R^*_{\mathrm{FT}} = R_{\mathrm{FT}}/R_{\mathrm{FT}}(0)$, $\delta^*_{\mathrm{FD1}} = \tanh(q_{\mathrm{FD1}} - \varphi_{\mathrm{F}})$, $\delta^*_{\mathrm{FD2}} = \tanh(q_{\mathrm{FD2}} - \varphi_{\mathrm{F}})$, $\delta^*_{\mathrm{FT}} = \tanh(q_{\mathrm{FT}} - \varphi_{\mathrm{F}})$, $R_{\mathrm{FD}}(0)$ represents the distance between the attack missile and the intercept missile at the beginning of the penetration, and $R_{\mathrm{FT}}(0)$ represents the initial distance between the attack missile and the target.

3.1.2. Definition of the Action Space

In selecting the action space, the missile's penetration acceleration is often directly used as the output. However, due to the small sampling step size typically used during training, directly outputting the acceleration can lead to significant fluctuations in the acceleration curve during penetration, which are difficult to realize in real-world scenarios. To mitigate this issue, we select the derivative of the missile's acceleration as the action space output:

$$A = \dot{a}_{F_{penetrate}} = [-(a_{F_{\max}}), a_{F_{\max}}] \tag{19}$$

3.1.3. Definition of the Reward Function

The reward function defines the immediate feedback provided by the environment after the agent takes an action in a particular state. It influences the agent's behavior and guides it toward achieving its goal. Therefore, a well-designed reward function directly impacts the generation of penetration commands. Unlike previous approaches that use the miss distance between the attack missile and the intercept missile as the reward function, this paper chooses to use the acceleration of the intercept missile as the reward function. When the interceptor's acceleration reaches a saturation point, it indicates that the interception task has surpassed the interceptor's operational capacity. Consequently, the goal of the attack missile's penetration is to drive the interceptor's acceleration toward saturation as much as possible, thus bypassing the interceptor and achieving successful penetration. At the same time, the attack missile should aim to minimize its maneuvering range. Based on this objective, we design the instantaneous reward function as follows:

$$R_1 = \begin{cases} \frac{|a_{\mathrm{D1}}| + |a_{\mathrm{D2}}|}{a_{F_{\max}}} & a_{\mathrm{D1}} < a_{F_{\max}} \ \text{ or } \ a_{\mathrm{D2}} < a_{F_{\max}} \\ 2 - \frac{|a_{F_{penetrate}}|}{a_{F_{\max}}} & |a_{\mathrm{D1}}| = a_{F_{\max}} \ \text{ and } \ |a_{\mathrm{D2}}| = a_{F_{\max}} \end{cases} \tag{20}$$

where $|a_{\mathrm{D1}}|$ and $|a_{\mathrm{D2}}|$ represent the accelerations sizes of two intercept missiles, respectively.

The terminal reward function is designed based on whether the task is successful:

$$R_2 = \begin{cases} -500 - J & R_{\mathrm{FD}} < R_{\mathrm{D}} \ \text{ or } \ \text{Miss} \ \ \text{Target} \\ 500 - J & R_{\mathrm{FT}} < R_{\mathrm{T}} \end{cases} \tag{21}$$

 doi:10.20944/preprints202508.0144.v2

where

$$J = \int \frac{1}{2}\left(\frac{a_F}{a_{F_{\max}}}\right)^2 dt$$

represents the energy consumption term, R_D represents the kill radius of the intercept missile, R_T represents the kill radius of the attack missile. When the attack missile is intercepted or the mission fails, a large penalty is applied. Conversely, when the attack missile successfully hits the target, a large reward is given.

Considering both the penetration effect and task completion status, the function is designed as follows:

$$R = \begin{cases} R_1 & \text{Penetrating Defense} \\ R_2 & \text{Penetrated Completed} \end{cases} \tag{22}$$

3.2. GAIL-PPO Algorithm

3.2.1. GAIL Training Network Construction

Generative Adversarial Imitation Learning (GAIL) learns a policy through a generative adversarial approach, aiming to make the generated behavior as similar as possible to the expert behavior. Its main structure is shown in Figure 3:

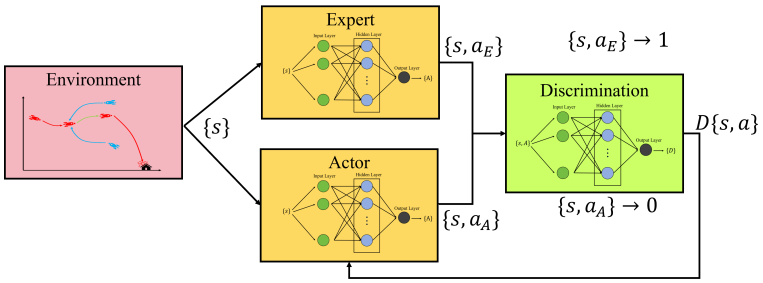


Figure 3. GAIL structure.

GAIL primarily consists of an Actor Network and a Discriminator Network. When the environment provides a state, both the Actor and the Expert generate corresponding actions. These state-action pairs are fed into the Discriminator, which outputs a real number between (0, 1). The Discriminator's task is to push the Expert's output closer to 0 and the Actor's output closer to 1, while the Actor's objective is to deceive the Discriminator as much as possible. Consequently, the loss functions for both the Actor and the Discriminator are formulated as follows:

$$\begin{aligned} L_{\text{Actor}} &= -E_{\tau\sim\tau_{\text{Expert}}}\left[\log D(\tau)\right] - E_{\tau\sim\tau_{\text{Actor}}}\left[1 - \log D(\tau)\right] \\ L_{\text{Discriminator}} &= -E_{\tau\sim\tau_{\text{Actor}}}\left[\log D(\tau)\right] \end{aligned} \tag{23}$$

where τ_{Expert} and τ_{Actor} denote the state-action pairs generated by the Expert and the Actor, respectively, and $D(\tau)$ represents the Discriminator's probability prediction that the state-action pair belongs to the Expert.

The Actor and the Discriminator form an adversarial process. Ultimately, the data distribution generated by the imitator policy will approach the real expert data distribution, achieving the goal of imitation learning.

3.2.2. PPO Training Network Construction

Proximal Policy Optimization (PPO) is a reinforcement learning algorithm designed for efficient and stable policy optimization [28]. Its core objective is to train agents that maximize cumulative rewards by updating policy parameters via gradient ascent. PPO-CLIP achieves this through two key innovations:

 doi:10.20944/preprints202508.0144.v2

PPO-CLIP uses a probability ratio to measure policy changes:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \tag{24}$$

where π_θ and $\pi_{\theta_{\text{old}}}$ denote the current policy and the old policy, respectively.

A clipped objective function controls update magnitude:

$$L_{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t\hat{A}_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{25}$$

where $\hat{A}_t$ is the advantage function and ϵ (typically 0.1-0.2) clips excessive policy changes. This prevents destructive updates while allowing monotonic improvement.

To reduce variance in advantage estimation, we use Generalized Advantage Estimation (GAE):

$$\hat{A}_t^{\text{GAE}} = \sum_{i=0}^{n} (\gamma\lambda)^i \delta_{t+i} \tag{26}$$

with temporal difference error:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{27}$$

where γ is the discount factor and λ balances bias-variance tradeoff.

The critic network $V_\phi(s_t)$ is optimized by minimizing mean squared error:

$$L_V(\phi) = \mathbb{E}_t\left[\left(V_\phi(s_t) - (r_t + \gamma V_\phi(s_{t+1}))\right)^2\right] \tag{28}$$

This provides stable value estimates for advantage calculation and policy updates.

In the GAIL-PPO algorithm, both networks share a single Actor Network. Taking into account the state space and action space of the model, the parameter settings for the three networks are summarized in Table 2 as follows:

Table 2. Network Architecture

Number of floors	Actor	Critic	Discriminator
Input Layer	6(states)	6(states)	7(states)
Hidden Layer 1	256	256	256
BatchNorm 1	ones	ones	ones
Activation Function 1	Relu	Relu	Relu
Hidden Layer 2	256	256	256
BatchNorm 1	ones	ones	ones
Activation Function 2	Relu	Relu	Relu
Output Layer 1	1(Mean output)	1(Value function)	1($D(\tau)$)
Output Layer 2	1(Standard deviation output)	–	–

where the BatchNorm layer reduces internal covariate shift by normalizing the input distribution. It accelerates training, enhances convergence stability, reduces sensitivity to weight initialization and learning rates, and alleviates the issues of gradient vanishing and exploding.

3.2.3. Training Procedure of the GAIL-PPO Algorithm

The training process of the GAIL-PPO algorithm is illustrated in Figure 4:

 doi:10.20944/preprints202508.0144.v2

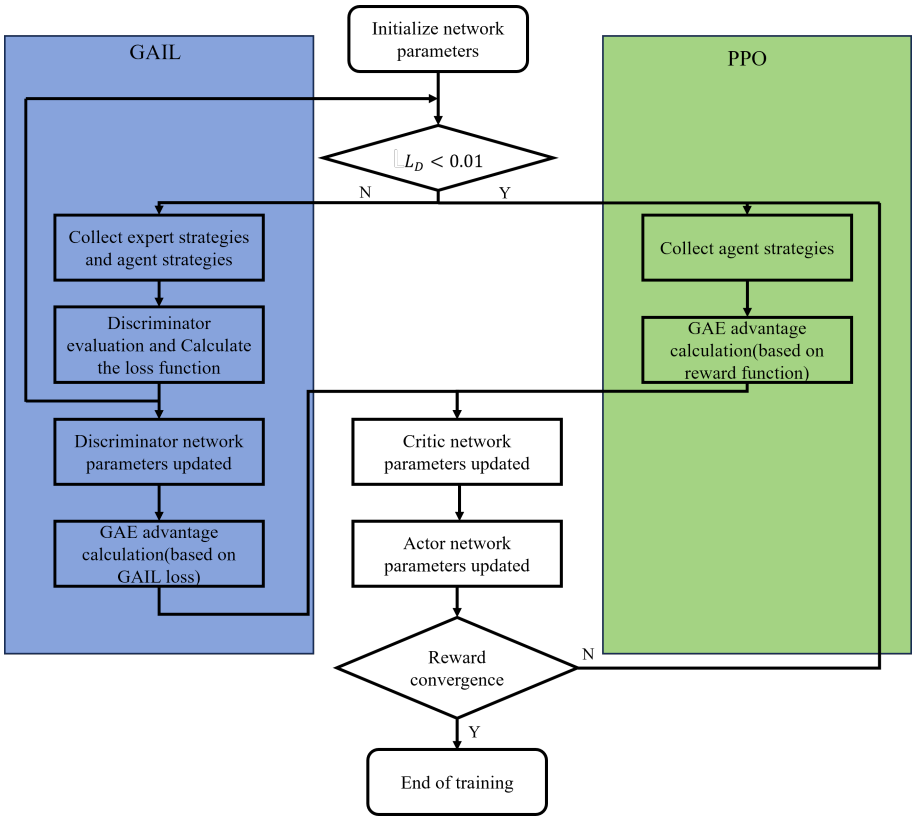


Figure 4. Training Procedure

First, the parameters of the Discriminator, Actor, and Critic networks are initialized. Subsequently, the training enters the GAIL pre-training phase. In this phase, when the Discriminator loss is $L_D \geq 0.01$, expert trajectories and agent trajectories are collected first. The Discriminator parameters are then updated by calculating the cross-entropy loss. Following this, the GAIL reward is computed based on the Discriminator's output, and the advantage function is estimated using GAE. The Critic parameters are updated with the Mean Squared Error (MSE) loss, while the Actor parameters are updated using the PPO clipped objective function. When the Discriminator loss reaches $L_D < 0.01$, the training transitions to the PPO fine-tuning phase. At this stage, only agent trajectories are collected, and the environment's true reward replaces the GAIL reward. The GAE advantage function is recomputed, and the Actor-Critic network is updated again using the same PPO clipped objective function. The training terminates when the average reward remains below a threshold for consecutive iterations.

The GAIL-PPO algorithm is shown in Table 3:

 doi:10.20944/preprints202508.0144.v2

Table 3. Simplified GAIL-PPO Algorithm Pseudocode

Algorithm 2 Simplified GAIL-PPO

1. Input: initial policy parameters θ_0, initial value function parameters ϕ_0, initial discriminator parameters ψ_0.
2. While $k <$ max_iter do
3. Compute $L_D[(\psi)]$ (Equation 23).
4. If $L_D(\psi) \geq \epsilon_D$
5. Collect $\mathcal{D}_{\text{agent}} = \{\tau_i\}$ via π_θ.
6. Update discriminator: $\psi \leftarrow \psi - \alpha_D \nabla_\psi L_D(\psi)$ (Adam).
7. Compute GAIL rewards: $r_t^{\text{gail}} = -\log(1 - D_\psi(s_t, a_t))$.
8. Else
9. Collect $\mathcal{D}_{\text{agent}} = \{\tau_i\}$ via π_θ.
10. Use environment rewards (Equation 22).
11. End if
12. Compute rewards-to-go:
$$\hat{R}_t = \sum_{i=0}^{T-t-1} \gamma^i r_{t+i+1} \text{ (use } r_t^{\text{gail}} \text{ or } r_t^{\text{env}}).$$
13. Compute GAE advantage (Equation 26).
14. Update policy via PPO-CLIP:
$$\theta_{k+1} = \arg\max_\theta \frac{1}{|\mathcal{D}_{\text{agent}}|T} \sum_{\tau,t} \min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \text{clip}(\cdot, 1-\epsilon, 1+\epsilon)\hat{A}_t \right).$$
15. Update value func. via MSE:
$$\phi_{k+1} = \arg\min_\phi \frac{1}{|\mathcal{D}_{\text{agent}}|} \sum_{\tau,t} (V_\phi(s_t) - \hat{R}_t)^2.$$
16. If $\text{avg}(R_t) < R_{\text{thres}}$ for N consecutive rounds, break.
17. $k \leftarrow k + 1$.
18. End while
19. Output: θ.

This training process leverages GAIL pre-training to rapidly learn expert behavior patterns and then employs PPO with true rewards to optimize the performance upper bound, combining the advantages of imitation learning and reinforcement learning.

4. Training Results and Performance Validation

In this section, the training method proposed in section 3 is first applied, using the BANG-BANG penetration strategy introduced in Section 2 as the expert experience to train the GAIL-PPO algorithm. Subsequently, the trained strategy is compared with traditional methods, and the effects of different parameters on penetration performance are evaluated. The engagement scenarios are consistent with those defined in Section II, and the parameter settings remain the same as in Section II unless otherwise specified.

4.1. Training Results

Before starting the training, the necessary parameters are set:

Table 4. Parameters for Training

Parameters	Value
Discount Factor	0.99
Clip Factor	0.1
Entropy Loss Weight	0.05
GAE Factor	0.95
Mini Batch Size	128
Experience Horizon	1024
Sample Time	0.01
Interceptor 1 Initial Position(m)	([45000,50000], -10000)
Interceptor 2 Initial Position(m)	([45000,50000], 10000)
Target Position(m)	([50000,55000], -10000)
LOS angle(rad)	$[\pi/3, \pi/2]$

 doi:10.20944/preprints202508.0144.v2

As shown in Figure 5 and Figure 6, the blue curve represents the per-step reward, while the red curve represents the moving average reward. The GAIL-PPO algorithm, leveraging expert trajectories from imitation learning, achieves significant improvements in training efficiency, final performance, and policy stability.In terms of convergence speed, the GAIL-PPO algorithm achieves an average reward of 300 at 1,700 training steps, while pure PPO requires 3,400 steps to reach the same reward level under identical hardware conditions. This represents a 50% reduction in training steps. This suggests that GAIL provides effective exploration priors for PPO by imitating expert behavior, thereby reducing ineffective attempts during random exploration. Regarding final performance, in the later stages of training, the average reward of GAIL-PPO stabilizes at 350, which is 94.4% higher than the 180 achieved by pure PPO. Additionally, the fluctuation range of the per-step reward for GAIL-PPO is significantly narrower, demonstrating more consistent action selection under similar states and reducing errors caused by random exploration. These characteristics directly correlate with task performance: higher average rewards with reduced fluctuations imply that GAIL-PPO can more stably reproduce high-reward successful penetration behaviors, resulting in a significantly higher penetration success rate compared to the pure PPO algorithm.

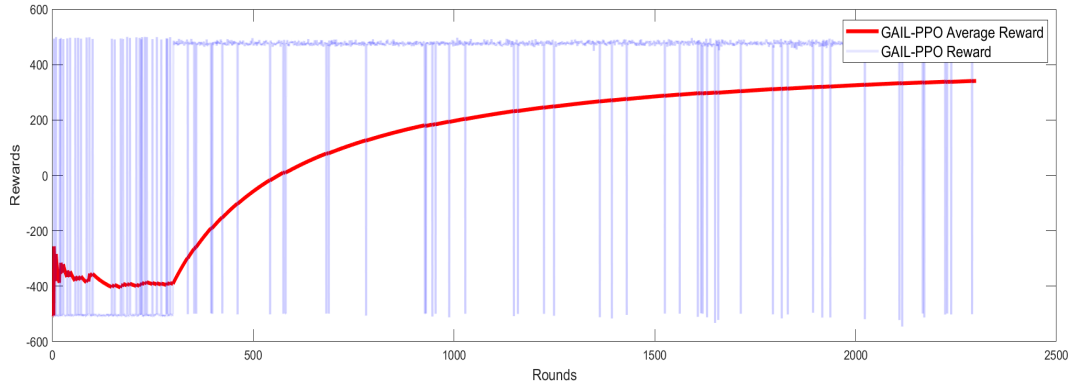


Figure 5. GAIL-PPO Training Reward

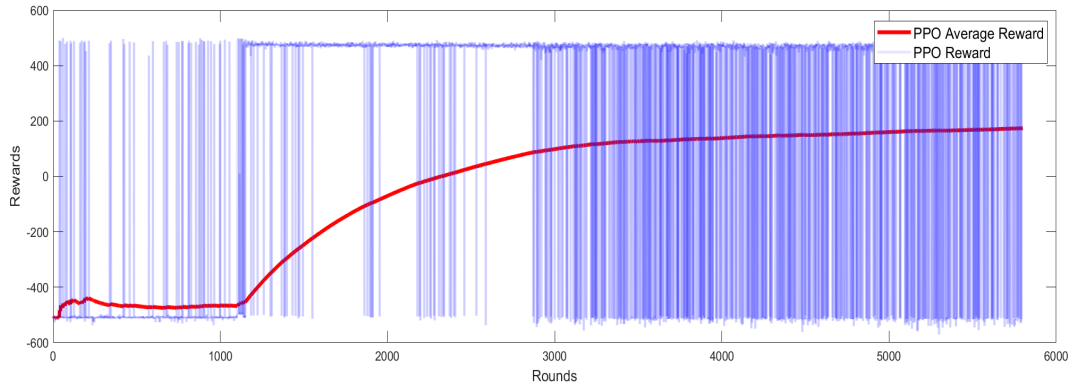


Figure 6. PPO Training Reward

The above results fully validate the effectiveness of the GAIL-PPO framework. Imitation learning provides RL with expert priors, addressing the core challenges in missile penetration tasks—sparse rewards (only successful penetration yields high rewards) and high exploration costs (incorrect actions incur heavy penalties). This significantly improves training efficiency, final performance, and policy stability. The imitation-reinforcement hybrid paradigm offers an optimized solution for training intelligent agents in complex tasks such as missile penetration.

4.2. Performance Validation

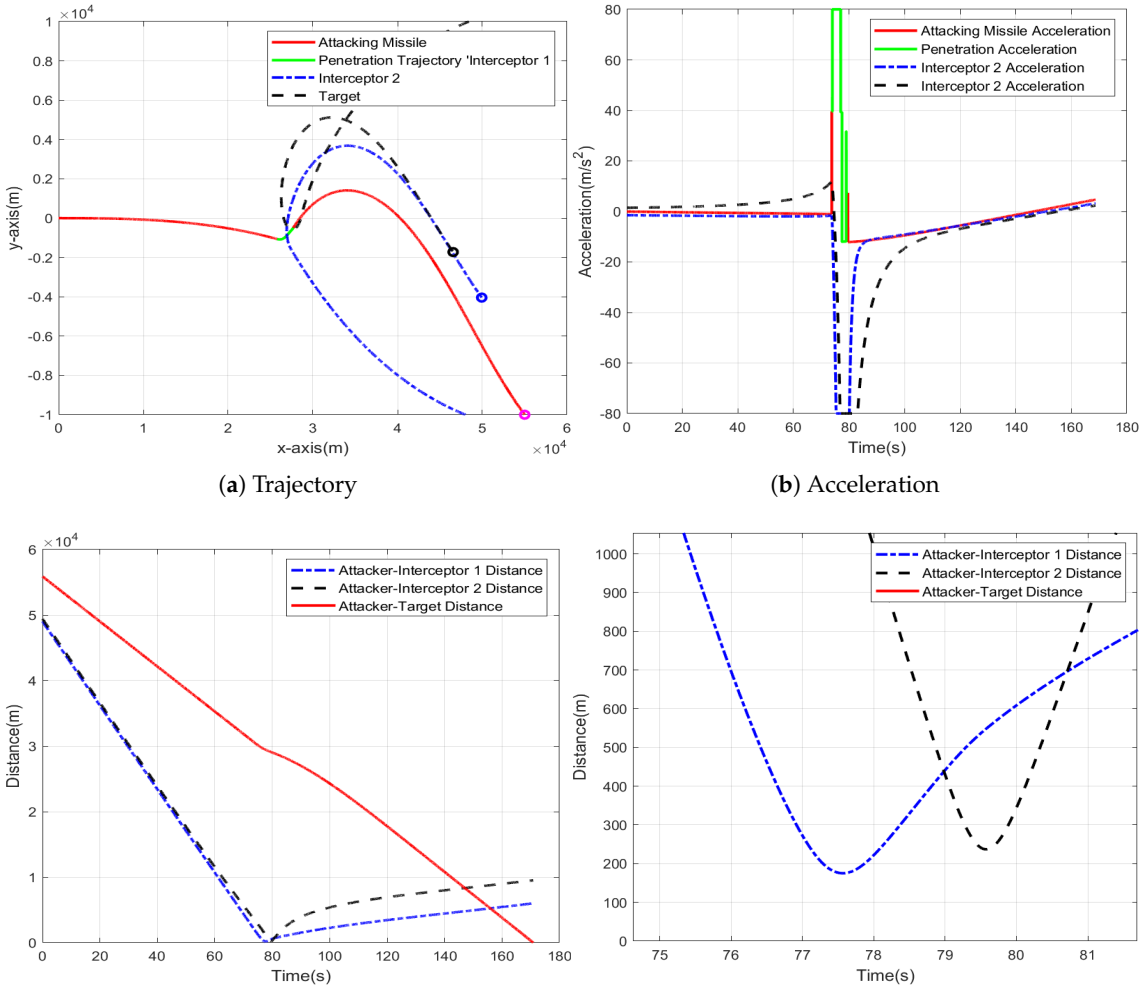
To validate the performance of the GAIL-PPO penetration strategy, the same simulation parameters as those used for the BANG-BANG strategy are adopted. The results are shown in Figure 7, while

 doi:10.20944/preprints202508.0144.v2

Figure 8 presents the energy consumption calculated based on Equation (29) over the entire mission for both strategies.

$$J = \frac{1}{2}\int a^2 dt \tag{29}$$

Compared to the BANG-BANG strategy, which requires stepwise maneuvers against each interceptor, the GAIL-PPO strategy achieves synchronized avoidance of multiple interceptor missiles through a single continuous maneuver. This eliminates the structural acceleration risk caused by sustained saturated acceleration in the BANG-BANG strategy. Additionally, the GAIL-PPO strategy significantly reduces the maneuvering range. This not only results in a 51% reduction in energy consumption compared to the BANG-BANG strategy but also creates more favorable conditions for guidance tasks following penetration.



(**a**) Trajectory (**b**) Acceleration

(**c**) Relative Distance (**d**) Relative Distance (part)

Figure 7. Simulation Results of Performance Validation

 doi:10.20944/preprints202508.0144.v2

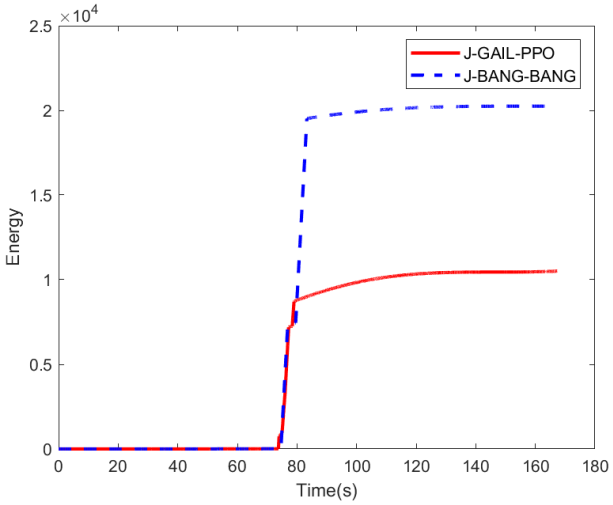


Figure 8. Energy Consumption

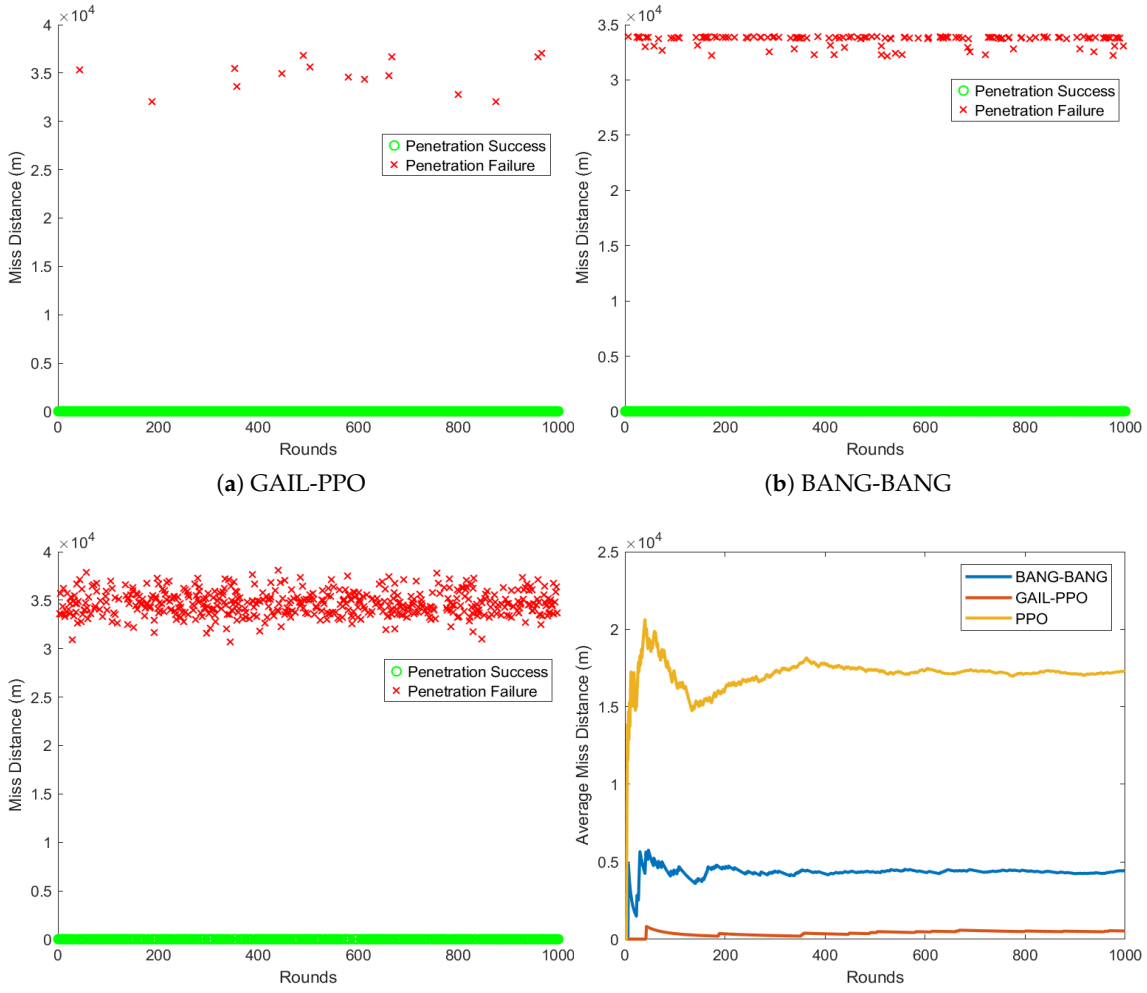
4.3. Monte Carlo Simulation

The previous section demonstrated, through a single-case simulation, the advantages of the GAIL-PPO strategy in terms of trajectory accuracy, maneuver efficiency, and energy consumption. However, it did not account for the inevitable uncertainties present in real penetration tasks. To systematically evaluate the robustness and statistical significance of the GAIL-PPO strategy, this section conducts 1,000 Monte Carlo simulations under both training parameters and non-training parameters. Key metrics, including penetration success rate, average energy consumption, mission time, and minimum interception distance, are compared between the GAIL-PPO and BANG-BANG strategies to quantitatively validate the comprehensive performance advantages of GAIL-PPO.

4.3.1. Testing in Training Parameters

Figure 9 shows the results of 1,000 Monte Carlo simulations, demonstrating that the GAIL-PPO strategy achieves a successful penetration rate of 98.5%, significantly higher than the 86.9% for the BANG-BANG strategy and 50.2% for the PPO strategy. In terms of the average miss distance over 1,000 simulations, the GAIL-PPO strategy stabilizes at 540 m, which is much lower than the 4,419 m for the BANG-BANG strategy and 17,273 m for the PPO strategy. These results indicate that the GAIL-PPO strategy exhibits significant advantages in three key metrics: penetration success rate, miss distance accuracy, and policy stability.

 doi:10.20944/preprints202508.0144.v2



(**a**) GAIL-PPO (**b**) BANG-BANG

(**c**) PPO (**d**) Average Miss Distance

Figure 9. Monte Carlo simulation results under the training scenario

4.3.2. Testing Under Non-Training Parameters

To evaluate the robustness and generalization ability of the proposed penetration strategy, it is applied to scenarios beyond the training data. The corresponding simulation parameters are summarized in Table 5. Untrained scenarios preserve the same dynamics but vary key parameters (initial relative position, LOS angle) to probe generalization. To enhance the penetration challenge, a disturbance error is introduced into the launch angle of the attacking missile. Furthermore, the target is no longer stationary but moves with a constant linear velocity along the horizontal plane in the direction of the attacking missile, simulating an evasive maneuver. The acceleration command for the intercept missile is governed by an interception guidance law specifically designed for high-speed maneuvering targets, as detailed in Reference [29].

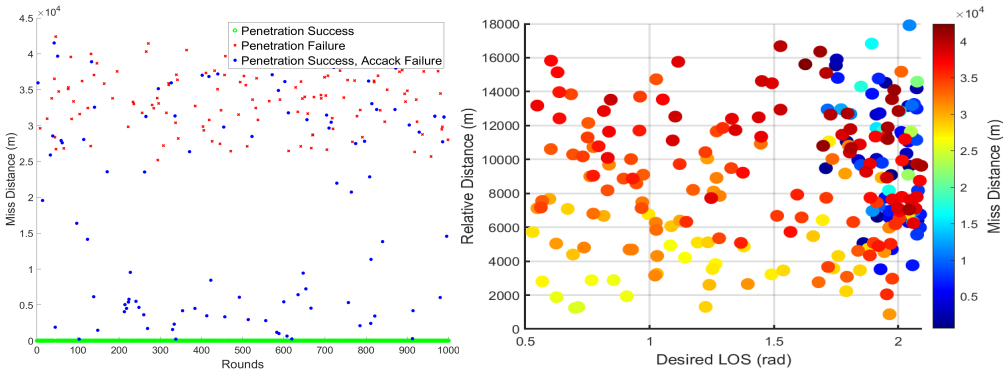
$$a_{\mathrm{D}} = \begin{cases} \min\left\{ \frac{NV_{\mathrm{D}}\dot{q}_{\mathrm{FD}}}{\cos(q_{\mathrm{FD}}-\varphi_{\mathrm{D}})}, \operatorname{sgn}\left(\frac{NV_{\mathrm{D}}\dot{q}_{\mathrm{FD}}}{|\cos(q_{\mathrm{FD}}-\varphi_{\mathrm{D}})|}\right)|a_{D\max}| \right\} & \left(q_{\mathrm{FD}}-\varphi_{\mathrm{D}} \neq \pm\frac{\pi}{2}\right) \\ 0 & \left(q_{\mathrm{FD}}-\varphi_{\mathrm{D}} = \pm\frac{\pi}{2}\right) \end{cases} \tag{30}$$

 doi:10.20944/preprints202508.0144.v2

Table 5. Non-Training Scenarios Parameter Settings

Parameters	Value
Interceptor 1 Initial Position(m)	([42000,52000], -10000)
Interceptor 2 Initial Position(m)	([42000,52000], 10000)
Target Position(m)	([50000,60000], -10000)
LOS angle(rad)	$[\pi/6, 2\pi/3]$
φ_F(rad)	$[-\pi/20, \pi/20]$
V_T(m/s)	20

Figure 10 displays the results of 1000 Monte Carlo simulation runs under the untrained scenario. Figure (a) presents the distribution of miss distances from the simulation results. It shows that the GAIL-PPO penetration strategy proposed in this paper performs well even in more complex adversarial scenarios. Despite facing more sophisticated interceptor missiles and certain disturbances, the penetration success rate reaches 86.3%. Figure (b) illustrates the relationship between miss distance, desired LOS, and the initial relative position of the target and interceptor. As shown in the figure, regions with a larger desired LOS angle are populated with numerous cases of smaller miss distances, which aligns with the distribution of cases where penetration is successful but the mission fails, as observed in Figure (a). This indicates that when dealing with large desired LOS angle, even if the attacking missile successfully penetrates, the subsequent strike mission becomes highly challenging due to insufficient altitude after penetration. Nevertheless, the attack success rate still achieves 77%.These simulation results validate that the comprehensive penetration guidance law, designed based on deep reinforcement learning in this paper, maintains excellent performance across various complex environments. The proposed method demonstrates strong robustness and good generalization capabilities, even when encountering adversarial scenarios with unknown characteristics, achieving a high mission success rate.



(**a**) Distribution of Miss Distances (**b**) Miss Distance under Different Initial Parameters

Figure 10. Monte Carlo simulation results under the Non-Training Scenarios

5. Conclusions

This paper proposes an intelligent penetration strategy combining optimal BANG-BANG law and DRL, aiming to address the penetration problem faced by missiles during the terminal guidance phase when confronting multiple interceptor missiles. The key research contributions are summarized as follows:

1. The BANG-BANG penetration strategy, which maximizes the miss distance in a one-on-one attacking missile-interceptor scenario, is derived and utilized as expert experience for GAIL training.
2. An MDP model tailored for penetration and guidance adversarial scenarios is established. A reward function is designed to reduce energy consumption while ensuring mission success, considering both penetration and guidance tasks comprehensively.

 doi:10.20944/preprints202508.0144.v2

3. A combined GAIL-PPO agent training method is proposed. Compared to the pure PPO algorithm, the convergence speed is improved by 50%.
4. Monte Carlo simulation results validate the effectiveness of the proposed strategy. In the trained parameter scenarios, the penetration success rate reaches 98.5%, significantly outperforming both the BANG-BANG strategy and the PPO strategy. Even in untrained scenarios, the strategy achieves a penetration success rate of 86.3% and a mission success rate of 77%, demonstrating its robustness and generalization ability.

References

1. Wright, D.; Tracy, C.L. Hypersonic Weapons: Vulnerability to Missile Defenses and Comparison to MaRVs. *Sci. Glob. Secur.* **2023**, *31*, 68–114. doi:10.1080/08929882.2023.2270292.
2. Jon, H.; Oscar, W. S-400 and S-500: Russia's Long-Range Air Defenders. *Jane's Int. Def. Rev.* **2019**, *52*, 56–60.
3. Yu, J.; Dong, X.; Li, Q. et al. Distributed Cooperative Encirclement Hunting Guidance for Multiple Flight Vehicles System. *Aerosp. Sci. Technol.* **2019**, *95*, 105475.
4. Yu, J.; Dong, X.; Li, Q. et al. Distributed Adaptive Cooperative Time-Varying Formation Tracking Guidance for Multiple Aerial Vehicles System. *Aerosp. Sci. Technol.* **2021**, *117*, 106925.
5. Zhan, Y.; Li, S.Y.; Zhou, D. Time-to-Go Based Three-Dimensional Multi-Missile Spatio-Temporal Cooperative Guidance Law: A Novel Approach for Maneuvering Target Interception. *ISA Trans.* **2024**, *149*, 178–195.
6. Jiang, Q.J.; Wang, X.G.; Bai, Y.L. et al. Intelligent Game-Maneuvering Policy for Reentry Glide Vehicle in Diving Phase. *J. Astronaut.* **2023**, *44*, 851–862.
7. Shen, Z.P.; Yu, J.L.; Dong, X.W. et al. Penetration Trajectory Optimization for the Hypersonic Gliding Vehicle Encountering Two Interceptors. *Aerosp. Sci. Technol.* **2022**, *121*, 107363.
8. Guo, R.; Ding, Y. et al. An Intelligent Penetration Guidance Law Based on DDPG for Hypersonic Vehicle. In *Proc. ICCES 2023: Comput. Exp. Simul. Eng.*, 2024; pp. 1349–1361.
9. Liu, S.X.; Liu, S.J. et al. Current Developments in Foreign Hypersonic Vehicles and Defense Systems. *Air Space Def.* **2023**, *6*, 39–51.
10. Guo, X. *Penetration Game Strategy for Hypersonic Vehicles*. Ph.D. Thesis, Northwestern Polytechnical University, Xi'an, China, 2018.
11. Ren, L.L.; Guo, W.L. et al. Deep Reinforcement Learning Based Integrated Evasion and Impact Hierarchical Intelligent Policy of Exo-Atmospheric Vehicles. *Chin. J. Aeronaut.* **2025**, *38*, 103193.
12. Liu, P.; Yin, H.; Wang, W.D. et al. Maneuvering Trajectory Planning During the Whole Phase Based on Piecewise Radau Pseudospectral Method. In *Proc. 37th Chin. Control Conf.*, Wuhan, China, 2018; pp. 4628–4632.
13. Zarchan, P. Proportional Navigation and Weaving Targets. *J. Guid. Control Dyn.* **1995**, *18*, 969–974.
14. Sahlholm, T.; Sahlholm, A.; Putaala, A. Simple Missile Models Against High-G Barrel Roll Maneuver. In *Proc. AIAA Guid. Navig. Control Conf.*, Portland, USA, 2011; pp. 1–12.
15. Singh, S.K.; Reddy, P.V. Dynamic Network Analysis of a Target Defense Differential Game With Limited Observations. *IEEE Trans. Control Netw. Syst.* **2023**, *10*, 308–320.
16. Segal, A.; Miloh, T. Novel Three-Dimensional Differential Game and Capture Criteria for a Bank-to-Turn Missile. *J. Guid. Control Dyn.* **1994**, *17*, 1068–1074.
17. Liang, H.Z.; Wang, J.Y. et al. Optimal Guidance Against Active Defense Ballistic Missiles via Differential Game Strategies. *Chin. J. Aeronaut.* **2020**, *33*, 978–989.
18. Liu, F.; Dong, X.W. et al. Cooperative Differential Games Guidance Laws for Multiple Attackers Against an Active Defense Target. *Chin. J. Aeronaut.* **2022**, *35*, 374–389.
19. Xie, R.H.; Ding, Y. et al. Research on a New Maneuver Penetration Strategy of Ballistic Missile. *Command Control Simul.* **2021**, *43*, 12–17.
20. Gavra, V.; Cook, A. et al. Missile Avoidance Using Reinforcement Learning. *AIAA SCITECH 2025 Forum*; `10.2514/6.2025-0105`.
21. Jacob, T.; Jay, P. Defender-Aware Attacking Guidance Policy for the Target-Attacker-Defender Differential Game. *J. Aerosp. Inf. Syst.* **2021**, *18*, 366–376.
22. Jiang, Q.J.; Wang, X.G. et al. Intelligent Game-Maneuvering Policy for Reentry Glide Vehicle in Diving Phase. *J. Aerosp. Inf.* **2023**, *44*, 851–862.
23. Gaudet, B.; Furfaro, R. Terminal Adaptive Guidance for Autonomous Hypersonic Strike Weapons via Reinforcement Meta Learning. *J. Spacecr. Rockets* **2023**, *60*, 286–298.

 doi:10.20944/preprints202508.0144.v2

24. Wang, X.F.; Gu, K.R. A Penetration Strategy Combining Deep Reinforcement Learning and Imitation Learning. *J. Astronaut.* **2023**, *44*, 914–925.
25. Yao, D.D.; Xia, Q.L. Finite-Time Convergence Guidance Law for Hypersonic Morphing Vehicle. *Aerospace* **2024**, *11*, 680.
26. Yan, T.; Jiang, Z.; Li, T.; et al. Intelligent maneuver strategy for hypersonic vehicles in three-player pursuit-evasion games via deep reinforcement learning. *Front Neurosci* **2024**, *18*, 1362303.
27. Wang, X.F.; Zhang, X. et al. Integrated Strategy of Penetration and Attack Based on Optimal Control. *Flight Dyn.* **2022**, *40*, 51–71.
28. Schulman, J.; Wolski, F. et al. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
29. Bai, G.Y.; Shen, H.R. et al. Study on Omni-Directional Interception Guidance Law for High-Speed Maneuvering Targets. *J. Equip. Acad.* **2016**, *27*, 75–80.