

Research on Offensive Autonomous Vehicle Trajectory Prediction Method Based on Improved Inverse Reinforcement Learning in Aircraft Active Defense Mode

ZHANG Hao[1], LIU Jianing[1], XU Zhi[1], YANG Yuanxin[1]

[1] School of Astronautics, Northwestern Polytechnical University, Xi'an 710129, China

Abstract

With the improvement of airborne fire control systems and situational awareness capabilities, the defense strategies against offensive autonomous vehicles are evolving from passive means such as jamming and deception to the active defense mode that intercepts incoming offensive autonomous vehicles via defensive interceptors. Nevertheless, interceptors feature low average velocity, limited defense range and insufficient overload ratio, which fail to meet the accurate collision requirements of traditional proportional navigation laws, thereby posing new challenges to the trajectory prediction of incoming offensive autonomous vehicles. Aiming at the high-probability prediction of interceptor guidance information in the three-body active defense scenario consisting of aircraft, incoming offensive autonomous vehicle and interceptor, this paper proposes an incoming offensive autonomous vehicle trajectory prediction method based on inverse reinforcement learning. Firstly, a mathematical model is established to extract the temporal maneuver characteristics of incoming offensive autonomous vehicles under the maximum causal entropy, and a behavioral strategy function library for incoming offensive autonomous vehicle guidance laws is constructed within the inverse reinforcement learning framework. Secondly, the calculation formula of inverse reinforcement learning strategy function based on quadratic form is derived to reduce the computational complexity of strategy functions in high-dimensional state spaces. Finally, the weighting coefficients of strategy functions are calculated online by using rolling window measured data, so as to realize real-time optimal selection and adaptive weighting of trajectory prediction distributions, and establish a real-time trajectory prediction model for incoming offensive autonomous vehicles. Simulation results verify that the proposed trajectory prediction algorithm possesses strong generalization ability outside the model and sample sets in three-body active defense scenarios. It exhibits favorable dynamic adaptability to complex target maneuvers and high prediction accuracy. The established high-probability trajectory prediction model can be applied to interceptor guidance, which presents certain theoretical application significance and engineering reference value.

Keywords: Three-body active defense scenario; Autonomous vehicles; Inverse Reinforcement Learning; Trajectory prediction.

1. Introduction

With the continuous performance improvement of offensive autonomous vehicles (OAVs), passive defense tactics including aerodynamic evasion and electromagnetic jamming gradually fail to satisfy modern aerial combat defense demands. Global research efforts are concentrated on airborne active defense systems, which deploy interceptors to counter incoming hostile vehicles actively [1,2]. In actual interception engagement, interceptors often possess no superiority in overload capacity. Intense line-of-sight variation, limited detection coverage and various practical constraints further raise interception difficulty, making precise trajectory forecasting an essential prerequisite for successful collision interception [3,4].

Current trajectory prediction studies predominantly focus on hypersonic near-space interception missions, where researchers summarize technical bottlenecks and developmental trends of interception guidance technologies [5]. Distinct from conventional combat modes, three-body active defense confrontations feature brief combat duration and fierce short-range maneuvering. Attack

intentions of incoming vehicles are relatively explicit, and battlefield observation data maintain steady quality. Restricted by small speed ratio and defensive boundary constraints, interceptors cannot acquire sufficient overload advantage against attacking targets [4]. Such performance limitations lead to divergent line-of-sight angular rates under classic proportional navigation guidance, producing large terminal miss distance and preventing effective lethal interception [6].

Core operational properties of three-body active defense scenarios can be concluded as follows [7,8]:

- Multi-layer interactive behaviors exist throughout the engagement, covering offensive strike launched by incoming OAVs, defensive interception implemented by interceptors and self-evasion maneuver conducted by protected aircraft;
- Incoming OAVs hold definite attack preference toward the protected aircraft, which narrows prediction boundary and effectively reduces trajectory forecasting uncertainty.

Nevertheless, the motion trajectories of non-cooperative attacking targets are driven by complicated nonlinear temporal evolution rules. Both offensive vehicles and defensive aircraft fly at high speed accompanied by drastic agile maneuvers. The flight state of incoming OAVs is strongly coupled with the real-time position and evasive actions of the defended platform. Conventional prediction methods lack effective modeling capability to mine implicit combat intentions and inherent guidance rules of hostile targets, meanwhile unreasonable hypothesis on target maneuver characteristics further lowers prediction precision. Existing algorithms cannot meet the practical requirements of three-body active defense missions. Dramatic battlefield dynamics and diversified guidance strategies of incoming vehicles bring new obstacles and challenges to reliable trajectory prediction under active defense scenarios.

It is specifically defined in this research that incoming OAVs have locked the aircraft as the exclusive strike objective. This paper mainly explores interception schemes for interceptors to eliminate threatening OAVs and safeguard the defended aircraft, centering on solving the trajectory prediction problem of uncooperative attacking targets.

Existing prediction approaches for incoming OAVs in three-body defense scenarios fall into two mainstream categories: state filtering estimation and deep neural network fitting. Model-based filtering and parameter identification methods integrate real-time measurement data to achieve optimal estimation of target states and time-varying unknown parameters, which suit targets adopting simple guidance laws. Fonod and Shima [9] adopted the Multiple Model Adaptive Estimator to identify typical guidance modes including proportional navigation and augmented proportional navigation. Multi-model collaborative frameworks are verified feasible to inversely deduce guidance parameters from limited kinematic observation data [10]. Frequency-domain optimized filtering methods are also designed to enhance anti-interference performance during guidance law recognition [11]. Nevertheless, these methods commonly adopt linearized collision geometry or uniform velocity simplification, resulting in tedious derivation process and unsatisfactory prediction accuracy.

In contrast, deep learning based methods reserve valid temporal information from high-dimensional trajectory sequences, fitting the short-period maneuver characteristics of incoming OAVs and supporting fast guidance identification and trajectory estimation. Hybrid algorithms combining Volterra series and ensemble learning can excavate latent motion laws from historical flight data to realize stable maneuver trajectory prediction [12]. Koopman operator based probabilistic prediction frameworks adapt to complex uncertain maneuver conditions and achieve reliable long-term trajectory deduction [13]. Improved recurrent network multi-model structures effectively boost dynamic parameter identification efficiency for proportional navigation guidance systems [14]. Hybrid strategies combining model filtering and attention enhanced neural networks show outstanding performance in distinguishing lateral maneuver features of high-speed flight vehicles [15]. Physics-embedded Transformer models restrain cumulative prediction errors and maintain stable output even when facing incomplete sampling data [16]. Even so, data-driven methods heavily rely on finite flight test and numerical simulation samples, leading to poor generalization capacity across diverse combat missions and insufficient modeling effect on uncertain target behaviors under small sample conditions.

Inverse reinforcement learning breaks the restriction of manually predefined reward functions, and infers implicit reward mechanisms and decision strategies directly from actual flight trajectory

samples [17]. Ziebart et al. [17] pioneered maximum entropy based strategy distribution optimization, matching learned tactical characteristics with real combat behaviors and extracting motion tendency from incomplete trajectory information. IRL driven behavioral modeling simulates target decision-making logic and acquires high-precision prediction results under uncertain flight states [18]. Meta-learning assisted inverse reinforcement learning explores inherent combat correlation and improves cross-scenario generalization performance of prediction models [19]. Multi-agent reinforcement learning and adversarial imitation learning are also applied to battlefield situation assessment and autonomous tactical decision generation for unmanned combat platforms [20,21].

Inverse reinforcement learning possesses prominent advantages in depicting target confrontation strategies under complex combat backgrounds, showing broad application prospect in intelligent air combat decision-making. Within three-body active defense scenes, IRL-based prediction methods can effectively dig out latent attack intentions and operational strategies of incoming OAVs, so as to improve prediction accuracy based on the fixed strike purpose of hostile targets. Two core difficulties remain when introducing IRL into this research: first, constructing reasonable reward function parameterization form to support inversion of incoming vehicle combat strategies; second, combining historical observation sequences and learned strategy sets to realize online intention reasoning and dynamic trajectory prediction of maneuvering targets.

Accordingly, this paper proposes a trajectory prediction method based on Maximum Causal Entropy Inverse Reinforcement Learning, with three core innovations summarized as follows:

- Dynamic Strategy Modeling: Construct a behavioral function library for incoming OAV guidance strategies via maximum causal entropy IRL. This method maximizes strategic entropy while matching expected trajectory features, which not only accommodates suboptimal and tentative maneuver behaviors [22,23], but also ensures the objectivity and uniqueness of learned strategies.
- Uncertainty Quantification: Dynamically quantify strategic uncertainty based on entropy criteria and learn implicit distributions embedded in expert trajectory strategies, which significantly enhances the model's generalization ability toward diverse maneuver modes of non-cooperative targets.
- Real-Time Prediction Optimization: Realize adaptive weighting and dynamic optimization of trajectory prediction distributions by combining sliding window measurement data and strategy function weight calculation, so as to strengthen the dynamic adaptability of the prediction algorithm against target maneuver changes.

2. Problem Formulation of Incoming OAV Trajectory Prediction

The basic scenario of three-body active defense is illustrated in Figure 1. The seeker of the incoming target OAV has locked onto the aircraft and entered the terminal attack phase. Upon receiving lock-on warning information, the aircraft launches defensive interceptors to implement active defense.

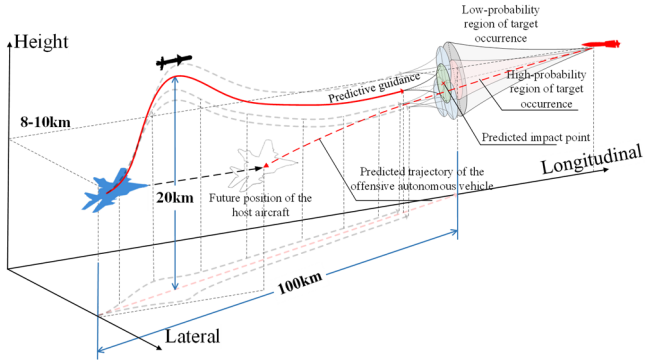


Figure 1 – Conceptual Diagram of the Three-Body Active Defense Scenario.

The involved operational processes and basic assumptions for research objects and scenarios are defined as follows:

- Assumptions for incoming OAVs: The OAV seeker maintains target lock status and attacks high-value single aircraft targets following a specific guidance law. All state vectors are

measured by airborne sensors with additive Gaussian white noise interference taken into consideration.

- Assumptions for aircraft: Once detecting incoming threats, the aircraft launches one interceptor for active defense, and subsequently executes evasive maneuvers under a fixed escape strategy. Meanwhile, its evasive maneuver strategy is transmitted and shared with the interceptor.
- Assumptions for interceptors: The interceptor receives the predicted trajectory sequence of the incoming OAV via aircraft-OAV data link for predictive guidance, and executes pre-emptive strike toward the predicted intercept points on the trajectory sequence.

In this scenario, the position and velocity vectors of the incoming OAV, aircraft and defensive interceptor are denoted as $\boldsymbol{r}_T$,$\boldsymbol{v}_T$, $\boldsymbol{r}_A$,$\boldsymbol{v}_A$ and $\boldsymbol{r}_D$,$\boldsymbol{v}_D$ respectively.

2.1 Dynamic Modeling of Incoming OAV

Neglecting the Earth's rotation, the particle dynamic equations of the incoming OAV in its passive flight phase are established in the local horizontal coordinate system as follows:

$$\begin{cases} \dot{\boldsymbol{r}}_T = \boldsymbol{v}_T \\ \dot{\boldsymbol{v}}_T = \boldsymbol{a}_T + \boldsymbol{g}\left(\boldsymbol{r}_T\right) \end{cases} \tag{1}$$

where $\boldsymbol{r}_T$ denotes the position vector of the incoming OAV in the local horizontal frame, $\boldsymbol{v}_T$ is its velocity vector, $\boldsymbol{a}_T$ represents the commanded acceleration vector, and $\boldsymbol{g}(r_T)$ is the gravitational acceleration related to the real-time position of the OAV. Maneuver features are extracted from the projected guidance acceleration commands in the semi-velocity coordinate system. The coordinate transformation relationship between accelerations in the local horizontal frame and semi-velocity frame is expressed as:

$$\boldsymbol{a}_T = \boldsymbol{T}_{\mathrm{HV}}^{\mathrm{G}} \boldsymbol{a}_T^{\mathrm{HV}} \tag{2}$$

in which $\boldsymbol{T}_{\mathrm{HV}}^{\mathrm{G}}$ stands for the coordinate transformation matrix formulated as:

$$\boldsymbol{T}_{\mathrm{HV}}^{\mathrm{G}} = \begin{bmatrix} v_x / v & -v_y / v_{\mathrm{g}} & -v_x v_z / (v v_{\mathrm{g}}) \\ v_y / v & v_x / v_{\mathrm{g}} & -v_y v_z / (v v_{\mathrm{g}}) \\ v_z / v & 0 & v_{\mathrm{g}} / v \end{bmatrix} \tag{3}$$

where v_x, v_y, v_z are three components of $\boldsymbol{v}_T$, v is the flight speed magnitude of the incoming OAV, and v_{g} is the magnitude of the horizontal projection of flight velocity.

To ensure that the learned behavioral strategy function can fully reflect the temporal variation law of relative states between the incoming OAV and aircraft, the state variables are defined as the relative position and velocity vectors of the aircraft with respect to the incoming OAV in the local horizontal frame:

$$\boldsymbol{x} = \left[x_A - x_T, y_A - y_T, z_A - z_T, v_{xA} - v_x, v_{yA} - v_y, v_{zA} - v_z \right]^{\mathrm{T}} \tag{4}$$

The control variables are selected as the projections of acceleration commands in the semi-velocity coordinate system:

$$\boldsymbol{u} = \boldsymbol{a}_T^{\mathrm{HV}} \tag{5}$$

Accordingly, the relative dynamic equations considering the motion of the aircraft relative to the incoming OAV are derived as:

$$\dot{\boldsymbol{x}} = \boldsymbol{f}\left[\boldsymbol{x}(t), \boldsymbol{u}(t)\right] = \begin{bmatrix} v_{xA} - v_{xT} \\ v_{yA} - v_{yT} \\ v_{zA} - v_{zT} \\ a_{xA} - \left(\mathbf{T}_{\mathrm{HV}}^{\mathrm{G}} \boldsymbol{a}_T^{\mathrm{HV}}\right)_1 \\ a_{yA} - \left(\mathbf{T}_{\mathrm{HV}}^{\mathrm{G}} \boldsymbol{a}_T^{\mathrm{HV}}\right)_2 \\ a_{zA} - \left(\mathbf{T}_{\mathrm{HV}}^{\mathrm{G}} \boldsymbol{a}_T^{\mathrm{HV}}\right)_3 \end{bmatrix} \tag{6}$$

where a_{xA}, a_{yA}, a_{zA} are determined by the preset programmed evasion sequence of the aircraft.

The observation variables of the incoming OAV, namely the relative position and velocity acquired

via data link, are written as:

$$\begin{aligned} z_k &= \mathbf{I}_6 \boldsymbol{x}_k + \boldsymbol{\nu}_k,\ \boldsymbol{\nu}_k \sim \mathcal{N}\left(\boldsymbol{\nu}_k; \mathbf{0}, \mathcal{R}_k\right) \\ &\triangleq \mathbf{H}_k \boldsymbol{x}_k + \boldsymbol{\nu}_k \end{aligned} \tag{7}$$

in which $\boldsymbol{\nu}_k$ denotes Gaussian measurement noise and $\mathcal{R}_k$ is the corresponding noise covariance matrix. Three predefined programmed maneuver modes are formulated as the evasion strategies of aircraft, as listed in Table 1.

Table 1 Escape strategies of aircraft

Maneuver Type	Maneuver Description
Constant-speed constant-altitude flight	Uniform straight flight at fixed altitude
S-shaped maneuver	S-shaped maneuver with constant overload magnitude
Tail-turn evasion	U-shaped descending maneuver with constant overload magnitude

2.2 Trajectory Prediction Model for Incoming OAVs

Six types of incoming OAV guidance laws (N=6, n=1,2,…,N) and three categories of aircraft escape strategies are adopted to construct simulation scenarios of three-body active defense. The expert trajectory dataset $\mathcal{D}^n$ for extracting maneuver behavioral features of incoming OAVs is established as:

$$\begin{cases} \mathcal{D}^n = \left\{\boldsymbol{\tau}_1^n, \boldsymbol{\tau}_2^n, \cdots, \boldsymbol{\tau}_J^n\right\} \\ \boldsymbol{\tau}_j^n = \left[[\boldsymbol{x}_1^j]^{\mathrm{T}}, [\boldsymbol{u}_1^j]^{\mathrm{T}}, \cdots, [\boldsymbol{x}_{t_f-1}^j]^{\mathrm{T}}, [\boldsymbol{u}_{t_f-1}^j]^{\mathrm{T}}, [\boldsymbol{x}_{t_f}^j]^{\mathrm{T}}, [\boldsymbol{u}_{t_f}^j]^{\mathrm{T}}\right]^{\mathrm{T}} \end{cases} \tag{8}$$

where $\boldsymbol{\tau}_j^n$ denotes the j-th trajectory generated under the n-th guidance law. The trajectory prediction problem of incoming OAVs is defined as follows: given prior target information $\mathcal{D}^n$ and partial observed trajectory segment $z^k = \left[z_0^{\mathrm{T}}, \cdots, z_{k-2}^{\mathrm{T}}, z_{k-1}^{\mathrm{T}}, z_k^{\mathrm{T}}\right]^{\mathrm{T}}$, predict the future trajectory sequence $\boldsymbol{\tau}_{\mathrm{pre}} = \left[[\boldsymbol{x}_k]^{\mathrm{T}}, [\boldsymbol{u}_k]^{\mathrm{T}}, \cdots, [\boldsymbol{x}_{t-1}]^{\mathrm{T}}, [\boldsymbol{u}_{t-1}]^{\mathrm{T}}, [\boldsymbol{x}_t]^{\mathrm{T}}, [\boldsymbol{u}_t]^{\mathrm{T}}\right]^{\mathrm{T}}$ from the current moment t_k to arbitrary time t. The proposed prediction model is formulated as:

$$\boldsymbol{f}_{\mathrm{pre}} : \left\{\mathcal{D}^n\right\}_{n=1}^N \times z^k \mapsto \boldsymbol{\tau}_{\mathrm{pre}} \tag{9}$$

in which $\mapsto$ represents the input-output mapping relationship of the prediction model. Taking the known aircraft evasive maneuver sequence $\boldsymbol{a}_A(t)$ as input, the predicted trajectory state variables are obtained by integrating the relative dynamic Equation (6):

$$\boldsymbol{x}(t) = \boldsymbol{x}(t_k) + \int_{t_k}^{t} \boldsymbol{f}\left[\boldsymbol{x}(\tau), \boldsymbol{u}(\tau); \boldsymbol{a}_A(\tau)\right] \mathrm{d}\tau \tag{10}$$

It is essential to determine the functional form of incoming OAV acceleration in Equation (5). The target acceleration is regarded as a function of relative state variables:

$$\boldsymbol{u} = \boldsymbol{u}\left[\boldsymbol{x}(t), \boldsymbol{\Theta}\right] \tag{11}$$

where $\boldsymbol{\Theta}$ denotes unknown parameters to be estimated from measured data z^k. Considering the uncertainty of guidance law types and relevant parameters, probability density functions are utilized to model maneuver characteristics of different guidance laws, thus constructing a model set of strategic functions $\mathcal{M} = \{\pi^{(n)}\}_{n=1}^N$. The strategic function corresponding to the n-th guidance law satisfies:

$$\begin{cases} \pi^{(n)} : \boldsymbol{u}_t \times \boldsymbol{x}_t \mapsto \mathbb{R}_{\geq 0} \\ \text{s.t.} \ \sum\limits_{\boldsymbol{u}_t} \pi^{(n)}\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right) = 1 \end{cases} \tag{12}$$

The weight coefficient set $\boldsymbol{\Theta} = \{w^{(n)}\}_{n=1}^N$ of measurement functions is updated via observed data z^k, and the current maneuver strategy of the incoming OAV is inferred by weighted fusion:

$$\pi(\boldsymbol{u}_t \mid \boldsymbol{x}_t;\boldsymbol{\Theta}) = \sum_{n=1}^{N} w^{(n)} \pi^{(n)}(\boldsymbol{u}_t \mid \boldsymbol{x}_t) \tag{13}$$

where $w^{(n)}$ is the weight coefficient corresponding to the n-th guidance law strategic function, satisfying the normalization constraint:

$$\sum_{n=1}^{N} w^{(n)} = 1 \tag{14}$$

Based on the inferred target strategic function $\pi(\boldsymbol{u}_t \mid \boldsymbol{x}_t;\boldsymbol{\Theta})$, combined with the current position state $\boldsymbol{x}_k$ of the incoming OAV and the relative dynamic Equation (6), the probability distribution function $p(\boldsymbol{x}(t))$of future target states can be derived by integrating the following Stochastic Differential Equation (SDE):

$$\begin{cases} \mathrm{d}\boldsymbol{x}(t) = \boldsymbol{f}\left[\boldsymbol{x}(t), \boldsymbol{u}(t)\right]\mathrm{d}t \\ \text{s.t.} \begin{cases} \boldsymbol{u}(t) \sim \pi(\boldsymbol{u}_t \mid \boldsymbol{x}_t;\boldsymbol{\Theta}) \\ \boldsymbol{x}(t_k) = \boldsymbol{x}_k \end{cases} \end{cases} \tag{15}$$

Eventually, the predicted trajectory sequence of the incoming OAV is acquired by extracting the mean value of the established prediction distribution.

3. Inverse Reinforcement Learning for Incoming OAVs Based on Maximum Causal Entropy

3.1 Overall Framework

The overall framework of incoming OAV trajectory prediction is illustrated in Figure 2, which mainly consists of two core modules: three-body active defense environment simulation, and behavioral strategy modeling as well as trajectory prediction based on maximum causal entropy inverse reinforcement learning.

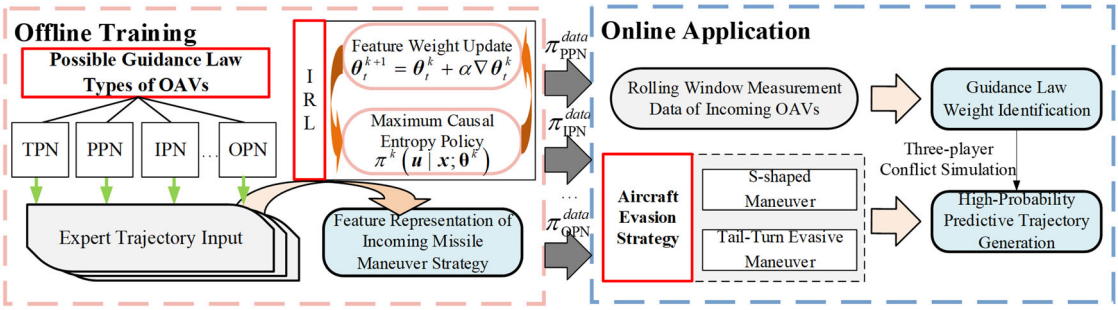


Figure 2 – Overall framework of incoming OAV trajectory prediction.

The specific implementation procedures are presented as follows:

- Three-body Active Defense Scenario Simulation: Based on sliding time window measurement data of incoming OAVs, a complete three-body active defense simulation environment is established. It integrates multiple typical incoming OAV guidance laws including True Proportional Navigation (TPN), Pure Proportional Navigation (PPN), Ideal Proportional Navigation (IPN) and Optimal Proportional Navigation (OPN), together with various aircraft evasion strategies such as S-shaped maneuver and tail-turn evasion. The interactive motion relationship among incoming OAVs, aircraft and interceptors is simulated to generate high-quality expert trajectory samples, which provide sufficient data foundation for subsequent behavioral strategy learning.

- Behavioral Strategy Modeling and Trajectory Prediction Based on Maximum Causal Entropy Inverse Reinforcement Learning: ① Guidance law behavioral strategy modeling of incoming OAVs. Within the inverse reinforcement learning framework, maximum causal entropy is taken as the optimization criterion, and the interpretability of incoming OAV maneuver behaviors is quantified into time-varying cost function parameters. ② Weight coefficient calculation of strategy functions and sliding window prediction architecture. Sliding windows are constructed for partial observed trajectory sequences. The weight coefficients are solved online in real time to realize adaptive weighted fusion of trajectory prediction distributions, and optimal prediction

results are finally selected and output, so as to effectively enhance the dynamic adaptability and anti-interference robustness against diverse target maneuver forms.

3.2 Maximum Causal Entropy Inverse Reinforcement Learning Formulation

This section formulates the behavioral strategy extraction problem of incoming OAVs based on maximum causal entropy inverse reinforcement learning, laying a foundation for subsequent weight coefficient calculation of behavioral strategies. Discrete inverse reinforcement learning is adopted to identify the guidance laws of incoming OAVs, and Equation (6) is discretized as:

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}_d\left(\boldsymbol{x}_t, \boldsymbol{u}_t\right) = \boldsymbol{x}_t + h\begin{bmatrix} \boldsymbol{x}_t^{4:6} \\ \boldsymbol{a}_{At} - \mathbf{T}_{\mathrm{HV}}^{\mathrm{G}}\boldsymbol{u}_t \end{bmatrix} \tag{16}$$

where h denotes the discrete time step, and $\boldsymbol{x}_t^{4:6}$ is the vector composed of the last three components of state variables.

The core idea of inverse reinforcement learning is to optimize the parameter vector θ_t of the reward function $r_t(\boldsymbol{x}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_t)$, so that the strategy function $\pi(\boldsymbol{u}_t \mid \boldsymbol{x}_t)$ derived from expert trajectories can be characterized as the optimal or suboptimal solution of optimal control problems. To measure the similarity between expert trajectories and strategy-generated trajectories, the trajectory feature function is defined as:

$$\boldsymbol{\phi}_t\left(\boldsymbol{x}_t, \boldsymbol{u}_t\right) : \boldsymbol{X}_t \times \boldsymbol{U}_t \mapsto \mathbb{R}^j \tag{17}$$

This function maps state and control variables into a j-dimensional vector, which describes artificially designed nonlinear trajectory features and is commonly used to construct the unknown reward function r_t hidden in expert trajectories. The core constraint of inverse reinforcement learning requires that the expectation values of trajectory feature functions are equal between strategy-generated trajectories and real expert trajectories, namely:

$$\mathop{\mathbb{E}}_{\pi}\left[\boldsymbol{\phi}_t(\boldsymbol{x}_t, \boldsymbol{u}_t)\right] = \mathop{\mathbb{E}}_{\mathbb{D}}\left[\boldsymbol{\phi}_t(\boldsymbol{x}_t, \boldsymbol{u}_t)\right] \ t = 0, 1, \cdots T \tag{18}$$

where $\mathbb{D}$ represents the expert trajectory set, and the trajectories generated by the strategy function are derived from the optimal strategy under the current reward function. When Equation (18) is satisfied and the reward function is linearly expressed by trajectory features $r_t(\boldsymbol{x}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_t) = -\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\phi}_t(\boldsymbol{x}_t, \boldsymbol{u}_t)$, the learned reward function $r_t(\boldsymbol{x}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_t)$ can well approximate the actual reward mechanism behind expert trajectories, which ensures that the learned target strategy is consistent with the real expert maneuver strategy.

Nevertheless, multiple different strategies can satisfy the feature matching constraint simultaneously, hence an additional selection criterion is necessary. From the perspective of maximum causal entropy inverse reinforcement learning, among all probability distributions satisfying expert trajectory feature matching, the distribution with the maximum causal entropy is optimal. Such distribution imposes the fewest extra prior assumptions and achieves the minimal predictive log loss. In particular, causal entropy further emphasizes chronological order for dynamic modeling of sequential behavioral maneuvers. Given the state-control trajectory sequence $\{\boldsymbol{x}_t, \boldsymbol{u}_t\}_{t=0,1,\cdots T}$ and policy $\pi_t(\boldsymbol{u}_t \mid \boldsymbol{x}_t)$, the causal entropy of policies along trajectory sequences is formulated as:

$$\begin{aligned} H\left(\boldsymbol{u}_{0:T-1} \,\|\, \boldsymbol{x}_{0:T-1}\right) &= \mathop{\mathbb{E}}_{(\boldsymbol{u}_{0:T-1}, \boldsymbol{x}_{0:T-1}) \sim \pi}\left[-\sum_{t=0}^{T-1} \ln \pi_t\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right)\right] \\ &= \sum_{t=0}^{T-1} \int_{\boldsymbol{u}_t} -\pi_t\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right) \ln \pi_t\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right) \mathrm{d}\boldsymbol{u}_t \\ &= \sum_{t=0}^{T-1} H\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right) \end{aligned} \tag{19}$$

Accordingly, the constrained optimization objective of maximum causal entropy inverse reinforcement learning for policy π is established as:

$$P:\begin{cases}\max\limits_{\pi} H\left(\boldsymbol{u}_{0:T-1} \| \boldsymbol{x}_{0:T-1}\right) \\ \text{s.t} \ \ \mathop{\mathbb{E}}\limits_{\pi}\left[\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]=\mathop{\mathbb{E}}\limits_{\mathbb{D}}\left[\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right] \ \ t=0,1,\cdots T \\ \text{且} \ \pi_t\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right)\geq 0, \ \sum\limits_{\boldsymbol{u}_t}\pi_t\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right)=1\end{cases} \tag{20}$$

The original optimization problem in Equation (20) belongs to infinite-dimensional constrained optimization in functional space, which brings great difficulties to numerical solution. Since the established problem is convex and satisfies strong duality, its dual problem shares the identical optimal solution with the primal problem. The dual transformation can convert the infinite-dimensional optimization into finite-dimensional optimization of Lagrange multipliers. The Lagrangian function is firstly constructed as:

$$\Lambda(\pi,\boldsymbol{\theta})=H\left(\boldsymbol{u}_{0:T-1} \| \boldsymbol{x}_{0:T-1}\right)+\left(\sum_{t=0}^{T-1}\mathop{\mathbb{E}}_{\mathbb{D}}\left[\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]-\mathop{\mathbb{E}}_{\pi}\left[\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]\right) \tag{21}$$

The corresponding dual optimization problem of Equation (20) is derived as:

$$\begin{aligned}&\overline{P}:\inf_{\boldsymbol{\theta}}\sup_{\pi}\Lambda(\pi,\boldsymbol{\theta}) \\ &=\inf_{\boldsymbol{\theta}}\left\{\sup_{\pi}\left(H\left(\boldsymbol{u}_{0:T-1} \| \boldsymbol{x}_{0:T-1}\right)-\mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{T-1}\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]\right)+\mathop{\mathbb{E}}_{\mathbb{D}}\left[\sum_{t=0}^{T-1}\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]\right\}\end{aligned} \tag{22}$$

Treating $\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)$ as the reward function $r_t(\boldsymbol{x}_t,\boldsymbol{u}_t)$, the optimal policy π^* corresponding to the supremum term $\sup_{\pi}\Lambda(\pi,\boldsymbol{\theta})$ in Equation (22) can be analytically solved via the Bellman equation of maximum causal entropy optimal control; the infimum term $\inf_{\theta}\Lambda\left(\pi^*,\boldsymbol{\theta}\right)$ can be solved by gradient descent to update Lagrange multipliers. Alternating iterative updating of the above two procedures enables accurate extraction of reward function parameters $\boldsymbol{\theta}_t$ that can fully characterize the maneuver behavioral features of incoming OAV expert trajectories.

3.3 Maximum Causal Entropy Policy Calculation and Behavioral Characteristic Parameter Optimization

By introducing a negative sign, the supremum term $\sup_{\pi}\Lambda(\pi,\boldsymbol{\theta})$ in Equation (22) can be converted into a maximum causal entropy optimal control problem, which essentially aims to solve the stochastic control policy $\pi_t^*\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right)$.

$$J\left(\pi;\boldsymbol{x}_0\right)=\mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{T}l_t^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]-H\left(\boldsymbol{u}_{0:T-1} \| \boldsymbol{x}_{0:T-1}\right) \tag{23}$$

where l_t^{cost} denotes the cost index function. Under linear feature representation, it is defined as the inner product of behavioral characteristic parameters $\boldsymbol{\theta}_t$ and trajectory feature function $\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)$:

$$\begin{cases}l_t^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t) \triangleq \boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)=-r_t(\boldsymbol{x}_t,\boldsymbol{u}_t) & t\leq T-1 \\ l_T^{\mathrm{cost}}(\boldsymbol{x}_T,\boldsymbol{u}_T)\triangleq\boldsymbol{\theta}_T^{\mathrm{T}}\boldsymbol{\phi}_T(\boldsymbol{x}_T) =-r_T(\boldsymbol{x}_T,\boldsymbol{u}_T) & t=T\end{cases} \tag{24}$$

The policy function satisfies $\pi_t\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right)\geq 0, \ \sum_{\boldsymbol{u}_t}\pi_t\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right)=1$. Define the value function $V_t\left(\boldsymbol{x}_t\right)\triangleq\min_{\pi}J\left(\pi;\boldsymbol{x}_t\right)$. The Bellman equation derived from the optimality principle is written as:

$$V_t\left(\boldsymbol{x}_t\right)=\min_{\pi}\left\{\mathop{\mathbb{E}}_{\pi}\left[l_t^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)+V_{t+1}\left(\boldsymbol{x}_{t+1}\right)\right]-H\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right)\right\} \tag{25}$$

Solving Equation (25) via the first-order optimality condition yields the optimal stochastic policy:

$$\begin{aligned}\pi_t\left(\boldsymbol{u}_t \mid \boldsymbol{x}_t\right)&=\frac{\exp\left(-\left[l_t^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)+V_{t+1}\left(\boldsymbol{x}_{t+1}\right)\right]\right)}{\int_{\boldsymbol{u}_t}\exp\left(-\left[l_t^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)+V_{t+1}\left(\boldsymbol{x}_{t+1}\right)\right]\right)\mathrm{d}\boldsymbol{u}_t} \\ &\triangleq Z_t^{-1}\left(\boldsymbol{x}_t\right)\exp\left(-\left[l_t^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)+V_{t+1}\left(\boldsymbol{x}_{t+1}\right)\right]\right)\end{aligned} \tag{26}$$

Substitute Equation (26) into Equation (25) as $Z_t\left(\boldsymbol{x}_t\right)=\exp\left(-V_t\left(\boldsymbol{x}_t\right)\right)$ and sort out the soft iterative

updating rule of the value function:

$$\begin{cases} Q_t(\boldsymbol{x}_t,\boldsymbol{u}_t)=l_t^{\text{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)+V_{t+1}(\boldsymbol{x}_{t+1}) & t\le T-1 \\ V_t(\boldsymbol{x}_t)=\begin{cases} -\ln\int_{\boldsymbol{u}_t}\exp\left[-Q_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]\mathrm{d}\boldsymbol{u}_t & t\le T-1 \\ l_T^{\text{cost}}(\boldsymbol{x}_T) & t=T \end{cases} \end{cases} \tag{27}$$

The term *soft* indicates that the traditional hard maximum operation $V_t(\boldsymbol{x}_t)=\max_{\boldsymbol{u}_t}Q_t(\boldsymbol{x}_t,\boldsymbol{u}_t)$ in value iteration is replaced by exponential transformation, summation and logarithmic operation, which fully takes into account suboptimal action values. Combining Equation (27) and Equation (26), the maximum causal entropy optimal policy is expressed as:

$$\pi_t^*(\boldsymbol{u}_t\mid\boldsymbol{x}_t)=\exp\left(-\left[Q_t(\boldsymbol{x}_t,\boldsymbol{u}_t)-V_t(\boldsymbol{x}_t)\right]\right) \tag{28}$$

Equation (28) is the explicit form of the policy function corresponding to the supremum in the dual problem (22). On this basis, the gradient of the infimum optimization problem $\inf_\theta \Lambda\left(\pi^*,\theta\right)$ with respect to behavioral characteristic parameter θ_t is derived as:

$$\nabla_{\theta_t}\Lambda\left(\pi^*,\boldsymbol{\theta}\right)=\mathop{\mathbb{E}}_{\pi^*}\left[\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]-\mathop{\mathbb{E}}_{\mathbb{D}}\left[\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right] \tag{29}$$

where $\mathbb{E}_{\pi^*}\left[\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]$ is the mean feature value of trajectories generated by the optimal policy under current θ_t, and $\mathbb{E}_{\mathbb{D}}\left[\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\right]$ is the mean feature value of real expert trajectories. The learned trajectory features can gradually approximate expert trajectory features by adjusting θ_t via gradient descent. The iterative updating formula for behavioral characteristic parameters is:

$$\boldsymbol{\theta}_t^{k+1}=\boldsymbol{\theta}_t^k-\eta_t\nabla_{\theta_t}\Lambda\left(\pi^*,\boldsymbol{\theta}\right) \tag{30}$$

where η_t denotes the descent step size. In this paper, line search based on Wolfe criterion and BFGS quasi-Newton algorithm are adopted to realize efficient and stable optimization of behavioral characteristic parameter θ_t.

3.4 Quadratic Form Based Strategy Function Calculation for Incoming OAVs

The calculation of the denominator term $Z(\boldsymbol{x}_t)$ in Equation (26) requires full integration over the domain of control variables. Affected by the strong nonlinearity of incoming OAV dynamics, this procedure suffers from excessive computational complexity when dealing with high-dimensional state variables. To address this issue, the value function *V* in Equation (27) is formulated in quadratic form:

$$V_t(\boldsymbol{x}_t)=\frac{1}{2}\begin{bmatrix}1\\ \boldsymbol{x}_t\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2V_{c,t}+2V_{H,t} & \mathbf{V}_{x,t}\\ \mathbf{V}_{x,t}^{\mathrm{T}} & \mathbf{V}_{xx,t}\end{bmatrix}\begin{bmatrix}1\\ \boldsymbol{x}_t\end{bmatrix} \tag{31}$$

and the action-value function *Q* is expressed as:

$$Q_t(\boldsymbol{x}_t,\boldsymbol{u}_t)=\frac{1}{2}\begin{bmatrix}1\\ \boldsymbol{x}_t\\ \boldsymbol{u}_t\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2Q_{c,t} & \mathbf{Q}_{x,t}^{\mathrm{T}} & \mathbf{Q}_{u,t}^{\mathrm{T}}\\ \mathbf{Q}_{x,t} & \mathbf{Q}_{xx,t} & \mathbf{Q}_{xu,t}\\ \mathbf{Q}_{u,t} & \mathbf{Q}_{ux,t} & \mathbf{Q}_{uu,t}\end{bmatrix}\begin{bmatrix}1\\ \boldsymbol{x}_t\\ \boldsymbol{u}_t\end{bmatrix} \tag{32}$$

Such formulation enables the integration and iteration in Equation (27) to adopt analytical matrix recursion forms. Inspired by the complexity reduction mechanism of iterative linear-quadratic regulator (iLQR) in indirect trajectory optimization, the dynamic Equation (16) is linearized via first-order Taylor expansion along expert trajectory $\{\bar{\boldsymbol{x}}_t,\bar{\boldsymbol{u}}_t\}_{t=0,1,\cdots T}$:

$$\begin{aligned}\boldsymbol{x}_{t+1}&=\bar{\boldsymbol{x}}_{t+1}+\boldsymbol{f}_{dx}\left(\bar{\boldsymbol{x}}_t,\bar{\boldsymbol{u}}_t\right)\left(\boldsymbol{x}_t-\bar{\boldsymbol{x}}_t\right)+\boldsymbol{f}_{du}\left(\bar{\boldsymbol{x}}_t,\bar{\boldsymbol{u}}_t\right)\left(\boldsymbol{u}_t-\bar{\boldsymbol{u}}_t\right)\\ &\triangleq\bar{\boldsymbol{\xi}}_t+\boldsymbol{f}_{dxt}\boldsymbol{x}_t+\boldsymbol{f}_{dut}\boldsymbol{u}_t\end{aligned} \tag{33}$$

where $\bar{\boldsymbol{\xi}}_t=\bar{\boldsymbol{x}}_{t+1}-\boldsymbol{f}_{dxt}\bar{\boldsymbol{x}}_t-\boldsymbol{f}_{dut}\bar{\boldsymbol{u}}_t$

$$\boldsymbol{f}_{dxt}=\begin{bmatrix}\boldsymbol{I}_3 & h\boldsymbol{I}_3\\ \mathbf{0}_{3\times3} & \boldsymbol{I}_3\end{bmatrix},\boldsymbol{f}_{dut}=\begin{bmatrix}\mathbf{0}_{3\times3}\\ -h\boldsymbol{T}_{\mathrm{HV}}^{\mathrm{G}}\end{bmatrix} \tag{34}$$

and *h* denotes the discrete time step. The trajectory feature function defined in Equation (17) is extended to matrix form $\boldsymbol{\phi}_t \in \mathbb{R}^{i\times j}$, so that the trajectory features of incoming OAVs are modeled by linear and quadratic terms of trajectory sequences:

$$\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)=\begin{cases}\begin{bmatrix}1 & \boldsymbol{x}_t^{\mathrm{T}} & \boldsymbol{u}_t^{\mathrm{T}}\\ \boldsymbol{x}_t & \boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}} & \boldsymbol{x}_t\boldsymbol{u}_t^{\mathrm{T}}\\ \boldsymbol{u}_t & \boldsymbol{u}_t\boldsymbol{x}_t^{\mathrm{T}} & \boldsymbol{u}_t\boldsymbol{u}_t^{\mathrm{T}}\end{bmatrix} & t\le T-1\\ \begin{bmatrix}1 & \boldsymbol{x}_T^{\mathrm{T}}\\ \boldsymbol{x}_T & \boldsymbol{x}_T\boldsymbol{x}_T^{\mathrm{T}}\end{bmatrix} & t=T\end{cases} \tag{35}$$

Meanwhile, the cost index function is redefined as $l_t^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)=\langle\boldsymbol{\theta}_t,\boldsymbol{\phi}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)\rangle$, in which $\langle\cdot,\cdot\rangle$ represents matrix inner product. The equivalent quadratic expression of the instantaneous cost function is obtained as:

$$\begin{cases}l_t^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)=\dfrac{1}{2}\begin{bmatrix}1\\ \boldsymbol{x}_t\\ \boldsymbol{u}_t\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2n_t & \boldsymbol{q}_t^{\mathrm{T}} & \boldsymbol{r}_t^{\mathrm{T}}\\ \boldsymbol{q}_t & \mathbf{Q}_t & \mathbf{M}_t\\ \boldsymbol{r}_t & \mathbf{M}_t^{\mathrm{T}} & \mathbf{R}_t\end{bmatrix}\begin{bmatrix}1\\ \boldsymbol{x}_t\\ \boldsymbol{u}_t\end{bmatrix} & t\le T-1\\ l_T^{\mathrm{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)=\dfrac{1}{2}\begin{bmatrix}1\\ \boldsymbol{x}_T\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2n_T & \boldsymbol{q}_T^{\mathrm{T}}\\ \boldsymbol{q}_T & \mathbf{Q}_T\end{bmatrix}\begin{bmatrix}1\\ \boldsymbol{x}_T\end{bmatrix} & t=T\end{cases} \tag{36}$$

Substituting Equation (33) and Equation (36) into the first part of the soft value iteration in Equation (27), all quadratic matrix components of the action-value function *Q* can be derived:

$$\begin{cases}\begin{bmatrix}\mathbf{Q}_{xx,t} & \mathbf{Q}_{xu,t}\\ \mathbf{Q}_{ux,t} & \mathbf{Q}_{uu,t}\end{bmatrix}=\begin{bmatrix}\boldsymbol{f}_{dxt}^{\mathrm{T}}\\ \boldsymbol{f}_{dut}^{\mathrm{T}}\end{bmatrix}\mathbf{V}_{xx,t+1}\begin{bmatrix}\boldsymbol{f}_{dxt}^{\mathrm{T}}\\ \boldsymbol{f}_{dut}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}+\begin{bmatrix}\mathbf{Q}_t & \mathbf{M}_t\\ \mathbf{M}_t^{\mathrm{T}} & \mathbf{R}_t\end{bmatrix}\\ \begin{bmatrix}\mathbf{Q}_{x,t}\\ \mathbf{Q}_{u,t}\end{bmatrix}=\begin{bmatrix}\boldsymbol{f}_{dxt}^{\mathrm{T}}\\ \boldsymbol{f}_{dut}^{\mathrm{T}}\end{bmatrix}\left(\mathbf{V}_{xx,t+1}\bar{\boldsymbol{\xi}}_t+\mathbf{V}_{x,t+1}\right)+\begin{bmatrix}\boldsymbol{q}_t\\ \boldsymbol{r}_t\end{bmatrix}\\ Q_{c,t}=0.5\bar{\boldsymbol{\xi}}_t^{\mathrm{T}}\mathbf{V}_{xx,t+1}\bar{\boldsymbol{\xi}}_t+\bar{\boldsymbol{\xi}}_t^{\mathrm{T}}\mathbf{V}_{x,t+1}+V_{c,t+1}+V_{H,t+1}+n_t\end{cases} \tag{37}$$

For the integral term in the second part of Equation (27), analytical solution is realized via Gaussian variable completion:

$$V_t(\boldsymbol{x}_t)=\frac{1}{2}\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{Q}_{xx,t}\boldsymbol{x}_t+\boldsymbol{Q}_{x,t}^{\mathrm{T}}\boldsymbol{x}_t+Q_{c,t}-\frac{1}{2}\left(\boldsymbol{u}_t^*\right)^{\mathrm{T}}\boldsymbol{Q}_{uu,t}\boldsymbol{u}_t^*-\ln\left((2\pi)^{n_u/2}\left|\boldsymbol{Q}_{uu,t}^{-1}\right|^{1/2}\right) \tag{38}$$

Thereby, the quadratic parameters of value function in Equation (31) satisfy the backward recursion for *t*<*T*:

$$\begin{cases}\boldsymbol{V}_{c,t}=Q_{c,t}-0.5\boldsymbol{k}_t^{\mathrm{T}}\boldsymbol{Q}_{uu,t}\boldsymbol{k}_t\\ \boldsymbol{V}_{H,t}=0.5\left[\ln\left|\boldsymbol{Q}_{uu,t}\right|-n_u\ln(2\pi)\right]\\ \boldsymbol{V}_{x,t}=\boldsymbol{Q}_{x,t}-\boldsymbol{K}_t^{\mathrm{T}}\boldsymbol{Q}_{uu,t}\boldsymbol{k}_t\\ \boldsymbol{V}_{xx,t}=\boldsymbol{Q}_{xx,t}-\boldsymbol{K}_t^{\mathrm{T}}\boldsymbol{Q}_{uu,t}\boldsymbol{K}_t\end{cases} \tag{39}$$

and the terminal condition at *t*=*T* is:

$$\boldsymbol{V}_{xx,T}=\boldsymbol{Q}_T,\boldsymbol{V}_{x,T}=\boldsymbol{q}_T,V_{c,T}=n_T,V_{H,T}=0 \tag{40}$$

The exponential term in Equation (28) is further sorted out as:

$$-\left[Q_t(\boldsymbol{x}_t,\boldsymbol{u}_t)-V_t(\boldsymbol{x}_t)\right]=-\frac{1}{2}\left(\boldsymbol{u}_t-\boldsymbol{u}_t^*\right)^{\mathrm{T}}\left(\boldsymbol{Q}_{uu,t}^{-1}\right)^{-1}\left(\boldsymbol{u}_t-\boldsymbol{u}_t^*\right)-\ln\left((2\pi)^{n_u/2}\left|\boldsymbol{Q}_{uu,t}^{-1}\right|^{1/2}\right) \tag{41}$$

where $\boldsymbol{u}_t^*$ denotes the deterministic optimal control law:

$$\begin{aligned}\boldsymbol{u}_t^*&=-\boldsymbol{Q}_{uu,t}^{-1}\left(\boldsymbol{Q}_{ux,t}\boldsymbol{x}_t+\boldsymbol{Q}_{u,t}\right)\\ &\triangleq\boldsymbol{K}_t\boldsymbol{x}_t+\boldsymbol{k}_t\end{aligned} \tag{42}$$

Consequently, the policy function in Equation (28) obeys Gaussian distribution:

$$\pi_t(\boldsymbol{u}_t \mid \boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{u}_t; \boldsymbol{K}_t \boldsymbol{x}_t + \boldsymbol{k}_t, \boldsymbol{Q}_{uu,t}\right) \triangleq \mathcal{N}\left(\boldsymbol{u}_t; \boldsymbol{u}_t^*, \boldsymbol{Q}_{uu,t}\right) \tag{43}$$

In summary, by adopting quadratic-form value function *V* and action-value function *Q*, the soft value iteration process in Equation (27) is transformed into backward recursive calculation of quadratic matrix terms as shown in Equation (37) and Equation (39). This method greatly reduces the solving difficulty of maximum causal entropy optimal policy, and is well applicable to high-dimensional dynamic systems in air combat scenarios.

4. Dynamic Trajectory Sequence Prediction Based on Probability Weighting

So far, a model set $\mathcal{M}$ corresponding to different guidance laws of incoming OAVs can be established via inverse reinforcement learning with quadratic trajectory features. To infer the unknown target strategy π using actual measured data z^k of incoming OAVs, this section proposes a strategy function based on weighted fusion of multiple model strategies. The initial weights are firstly determined by the value functions of all strategies in the model set, and then the weights are dynamically updated utilizing data within the sliding time window.

4.1 Initial Weight Calculation of Target Strategy

Considering the dynamic influence of incoming OAV maneuvers on weight solution, the current maneuver strategy of the incoming OAV is inferred via weighted fusion of strategies in the model set, as shown in Equation (13). The initial weight coefficients $w_{t|0}^{(n)}$ are calculated from the value function $V_0^{(n)}$ of each strategy function in the model set. Firstly, the partition function of each strategy function within the model set is defined as:

$$Z_0^{(n)}(\boldsymbol{x}_0) = \exp\left(-V_0^{(n)}(\boldsymbol{x}_0)\right) \tag{44}$$

where $\boldsymbol{x}_0$ denotes the OAV-target relative state at the initial measurement instant. Let the partition function of the weighted mixed strategy at time *t+1* satisfy:

$$Z_{t+1}(\boldsymbol{x}_{t+1}) = \sum_{n=1}^{N} Z_{t+1}^{(n)}(\boldsymbol{x}_{t+1}) \tag{45}$$

Combined with Equation (26), it yields:

$$\begin{aligned} Z_t(\boldsymbol{x}_t) &= \sum_{n=1}^{N} \int_{\boldsymbol{u}_t} \exp\left(-l_t^{\text{cost}}(\boldsymbol{x}_t, \boldsymbol{u}_t)\right) Z_{t+1}^{(n)}(\boldsymbol{x}_{t+1}) \mathrm{d}\boldsymbol{u}_t \\ &= \sum_{n=1}^{N} Z_t^{(n)}(\boldsymbol{x}_t) \end{aligned} \tag{46}$$

It is assumed that each single-strategy policy $\pi(\boldsymbol{u}_t \mid \boldsymbol{x}_t)$ is derived from the maximum causal entropy optimal control problem and conforms to Equation (26). Expanding the mixed strategy formula Equation (13) obtains:

$$\begin{aligned} \pi_t(\boldsymbol{u}_t \mid \boldsymbol{x}_t) &= \frac{Z_{t+1}(\boldsymbol{x}_{t+1}) \exp\left(-l_t^{\text{cost}}(\boldsymbol{x}_t, \boldsymbol{u}_t)\right)}{Z_t(\boldsymbol{x}_t)} \\ &= \sum_{n=1}^{N} w_{t|0}^{(n)}(\boldsymbol{x}_t) \frac{Z_{t+1}^{(n)}(\boldsymbol{x}_{t+1}) \exp\left(-l_t^{\text{cost},(n)}(\boldsymbol{x}_t, \boldsymbol{u}_t)\right)}{Z_t^{(n)}(\boldsymbol{x}_t)} \end{aligned} \tag{47}$$

The initial weight coefficient $w_{t|0}^{(n)}(\boldsymbol{x}_t)$ is set as follows:

$$w_{t|0}^{(n)}(\boldsymbol{x}_t) = \frac{Z_t^{(n)}(\boldsymbol{x}_t)}{Z_t(\boldsymbol{x}_t)} \tag{48}$$

Under this setting, the cost function corresponding to the mixed strategy $\pi(\boldsymbol{u}_t \mid \boldsymbol{x}_t)$ is equivalent to the weighted combination of future residual costs of all individual strategies in the exponential domain. This indicates that the incoming OAV tends to adopt maneuver strategies with lower residual cost, which can be expressed as:

$$\exp\left(-l_t^{\text{cost}}(\boldsymbol{x}_t,\boldsymbol{u}_t)\right)=\sum_{n=1}^{N}w_{t+1|0}^{(n)}(\boldsymbol{x}_{t+1})\exp\left(-l_t^{\text{cost},(n)}(\boldsymbol{x}_t,\boldsymbol{u}_t)\right) \tag{49}$$

4.2 Probabilistic Weighting of Target Strategy Function Based on Observation Sequences

The dynamic weight calculation takes *w* groups of measured data $z_w^k=\left[\boldsymbol{z}_{k-w+1}^{\mathrm{T}},\cdots,\boldsymbol{z}_{k-p+1}^{\mathrm{T}},\boldsymbol{z}_{k-p+2}^{\mathrm{T}},\cdots,\boldsymbol{z}_k^{\mathrm{T}}\right]^{\mathrm{T}}$ ranging from *t=k−w+1* to the current moment *t=k* as inputs, and outputs the weight coefficients $\{w_k^{(n)}\}_{n=1}^{N}$ corresponding to each guidance strategy in the model set.

Let event M_k^n denote that the incoming OAV adopts the n-th guidance law at current time *t=k*. The initial weight of each guidance law calculated by Equation (48) is expressed as:

$$P\left\{M_k^n\right\}=w_{k|0}^{(n)} \tag{50}$$

In accordance with Bayesian theorem, the updated weight coefficients to be solved are derived as:

$$w_k^{(n)}\triangleq P\left\{M_k^n\mid z_w^k\right\}\propto p\left(z_w^k\mid M_k^n\right)P\left\{M_k^n\right\} \tag{51}$$

The observation likelihood $p\left(z_w^k\mid M_k^n\right)$ is decomposed into sequential conditional likelihood terms:

$$\begin{aligned}L_k^n&\triangleq p\left(z_w^k\mid M_k^n\right)\\&=p\left(\boldsymbol{z}_k\mid\boldsymbol{z}_{k-w+1:k-1},M_k^n\right)p\left(\boldsymbol{z}_{k-w+1:k-1}\mid M_k^n\right)\\&\triangleq l_{k-1}^nL_{k-1}^n\end{aligned} \tag{52}$$

where the single-step observation likelihood l_{k-1}^n satisfies:

$$\begin{aligned}l_{k-1}^n&\triangleq p\left(\boldsymbol{z}_k\mid\boldsymbol{z}_{k-w+1:k-1},M_k^n\right)\\&=\sum_{\boldsymbol{x}_k}\left[p\left(\boldsymbol{z}_k\mid\boldsymbol{x}_k\right)p\left(\boldsymbol{x}_k\mid\boldsymbol{z}_{k-w+1:k-1},M_k^n\right)\right]\end{aligned} \tag{53}$$

The first term on the right-hand side of Equation (53) is determined by the observation equation (7):

$$p\left(\boldsymbol{z}_k\mid\boldsymbol{x}_k\right)=\mathcal{N}\left(\boldsymbol{z}_k;\boldsymbol{H}_k\boldsymbol{x}_k,\mathcal{R}_k\right) \tag{54}$$

The second term is obtained via one-step state prediction:

$$p\left(\boldsymbol{x}_k\mid\boldsymbol{z}_{k-w+1:k-1},M_k^n\right)\approx\mathcal{N}\left(\boldsymbol{x}_k;\hat{\boldsymbol{x}}_{k|k-1}^{(n)},\hat{\boldsymbol{P}}_{k|k-1}^{(n)}\right) \tag{55}$$

The mean value and covariance matrix of the predicted state are calculated as:

$$\begin{cases}\hat{\boldsymbol{x}}_{k|k-1}^{(n)}=\boldsymbol{f}_d\left(\hat{\boldsymbol{x}}_{k-1}^{(n)},\boldsymbol{K}_{k-1}^{(n)}\hat{\boldsymbol{x}}_{k-1}^{(n)}+\boldsymbol{k}_{k-1}^{(n)}\right)\\\hat{\boldsymbol{P}}_{k|k-1}^{(n)}=\boldsymbol{F}_{k-1}^{(n)}\hat{\boldsymbol{P}}_{k-1}^{(n)}\boldsymbol{F}_{k-1}^{(n)\mathrm{T}}+\boldsymbol{B}_{k-1}^{(n)}\boldsymbol{Q}_{uu,k-1}^{(n)}\boldsymbol{B}_{k-1}^{(n)\mathrm{T}}\\\boldsymbol{F}_{k-1}^{(n)}=\boldsymbol{A}_{k-1}^{(n)}+\boldsymbol{B}_{k-1}^{(n)}\boldsymbol{K}_{k-1}^{(n)}\end{cases} \tag{56}$$

in which $\boldsymbol{A}_{k-1}^{(n)},\boldsymbol{B}_{k-1}^{(n)}$ represents the Jacobian matrix of the discrete dynamic equation (16) with respect to state and control variables at moment *k−1*. Based on the properties of Gaussian distribution, Equation (53) can be further simplified as:

$$p\left(\boldsymbol{z}_k\mid\boldsymbol{z}_{k-w+1:k-1},M_k^n\right)=\mathcal{N}\left(\boldsymbol{z}_k;\boldsymbol{H}_k\hat{\boldsymbol{x}}_{k|k-1}^{(n)},\boldsymbol{H}_k\hat{\boldsymbol{P}}_{k|k-1}^{(n)}\boldsymbol{H}_k^{\mathrm{T}}+\mathcal{R}_k\right) \tag{57}$$

The decomposition formula (52) is recursively solved in reverse order within the sliding time window, and define the recursive likelihood term as $L_k^n=l_{k-1}^nl_{k-2}^n\cdots L_{k-w+1}^n$. Define

$$L_{k-w+1}^n=1 \tag{58}$$

Finally, trajectory prediction distributions are constructed in three different forms, and the target predicted trajectory sequences are obtained by extracting the mean values of these distributions. The three types of prediction distributions are specified as follows:

- Trajectory prediction distribution under individual modes.

$$p\left(\boldsymbol{x}_{(k-w+1):(k+\Delta T_{\max})}\mid z_w^k,M_k^n\right) \tag{59}$$

- Mode-weighted fused trajectory prediction distribution.

$$p\left(\boldsymbol{x}_{(k-w+1):(k+\Delta T_{\max})} \mid \boldsymbol{z}_w^k\right)=\sum_{n=1}^{N}\left[w_k^{(n)} p\left(\boldsymbol{x}_{k-w+1:\Delta T_{\max}} \mid \boldsymbol{z}_w^k, M_k^n\right)\right] \tag{60}$$

- Strategy-weighted trajectory prediction distribution.

$$\begin{cases}\pi\left(\boldsymbol{u}_i \mid \boldsymbol{x}_i\right)=\sum_{n=1}^{N} w_k^{(n)} \pi^{(n)}\left(\boldsymbol{u}_i \mid \boldsymbol{x}_i\right) \\ p\left(\boldsymbol{x}_{(k-w+1):(k+\Delta T_{\max})} \mid \boldsymbol{z}_w^k\right)=\prod_{i=k-w+1}^{k+\Delta T_{\max}} p\left(\boldsymbol{x}_{i+1} \mid \boldsymbol{x}_i, \boldsymbol{u}_i\right) \pi\left(\boldsymbol{u}_i \mid \boldsymbol{x}_i\right) p\left(\boldsymbol{x}_{k-w+1} \mid \boldsymbol{z}_w^k\right)\end{cases} \tag{61}$$

To obtain high-precision target predicted trajectories, the mean values of all trajectory prediction distributions are taken as candidate predicted sequences. The prediction errors are calculated within the sliding data window, and the sequence with the minimum prediction error is determined as the final output of the trajectory prediction module. The standardized root mean square error (RMSE) is adopted to quantitatively evaluate the trajectory prediction accuracy within the time window, which is defined as:

$$J=\sqrt{\frac{1}{w} \sum_{i=k-w+1}^{k}\left\|\boldsymbol{z}_i-\boldsymbol{h}_i\left(\boldsymbol{x}_i\right)\right\|_{\boldsymbol{R}_i^{-1}}^2} \tag{62}$$

5. Simulation Verification

Both the learning of expert trajectories of incoming OAVs and the verification of trajectory prediction algorithms rely on the establishment of the three-body active defense simulation scenario shown in Figure 1. The typical three-body active defense combat scenario is defined as Working Condition 1. In this scenario, the initial distance between the incoming OAV and the aircraft is 200 km, with the OAV flying at an altitude of 10 km and a speed of 1.5 Mach towards the aircraft. The initial radial, altitude and lateral positions as well as trajectory inclination and deflection angles of incoming OAVs are sampled following uniform distribution to simulate combat scenarios under diverse initial states.

The full flight trajectory of the incoming OAV is divided into the powered phase and unpowered phase. In the powered phase, the OAV climbs and accelerates under programmed angle-of-attack commands. The engine works for 11 seconds with a thrust of 100 kN, and the terminal flight speed reaches 8 Mach. In the unpowered phase, the OAV attacks the aircraft adopting proportional navigation series guidance laws, with guidance constants set as 3, 4 and 5 respectively. The maximum available angle of attack is 35°, and the maximum longitudinal and lateral overload limit is 50*g*. The detailed initial parameters are listed in Table 2.

Table 2 Initial Conditions of Incoming OAV

Initial State	Value
Initial OAV-target distance / km	200
Initial flight altitude / km	10
Terminal speed of powered phase	8 Mach
Dispersion of radial/altitude/lateral position / km	±20 / ±2 / ±20
Dispersion of trajectory inclination / deflection angle / (°)	±3 / ±20

The aircraft keeps an initial altitude of 10 km and flies head-on against the incoming OAV at a speed of 2 Mach. Once detecting the launch of incoming OAVs, the aircraft immediately switches to evasive maneuver mode. Three optional evasive strategies are available, namely uniform straight flight, S-shaped maneuver with 10g overload, and tail-turn evasive maneuver with 2 km altitude drop.

Restricted by onboard computing resources, the proposed trajectory prediction algorithm is designed to run on airborne computers. During algorithm execution, the aircraft cannot acquire the exact guidance law adopted by incoming OAVs, yet it is confirmed to belong to the established model set $\mathcal{M}$. The evasive maneuver of the aircraft follows preset programmed motions. The airborne sensors continuously acquire the position and velocity observation sequences of incoming OAVs, and the predicted trajectories are transmitted to interceptors in real time via data link. Combined with radar-infrared fused measurement and target tracking results based on constant velocity (CV) motion

model, pseudo observation conditions for the relative states z_k between incoming OAVs and aircraft are constructed with a sampling interval of 50 ms, as shown in Table 3.

Table 3 Observation conditions of aircraft

Observed Quantity	Noise Level (1σ Standard Deviation)
Relative longitudinal range / m	10
Relative altitude / m	8
Relative lateral range / m	8
Relative longitudinal velocity / (m·s⁻¹)	1.0
Relative vertical velocity / (m·s⁻¹)	0.5
Relative lateral velocity / (m·s⁻¹)	0.5

5.1 Learning and Training Process of Behavioral Characteristic Parameters

A total of 600 expert trajectories are generated for each proportional navigation-based guidance law in the model set $\mathcal{M}$, as illustrated in Fig. 3. Among them, 200 trajectories correspond to each of the three aircraft escape strategies. Taking the learning of the PPN (Pure Proportional Navigation) strategy function as an example, 450 out of the 600 trajectories are used as the training set for the PPN strategy function, and the remaining 150 trajectories are applied to verify the trajectory prediction algorithm.

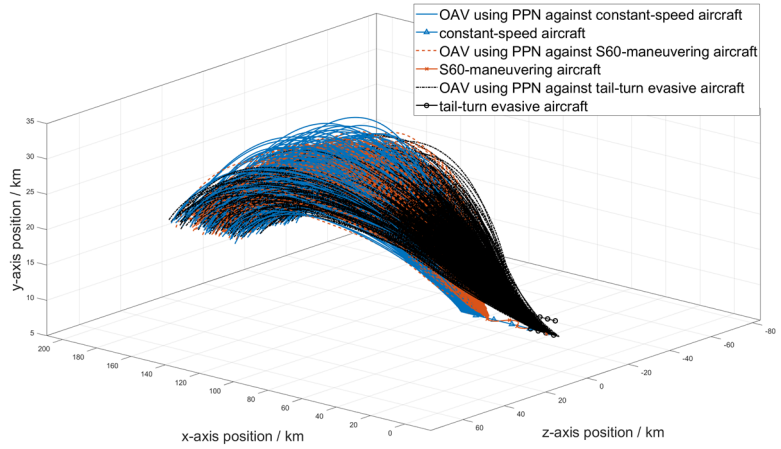


Figure 3 – Expert trajectories of proportional navigation guidance law

To improve the numerical stability and computational efficiency of the training process, the state variables and control variables are normalized as follows:

$$\begin{cases} \tilde{\boldsymbol{x}}_k = \boldsymbol{C}^{-1}\boldsymbol{x}_k \\ \tilde{\boldsymbol{u}}_k = \boldsymbol{u}_k / A_{\text{scale}} \end{cases} \tag{63}$$

where

$$\begin{cases} \boldsymbol{C} = \begin{pmatrix} L_{\text{scale}}\boldsymbol{I}_3 & \boldsymbol{0}_3 \\ \boldsymbol{0}_3 & V_{\text{scale}}\boldsymbol{I}_3 \end{pmatrix} \\ A_{\text{scale}} = V_{\text{scale}}^2 / L_{\text{scale}} \end{cases} \tag{64}$$

The normalization coefficients are determined by the maximum absolute values of the state and control variables extracted from the expert trajectories. Furthermore, the behavioral characteristic parameters $\boldsymbol{\theta}_t$, namely the elements of the matrices in Equation (36), are parameterized using Chebyshev polynomials. The expression of the p-th order Chebyshev polynomial is:

$$T_p(t) = \cos\left[p \cdot \arccos(t)\right] \tag{65}$$

The polynomial functions of different orders form a vector function:

$$\boldsymbol{\phi}(t) = \left[T_0(t), T_1(t), \cdots, T_n(t)\right]^{\mathrm{T}} \tag{66}$$

Thus, the values of the behavioral characteristic parameters $\boldsymbol{\theta}_t \in \mathbb{R}^{m\times 1}$ at each time step are represented as:

$$\boldsymbol{\theta}_t = \tilde{\boldsymbol{\theta}} \cdot \boldsymbol{\phi}(t) \quad \tilde{\boldsymbol{\theta}} \in \mathbb{R}^{m\times n} \tag{67}$$

The gradient-based update rule for behavioral characteristic parameters in Equation (30) is modified to:

$$\tilde{\boldsymbol{\theta}}^{k+1} = \tilde{\boldsymbol{\theta}}^{k} - \eta \sum_{t} \left[\boldsymbol{\phi}(t) \frac{\partial \Lambda}{\partial \boldsymbol{\theta}_t^{\mathrm{T}}} \right]^{\mathrm{T}} \tag{68}$$

Balancing computational efficiency and learning performance, the order p is set to 10. The training process is implemented using the unconstrained nonlinear optimization solver fminunc in MATLAB R2023b. The configuration adopts a quasi-Newton method with analytical gradients, the BFGS algorithm for Hessian matrix approximation, and the optimization termination tolerance is set to 1×10^{-4}. The simulation platform specifications are: CPU: Intel Core i9-13900HX, RAM: 40 GB, GPU: NVIDIA GeForce RTX 4060 (8 GB).

The training curves are depicted in Fig. 4. As the training progresses, the objective function (21) gradually decreases and converges, while the gradient norm of the problem approaches zero. This indicates that the learned behavioral characteristic parameters align the trajectory features of the strategy function with those of the expert trajectories, which occurs if and only if the equality constraint in Equation (18) is satisfied.

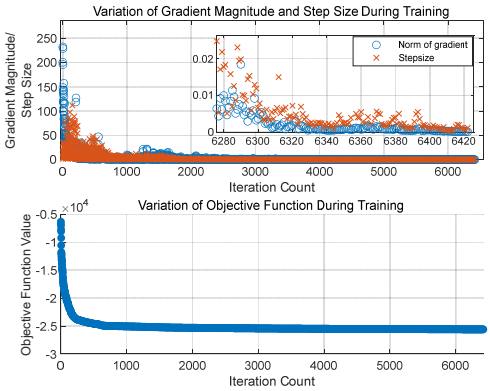


Figure 4 – Variation of metrics and gradient magnitudes during training.

To comprehensively evaluate the performance of the proposed incoming OAV trajectory prediction algorithm in the three-body active defense scenario, three core aspects are mainly verified:

- The long-term trajectory prediction accuracy of inverse reinforcement learning;
- The dynamic adjustment capability of weighting coefficients;
- The improvement effect of the trajectory prediction algorithm on guidance performance, including policy optimality and miss distance.

5.2 Analysis of Prediction Accuracy and Weight Coefficient Influence

To compare the prediction accuracy of different trajectory prediction distributions and verify the calculation performance of target guidance law weighting coefficients within sliding time windows, simulation evaluations are carried out on trajectory data of the PPN test set. The sliding window size is set as w=201, corresponding to a total duration of 10 seconds. Tests are firstly conducted on incoming OAV trajectories under the condition that the aircraft flies at a constant speed in straight line. Starting from the moment when the target OAV exits the boost phase, target observation sequences from the 11th second to the end of the whole trajectory are continuously collected via the sliding time window. The likelihood function values within the window are calculated using the mean values of various prediction distributions, and the estimated weights of corresponding guidance laws are consequently obtained, as shown in Figure 5.

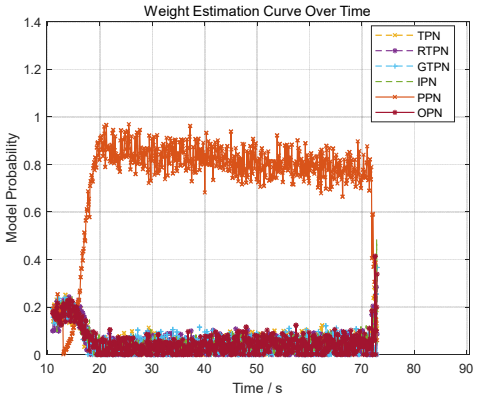
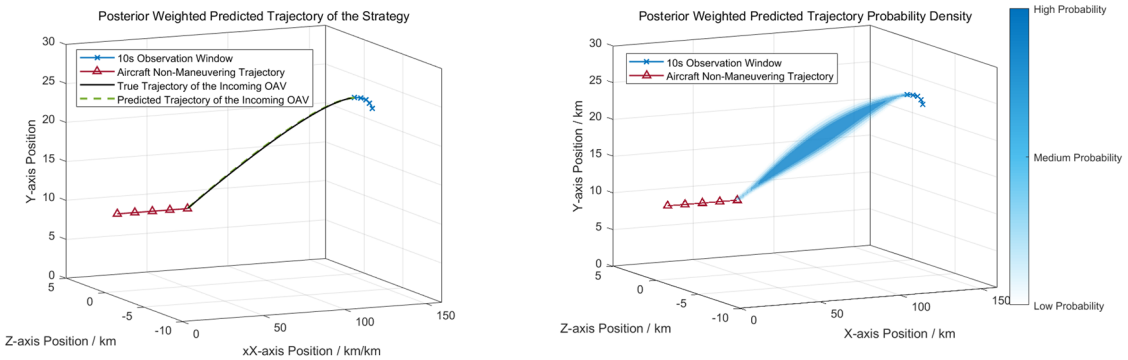


Figure 5 – Analysis of behavioral feature parameter learning performance.

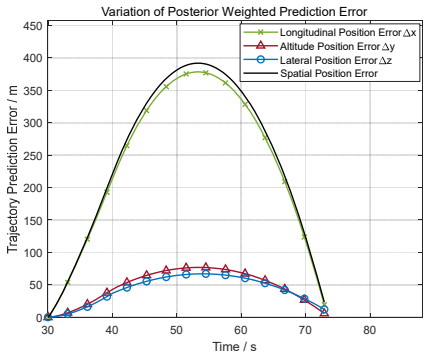
In Figure 5, the estimated weight of PPN gradually rises from uniform distribution to 80% within the first 10 seconds, which is restricted by the insufficient effective data length in the initial sliding window. Afterwards, the identified weight of PPN remains above 65%, showing a distinct difference from other guidance laws. The proposed method reinitializes weights at each time step and calculates observation likelihood merely within a fixed-length sliding window of size w, so as to guarantee favorable dynamic adaptability to violent maneuvers of incoming OAVs. The steady-state weight converges to approximately 80% in this method, while MMAE can approach nearly 100%. This phenomenon results from the adoption of abundant candidate guidance laws in weight iteration. Under the interference of measurement noise, guidance laws with similar motion characteristics interfere with steady weight values. Nevertheless, this mechanism enables the algorithm to reserve low-probability guidance strategies, making the fused prediction strategy constructed by weight combination more adaptive and robust.

This paper models trajectory characteristics under diverse guidance laws and introduces corresponding strategy functions via inverse reinforcement learning, which endows the algorithm with superior universality and effectively improves overall prediction accuracy.

Figure 6 presents the trajectory prediction results within the time interval [30,72.9]s using observation data collected from 20 s to 30 s, denoted as R(20,[30,72.9]). Figure 6(a) and Figure 6(b) separately illustrate the mean trajectory and spatial dispersion of the predicted distribution, where dark blue represents high-probability regions and light yellow denotes low-probability regions. Figure 6(c) shows the prediction errors in radial, altitude, lateral directions and overall spatial position for R(20,[30,72.9]). It can be observed that the maximum overall spatial extrapolation error reaches about 391 meters around the 54th second. Thereafter, driven by the inherent attack intention of the incoming OAV toward the aircraft, the spatial position prediction error gradually decreases and drops to roughly 28 meters at the terminal moment.



(a) Predicted trajectory (b) Predicted trajectory probability density



(c) Three-Axis spatial position prediction error

Figure 6 – Predicted trajectory of incoming OAV when aircraft is not maneuvering.

To verify the dynamic adaptability of the proposed algorithm, prediction results under continuous sliding input conditions are evaluated, namely R(11+0ΔT,[21+0ΔT,72.9]), R(11+1ΔT,[21+1ΔT,72.9]), …, R(11+ n_{NUM} ΔT,[21+ n_{NUM},72.9]). The dynamic variation curve of the maximum spatial position prediction error within each prediction interval is plotted in Figure 7. In the formula, n_{NUM} =|(72.9-21)/ ΔT |. Considering the real-time requirement of trajectory integral calculation, the time step is set as ΔT=0.1 s.

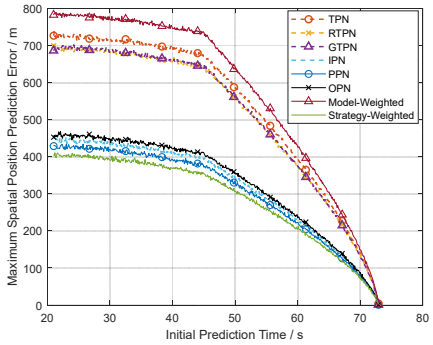


Figure 7 – Dynamic curves of maximum spatial position prediction error when aircraft is at constant speed and height.

Figure 7 indicates that the maximum spatial position prediction error gradually declines as the remaining flight time decreases. Among all prediction distributions, the strategy-weighted distribution achieves the minimum prediction error, followed by the single PPN mode prediction distribution.

Furthermore, tests are conducted on incoming OAV trajectories when the aircraft performs continuous S-shaped maneuvers. According to the proposed algorithm, the likelihood function within the sliding window is calculated using the mean values of different prediction distributions, and the corresponding estimated weights of each guidance law are obtained as shown in Figure 8.

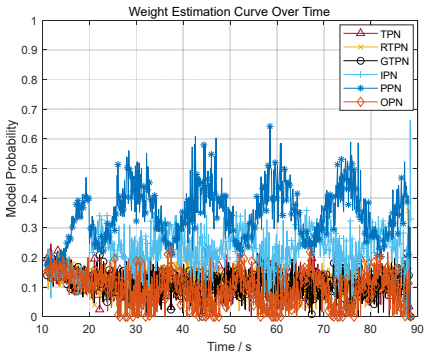
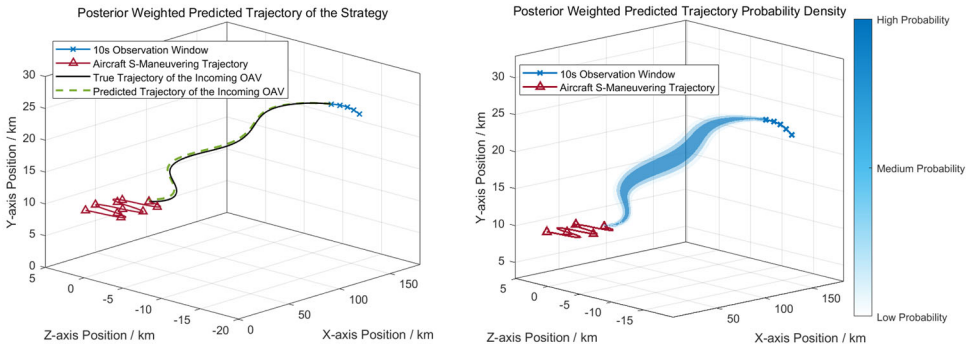


Figure 8 – Weight estimation curves during aircraft S-maneuver.

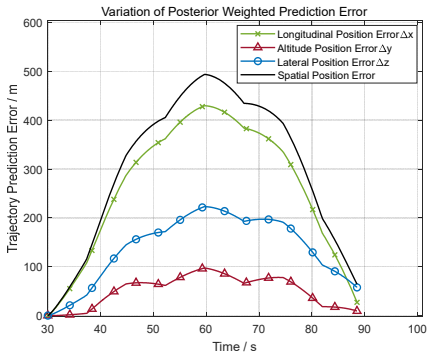
Figure 8 shows that the estimated weight of PPN calculated by the proposed algorithm gradually rises from uniform distribution to 40% within the first ten seconds, which is limited by insufficient valid data in the initial sliding window. From then on until the end of the target trajectory, the identified PPN weight fluctuates around 35%, and is slightly higher than those of IPN and OPN guidance laws

in most cases, which is affected by the maneuver acceleration amplitude of incoming OAVs and observation noise.

Figure 9 presents the trajectory prediction results of R(20,[30,88.5]). Figure 9(a) and Figure 9(b) respectively depict the mean value and spatial dispersion of trajectory prediction distributions, with dark blue for high-probability areas and light yellow for low-probability areas. Figure 9(c) illustrates the variation curves of prediction errors in radial, altitude, lateral directions and overall spatial position against extrapolation duration. It can be seen that the maximum overall spatial extrapolation error reaches approximately 490 meters at around 60 seconds. Guided by the OAV's inherent strike intention, the spatial position prediction error gradually decreases to about 60 meters at the terminal moment.



(a) 来袭弹预测轨迹　　(b) 轨迹概率密度分布的空间可视化



(c)三轴及空间位置预测误差

Figure 9 – Predicted trajectory of incoming OAV during aircraft S-maneuver.

Prediction performances under sliding input mode are assessed via a series of prediction schemes: R(11+0ΔT,[21+0ΔT,88.5]), R(11+1ΔT,[21+1ΔT,88.5]), …, R(11+ n_{NUM} ΔT,[21+ n_{NUM} ΔT,88.5]).

Herein, n_{NUM} =|(88.5-21)/ Δ T|. The dynamic variation curve of the maximum spatial position prediction error within each prediction interval is plotted in Figure 10.

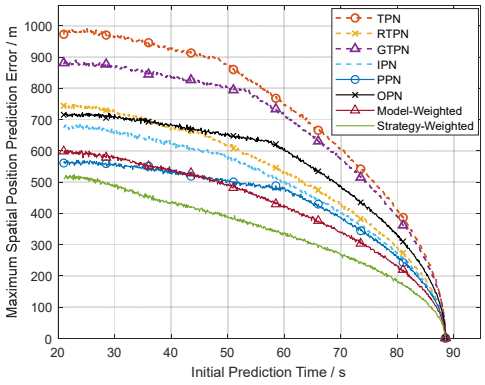


Figure 10 – Dynamic curves of maximum spatial position prediction error during S-shaped maneuver of the aircraft.

It can be seen from Figure 10 that the maximum spatial position prediction error decreases gradually with the reduction of remaining flight time. Nevertheless, the error declines more slowly compared

with the scenario where the aircraft flies at a constant speed and fixed altitude, which is mainly caused by the intense maneuvering of the incoming OAV. Among all methods, the strategy-weighted distribution achieves the optimal prediction performance.

5.3 Evaluation on Prediction Accuracy and Scene Adaptability of the Algorithm

Existing research on incoming OAV trajectory prediction is mainly divided into filtering identification algorithms and deep learning network models. To quantitatively evaluate the prediction performance, four methods including interactive multi-model filtering identification algorithm, adversarial inverse reinforcement learning (AIRL), LSTM-Transformer hybrid model and the proposed algorithm are adopted to conduct trajectory prediction under the three-body active defense scenario. The root mean square error (RMSE) of the maximum spatial position prediction error within 10 seconds and 40 seconds (near collision moment) on the test set is calculated and compared. The interactive multi-model (IMM) filtering algorithm takes the three-axis position, velocity and guidance coefficient of incoming OAVs as state variables, and contains PN, IPN, APN and OPN guidance models. Its predicted trajectories are extrapolated according to the model with the highest probability and the estimated guidance coefficients.

As shown in Table 4, the proposed algorithm outperforms the other three methods in both short-term 10-second and medium-term 40-second prediction tasks. The core reason lies in that the proposed method accurately characterizes target maneuver characteristics such as aerodynamic properties and cost function constraints affecting trajectory sequences into behavioral characteristic parameters θ_t via inverse reinforcement learning. Moreover, the strategy weighting mechanism constructed based on sliding time window effectively weakens the dynamic interference caused by drastic target maneuvers, thus achieving more stable and accurate long-term trajectory prediction.

Table 4 Different prediction mode performance

Aircraft Maneuver Mode	Prediction Algorithm	RMSE of Maximum Spatial Position Prediction Error within 10 s / m	RMSE of Maximum Spatial Position Prediction Error within 40 s / m
Uniform straight flight	IMM filtering identification algorithm	274.2131	550.465
	AIRL method	218.457	439.824
	LSTM-Transformer hybrid model	211.052	424.968
	Proposed algorithm	**162.065**	**394.599**
S-shaped maneuver	IMM filtering identification algorithm	258.143	677.651
	AIRL method	202.556	561.726
	LSTM-Transformer hybrid model	183.210	507.288
	Proposed algorithm	**157.624**	**494.523**
Tail-turn evasive maneuver	IMM filtering identification algorithm	234.227	501.584
	AIRL method	213.385	452.283
	LSTM-Transformer hybrid model	208.617	431.499
	Proposed algorithm	**170.706**	**393.720**

To assess the scene adaptability of the proposed prediction algorithm and its robustness against extreme conditions with abrupt violent maneuvers, two additional sets of violent maneuver simulation conditions are supplemented, forming a total of three working conditions as listed in Table 5.

Table 5 Designed working conditions for algorithm assessment

Working Condition Type	Key Parameters and Trigger Logic	Remarks
Condition 1	Set as the typical three-body active defense scenario (refer to the	

Working Condition Type	Key Parameters and Trigger Logic	Remarks
	textual content in Table 2)	
Condition 2	The aircraft executes extreme S-shaped evasive maneuvers with overloads of 10 *g*, 12.5 *g* and 15 *g*	Other simulation settings are consistent with Condition 1
Condition 3	The guidance strategy of the incoming OAV switches during flight with specified switching conditions $r_{AT} > R_{AT}, R_{AT} \sim U[0.5r_{AT0}, 0.8r_{AT0}]$	

For working condition 3, the relevant threshold value follows a uniform distribution $U[0.5r_{AT0}, 0.8r_{AT0}]$ ranging from $0.5r_{AT0}$ to $0.8r_{AT0}$, and all other simulation parameters remain consistent with working condition 1.

A total of 500 Monte Carlo simulations are carried out under working condition 1. Figure 11 presents the maximum spatial position prediction errors within 10-second and 40-second extrapolation ranges obtained by the proposed algorithm, filtering algorithm, AIRL method and LSTM-Transformer model respectively.

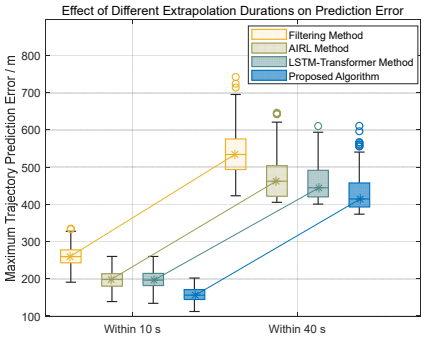
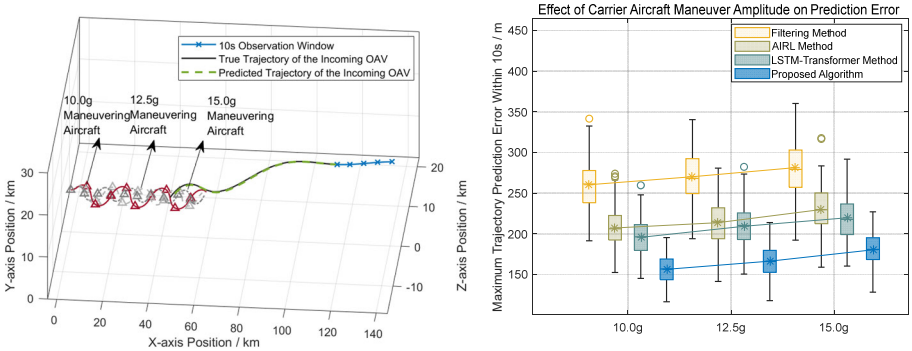


Figure 11 – Prediction error statistics for Condition 1

It can be seen from Figure 11 that compared with the filtering algorithm, AIRL method and LSTM-Transformer model, the box plot of the proposed algorithm is overall lower. Its RMSE values of maximum spatial prediction error within 10 s and 40 s are 158.037 m and 431.742 m respectively. The maximum spatial prediction errors of all four models rise with the increase of extrapolation time, while the proposed algorithm always maintains lower prediction error with narrower box distribution, showing more stable prediction performance.

For working condition 2, to explore the influence of different maneuver amplitudes of the aircraft on algorithm performance, 200 Monte Carlo simulations are conducted for each prediction model under combined S-shaped evasive maneuvers of 10 *g*, 12.5 *g* and 15 *g*. The maximum spatial position prediction error within 10 seconds is taken as the dependent variable to draw the box plots, as shown in Figure 12.



(a)不同载机机动下的来袭弹轨迹预测　　(b)载机机动幅值对预测误差的影响

Figure 12 – Prediction error statistics for Condition 2

It can be observed from Figure 12(b) that the box plot of the proposed algorithm is obviously lower

than those of the filtering algorithm, AIRL method and LSTM-Transformer model, which indicates its superior prediction accuracy. As the evasive overload of the aircraft increases from 10 *g*, 12.5 *g* to 15 *g*, the prediction errors of all four methods increase correspondingly. Nevertheless, the proposed algorithm keeps the lowest error level, with the maximum spatial position prediction errors within 10 s being 156.949 m, 166.310 m and 181.037 m respectively. This fully verifies its outstanding adaptability to intense high-overload evasive maneuvers of the aircraft.

For working condition 3, we mainly analyze the temporal response characteristics of prediction errors after the guidance strategy of the incoming OAV switches. After the guidance law switching moment, the prediction accuracy of the four algorithms is counted at an interval of 0.3 s, and the statistical results are shown in Figure 13.

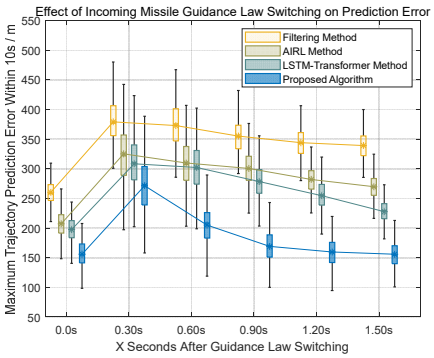


Figure 13 – Prediction error statistics for Condition 3

In Figure 13, the box plot of the proposed algorithm is positioned lower, signifying smaller prediction errors; meanwhile, its narrower box range reflects limited error fluctuation and superior prediction stability. All methods suffer a sharp abrupt rise in prediction error after guidance law switching, followed by a gradual decline. The proposed algorithm exhibits a more distinct error reduction trend, which proves that it possesses faster adaptive capability under Working Condition 3 and can effectively cope with drastic state changes caused by guidance law switching of incoming OAVs.

6. Conclusions

This paper investigates the trajectory prediction of incoming OAVs in three-body active defense scenarios, and proposes an improved inverse reinforcement learning-based trajectory prediction method for incoming OAVs. This method addresses the low prediction accuracy caused by dynamic variations in the maneuver modes of highly maneuverable targets, which hinders the practical application of three-body active defense. The main research conclusions are drawn as follows:

- Under working condition 1, the proposed algorithm achieves higher prediction accuracy than the filtering algorithm, AIRL method and LSTM-Transformer model, with the RMSE values of maximum spatial position prediction error within 10 s and 40 s being 158.037 m and 431.742 m respectively. Under working condition 2 featuring intense evasive maneuvers of the aircraft at 10 *g*, 12.5 *g* and 15 *g* overloads, the proposed method yields smaller prediction errors than the other three methods, and its maximum spatial position prediction errors within 10 s are 156.949 m, 166.310 m and 181.037 m correspondingly. Under working condition 3, the proposed algorithm presents a narrower error distribution within 0 s to 1.5 s after the guidance law switching of incoming OAVs, and exhibits a more remarkable error decline after sudden error fluctuation, thus possessing favorable adaptability.

- When constructing multiple behavioral strategy functions for incoming OAVs, the proposed prediction method quantifies strategy uncertainty based on the maximum causal entropy criterion. Combined with rolling observation data, it realizes adaptive weighting and dynamic optimization of prediction distributions, and thereby achieves high-probability trajectory prediction of incoming OAVs.

In the future, three-body active defense will face challenges such as highly dynamic air combat situations and multi-target game confrontations, and the guidance strategies of incoming OAVs will incorporate more complex decision-making intentions. In the context of active defense for future

advanced combat platforms, further in-depth research is required on trajectory prediction under three-body game confrontation and prediction technology oriented to multi-carrier-aircraft scenarios.

7. Contact Author Email Address

ZHANG HAO: zhanghao1117@mail.nwpu.edu.cn

8. Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] Xie J and Ma K. Distributed defending strategies in target-attacker-defender game with applications to cooperative guidance. *Aerospace Science and Technology*, Vol.173, pp 1-12, 2026.

[2] Chao T, Wang X, Wang S and others. Linear-quadratic and norm-bounded differential game combined guidance strategy against active defense aircraft in three-player engagement. *Chinese Journal of Aeronautics*, Vol.36, No.8, pp 331-350, 2023.

[3] Wang X, Chao T, Hou M and others. Two-phase differential game guidance scheme against active defensive aircraft in three-body engagement. *Chinese Journal of Aeronautics*, Vol.39, No.3, pp 1-12, 2026.

[4] Liu S, Lin Z, Wang Y and others. Three-body cooperative active defense guidance law with overload constraints: A small speed ratio perspective. *Chinese Journal of Aeronautics*, Vol.38, No.2, pp 1-11, 2025.

[5] Wang F, Qu Q, Yang H and others. Current status and prospects of guidance techniques for intercepting hypersonic glide vehicles: A review. *Defence Technology*, Vol.58, pp 241-259, 2026.

[6] Dou D, Song T, Tao H and others. Adaptive-dynamic-programming-based cooperative active defense guidance against a guided attacker. *Aerospace Science and Technology*, Vol.177, pp 1-11, 2026.

[7] Wang X, Yang Z, Piao H and others. Intelligent recognition method of target tactical behavior intention in air combat based on deep learning. *Engineering Applications of Artificial Intelligence*, Vol.138, pp 1-12, 2024.

[8] Jiang F, Cui H, Li Y and others. Long-range masked autoencoder for pre-extraction of trajectory features in within-visual-range maneuver recognition. *Defence Technology*, Vol.55, pp 301-315, 2026.

[9] Fonod R and Shima T. Multiple Model Adaptive Evasion Against a Homing Missile. *Journal of Guidance, Control, and Dynamics*, Vol.39, No.7, pp 1578-1592, 2016.

[10] Wang Y, Fan S, Wang J and others. Quick identification of guidance law for an incoming missile using multiple-model mechanism. *Chinese Journal of Aeronautics*, Vol.35, No.9, pp 282-292, 2022.

[11] Cai Y, Zhuang X, Zhong Y and others. Convolutional interactive learning network with auxiliary features for hypersonic glide vehicle trajectory prediction under data missing scenarios. *Aerospace Science and Technology*, Vol.177, pp 1-11, 2026.

[12] Zhi-Fei X, Ying-Xin K, Zhan-Wu L and others. Air combat target maneuver trajectory prediction based on robust regularized Volterra series and adaptive ensemble online transfer learning. *Defence Technology*, Vol.20, pp 187-206, 2023.

[13] Liu Z, Hu Y, Gao C and others. Probabilistic trajectory prediction for reentry glider vehicles using koopman operators under uncertainty. *Aerospace Science and Technology*, Vol.168, pp 1-11, 2026.

[14] Wang Y, Wang J and Fan S. Parameter Identification of a PN-Guided Incoming Missile Using an Improved Multiple-Model Mechanism. *IEEE Transactions on Aerospace and Electronic Systems*, Vol.59, No.5, pp 5888-5899, 2023.

[15] Xu H, Liu Y, Xing Y and others. Lateral Maneuver Discrimination for Hypersonic Glide Vehicles: A Hybrid Approach Combining Model-Driven and Data-Driven Methods. *IEEE Sensors Journal*, Vol.24, No.7, pp 11425-11437, 2024.

[16] Ren J, Wu X, Liu Y and others. Long-Term Trajectory Prediction of Hypersonic Glide Vehicle Based on Physics-Informed Transformer. *IEEE Transactions on Aerospace and Electronic Systems*, Vol.59, No.6, pp 9551-9561, 2023.

[17] Snoswell A J, Singh S P N and Ye N. Revisiting Maximum Entropy Inverse Reinforcement Learning: New Perspectives and Algorithms. *Proc IEEE Symposium Series on Computational Intelligence, Virtual Conference*, pp 241-249, 2020.

[18] Jiang C, Wang H and Ai J. Autonomous maneuver decision-making algorithm for UCAV based on generative adversarial imitation learning. *Aerospace Science and Technology*, Vol.164, pp 1-11, 2025.

[19] Yang B, Lu Y, Wan R and others. Meta-IRLSOT++: A meta-inverse reinforcement learning method for fast adaptation of trajectory prediction networks. *Expert Systems with Applications*, Vol.240, pp 1-12, 2024.

[20]Zhou Y, Yang F, Zhang C and others. Cooperative decision-making algorithm with efficient convergence for UCAV formation in beyond-visual-range air combat based on multi-agent reinforcement learning. *Chinese Journal of Aeronautics*, Vol.37, No.8, pp 311-328, 2024.

[21]Wang L, Zheng S, Tai S and others. UAV air combat autonomous trajectory planning method based on robust adversarial reinforcement learning. *Aerospace Science and Technology*, Vol.153, pp 1-11, 2024.

[22]Levine S and Koltun V. Continuous Inverse Optimal Control with Locally Optimal Examples. *ArXiv*, pp 1-15, 2012.

[23]Kim J and Yang I. Maximum Entropy Optimal Control of Continuous-Time Dynamical Systems. *IEEE Transactions on Automatic Control*, Vol.68, No.4, pp 2018-2033, 2023.