SCIENCE CHINA
Information Sciences

• **Supplementary File** •

A unified intelligent control strategy synthesizing multi-constrained guidance and avoidance penetration

SiBo ZHAO[1], JianWen ZHU[1,2*], WeiMin BAO[1,3] & XiaoPing LI[1]

[1]*School of Aerospace Science and Technology, Xidian University, Xi'an* 710126*, China;*
[2]*College of Missile Engineering, Rocket Force University of Engineering, Xi'an* 710025*, China;*
[3]*China Aerospace Science and Technology Corporation, Beijing* 100048*, China*

This supplementary file mainly includes four appendices. Appendix A provides supplementary explanations on the basic guidance law used in this study. Appendix B provides supplementary explanations on predicting the terminal velocity and heading error of the vehicle. Appendix C provides a theoretical explanation of the SAC algorithm, and supplements the pseudo code of the algorithm based on the flight data generation and network updating. Appendix D is the simulation section, where D.1 is the explanation of simulation conditions. D.2 is the verification of the correctness of the integrated guidance avoidance strategy, including the analysis of the training results of the SAC network and the analysis of the training efficiency after adding process rewards. D.3 is the adaptability verification of the algorithm. D.4 is a comparative experiment, which verifies the superiority of the SAC strategy by comparing with other DRL strategies. Besides, comparing with traditional methods of avoiding NFZs, the avoidance effect, guidance error, and energy loss are analyzed and verified.

Appendix A Optimal gliding guidance method

In a previous research [1], based on the quasi-equilibrium-gliding (QEG) condition and taking the required overload as the control amount, the performance index with the minimum energy loss is established. The optimal longitudinal and lateral overload are designed respectively, satisfying the constraints on terminal latitude, longitude, altitude and velocity. The required overload command is shown in Eq.(A1).

$$\begin{cases} n_y^* = k\left(C_h L_R - C_\theta\right) + 1 \\ n_z^* = \frac{\sigma_{LOS} - \sigma}{k\left(L_{Rf} - L_R\right)} \end{cases} \tag{A1}$$

where n_y^* and n_z^* are optimal longitudinal and lateral overload. $k = \frac{g_0}{v^2} \approx \frac{g}{v^2}$, where g_0 is the gravitational acceleration at sea level. L_R is the current range, and L_{Rf} is the total range of gliding phase. C_h and C_θ are the guidance coefficients based on optimal control, represented as Eq.(A2).

$$\begin{cases} C_h = \frac{6\left(\left(L_R - L_{Rf}\right)\left(\theta_f + \theta\right) - 2h + 2h_f\right)}{k^2\left(L_R - L_{Rf}\right)^3} \\ C_\theta = \frac{2\left(L_R L_{Rf}\left(\theta - \theta_f\right) - L_{Rf}^2\left(2\theta + \theta_f\right) + L_R^2\left(2\theta_f + \theta\right) + 3\left(L_{Rf} + L_R\right)\left(h_f - h\right)\right)}{k^2\left(L_R - L_{Rf}\right)^3} \end{cases} \tag{A2}$$

where θ is the velocity slope angle, θ_f is the terminal velocity slope angle constraint. h is flight height, and h_f is the terminal height constraint. Based on Eq.(A1), the control variable α and v are implicit in Eq.(A3).

$$\begin{cases} \frac{\rho v^2 S_m C_L(Ma,\alpha)}{2g_0} = \sqrt{n_y^{*2} + n_z^{*2}} \\ v = \arctan\left(\frac{n_z^*}{n_y^*}\right) \end{cases} \tag{A3}$$

where ρ is atmospheric density, S_m is vehicle reference area, and $C_L(Ma, \alpha)$ is the lift coefficient, which is related to the Mach number Ma and α.

Appendix B Calculation of terminal velocity and heading error

Based on the analysis of avoidance guidance mission, the process reward value is estimated by predicting the terminal velocity and position deviation of the vehicle. In this appendix, the methods of velocity prediction and position prediction are derived and calculated.

* Corresponding author (email: zhujianwen1117@163.com)

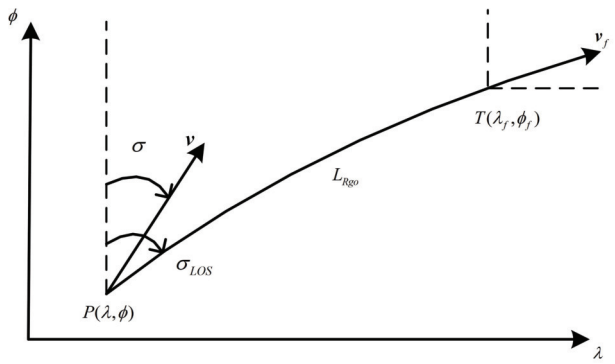

Figure B1 The lateral relative motion.

Appendix B.1 Terminal velocity prediction

Based on the current latitude and longitude (λ, ϕ) of vehicle and the terminal position (λ_f, ϕ_f), the remaining flight range is accurately calculated as

$$L_{Rgo} = R_e \arccos\left(\sin\phi \sin\phi_f + \cos\phi \cos\phi_f \cos\left(\lambda_f - \lambda\right)\right) \tag{B1}$$

where R_e is average radius of the earth. Combining the current flight velocity and the differential of velocity $\dot{v}$, the remaining flight time T_{go} can be predicted as Eq.(B2).

$$T_{go} \approx \frac{-v\cos\theta + \sqrt{v^2\cos^2\theta + 2\dot{v}L_{Rgo}}}{\dot{v}} \tag{B2}$$

In the second half of the gliding flight, the vehicle has enough lift to maintain the QEG flight, while the longitudinal force is basically balanced. Hence, the positive aerodynamic lift L is basically equal to negative gravitational force, as shown in Eq.(B3).

$$L \approx mg \tag{B3}$$

For gliding flight, the Lift to Drag ratio $R_{L/D}$ is always large and the variation range is less, which is set as a constant in the same guidance cycle, the aerodynamic drag can be indirectly expressed as

$$D = \frac{L}{R_{L/D}} \approx \frac{mg}{R_{L/D}} \tag{B4}$$

According to the simplified aerodynamic drag in Eq.(B4), $\dot{v}$ is transformed into

$$\dot{v} = -\frac{D}{m} - g\sin\theta = -\frac{g}{R_{L/D}} - g\sin\theta \tag{B5}$$

When the vehicle satisfies the gliding flight, the velocity slope angle is small and the differential are closed to zero, regarding the right-hand side of Eq.(B5) as a constant. The terminal predicted velocity is obtained by definite integral of Eq.(B6).

$$\begin{aligned} \int_t^{t_f} \dot{v}\mathrm{d}t &= \int_t^{t+T_{go}} \left(-\tfrac{g}{R_{L/D}} - g\sin\theta\right)\mathrm{d}t \\ \Rightarrow v_{fp} &= v - \left(\tfrac{g}{R_{L/D}} + g\sin\theta\right)T_{go} \end{aligned} \tag{B6}$$

Appendix B.2 Terminal Line-Of-Sight (LOS) angular rate prediction

The lateral relative motion relationship of vehicle is shown in Figure B1. The necessary and sufficient condition of satisfying terminal position constraint is controlling the heading error to zero, which is directly related to the LOS angular rate $\dot{\sigma}_{LOSF}$.

Lateral relative motion model is shown in Eq.(B7).

$$\begin{cases} \dot{L}_{Rgo} = -v\cos\Delta\sigma \\ L_{Rgo}\dot{\sigma}_{LOS} = v\sin\Delta\sigma \end{cases} \tag{B7}$$

The heading error is $\Delta\sigma$, calculated as $\Delta\sigma = \sigma - \sigma_{LOS}$. Taking the derivation of the second formula in Eq.(B7).

$$\dot{L}_{Rgo}\dot{\sigma}_{LOS} + L_{Rgo}\ddot{\sigma}_{LOS} = \dot{v}\sin\Delta\sigma + v\Delta\dot{\sigma}\cos\Delta\sigma \tag{B8}$$

Bring the heading error and first formula in Eq.(B7) into Eq.(B8), the rate of LOS angular rate is calculated by Eq.(B9).

$$\begin{aligned} &\dot{L}_{Rgo}\dot{\sigma}_{LOS} + L_{Rgo}\ddot{\sigma}_{LOS} = \dot{v}\sin\Delta\sigma + v\Delta\dot{\sigma}\cos\Delta\sigma \Rightarrow \\ &\dot{L}_{Rgo}\dot{\sigma}_{LOS} + L_{Rgo}\ddot{\sigma}_{LOS} = \tfrac{\dot{v}}{v}L_{Rgo}\dot{\sigma}_{LOS} - \dot{L}_{Rgo}\dot{\sigma}_{LOS} + \dot{L}_{Rgo}\dot{\sigma} \Rightarrow \\ &\ddot{\sigma}_{LOS} = \left(\tfrac{\dot{v}}{v} - \tfrac{2\dot{L}_{Rgo}}{L_{Rgo}}\right)\dot{\sigma}_{LOS} + \tfrac{\dot{L}_{Rgo}}{L_{Rgo}}\dot{\sigma} \end{aligned} \tag{B9}$$

Defining T_{goc} as the predicted remaining time of flight, T_{goc} is derived via the remaining flight range and variation in range, expressing in Eq.(B10).

$$T_{goc} = -\frac{L_{Rgo}}{\dot{L}_{Rgo}} \tag{B10}$$

Defining the value of state x=$\dot{\sigma}_{LOS}$ and the value of control $u = \dot{\sigma}_v$, the differential of LOS angular rate is obtained, which is shown in Eq.(B11).

$$\dot{x} = \left(\frac{\dot{v}}{v} + \frac{2}{T_{goc}}\right) x - \frac{1}{T_{goc}} u \tag{B11}$$

In the latter phase of glide flight, $\frac{\dot{v}}{v}$ is an order of magnitude smaller than $\frac{2}{T_{goc}}$. Eq.(B11) is further simplified to Eq.(B12).

$$\dot{x} = \frac{2}{T_{goc}} x - \frac{1}{T_{goc}} u \tag{B12}$$

The current remaining flight time T_{goc}, a certain time t of future flight starting from the current time, and the remaining flight time T_{gol} at time t satisfy the following relationship.

$$T_{goc}{=}T_{go1}{+}t \tag{B13}$$

dT_{go1}=dt represents the derivation of remaining flight time. For a given control input u, the definite integral of Eq.(B12) is solved, and the calculated result is shown in Eq.(B14).

$$\begin{aligned} x\left(t\right) &= \mathrm{e}^{\int \frac{2}{T_{go1}}\mathrm{d}t}\left(\int -\frac{u}{T_{go1}}\mathrm{e}^{-\int \frac{2}{T_{go1}}\mathrm{d}t}\mathrm{d}t + C\right) = \mathrm{e}^{\int \frac{2}{T_{goc}-t}\mathrm{d}t}\left(\int -\frac{u}{T_{goc}-t}\mathrm{e}^{-\int \frac{2}{T_{goc}-t}\mathrm{d}t}\mathrm{d}t + C\right) \\ &= \mathrm{e}^{-2\ln\left(T_{goc}-t\right)}\left(\int \frac{u}{t-T_{goc}}\mathrm{e}^{2\ln\left(t-T_{goc}\right)}\mathrm{d}t + C\right) = \frac{1}{\left(T_{goc}-t\right)^2}\left(\int u\left(t-T_{goc}\right)\mathrm{d}t + C\right) \\ &= \frac{1}{\left(T_{goc}-t\right)^2}\left(u\left(\frac{1}{2}t^2 - T_{goc}t\right) + C\right) \end{aligned} \tag{B14}$$

The LOS angular rate is $\dot{\sigma}_{LOS}$, at the current time $t = 0$ the constant C is expressed by Eq.(B15).

$$C{=}\dot{\sigma}_{LOS}T_{goc}^2 \tag{B15}$$

$\dot{\sigma}_{LOS}$ is obtained by Eq.(B16).

$$\dot{\sigma}_{LOS}\left(t\right) = \frac{1}{\left(T_{goc}-t\right)^2}\left(u\left(\frac{1}{2}t^2 - T_{goc}t\right) + \dot{\sigma}_{LOS}T_{goc}^2\right) \tag{B16}$$

$\dot{\sigma}_{LOS}$ at the terminal time t_f is shown in Eq.(B17).

$$\dot{\sigma}_{LOSF}{=}\dot{\sigma}_{LOS}\left(t_f\right) = \frac{1}{\left(T_{goc}-t_f\right)^2}\left(u\left(\frac{1}{2}t_f^2 - T_{goc}t_f\right) + \dot{\sigma}_{LOS}T_{goc}^2\right) \tag{B17}$$

Appendix C The solving of avoidance guidance law via SAC

In this appendix, the focus is on analyzing the principle of SAC. The pseudo-code of SAC networks is shown in the Algorithm C1.

To enhance the adaptability of avoided guidance algorithm under dynamically complex environments, the position of NFZs and target varies randomly within a certain range (1°) during the training process. When the vehicle has not reached the terminal state, the study introduces prediction method to calculate terminal states, and adding process reward to accelerate training.

For updating critic net parameter ω, critic net outputs the action value $Q_{soft}(\mathbf{s}_t, \mathbf{a}_t)$, based on samples, and actor net outputs the action probability, which is used to calculate entropy $\mathcal{H}\left(\pi\left(\cdot|\,\mathbf{s}_t\right)\right)$. The value function is conducted and shown in Eq.(C1).

$$Q_{soft}(\mathbf{s}_t, \mathbf{a}_t) = \underset{(\mathbf{s}_t, \mathbf{a}_t)\sim\rho_\pi}{E}\left[\sum_{t=0}^{\infty}\gamma^t r(\mathbf{s}_t, \mathbf{a}_t) + \tau\sum_{t=1}^{\infty}\gamma^t\mathcal{H}(\pi(\cdot|\mathbf{s}_t))\right] \tag{C1}$$

where $r(\mathbf{s}_t, \mathbf{a}_t)$ represents the reward value corresponding to the state $\mathbf{s}_t$ and action $\mathbf{a}_t$, γ is discount factor, π is strategy of actor net. Further obtain the Bellman equation.

$$Q_{soft}(\mathbf{s}_t, \mathbf{a}_t) = \underset{\substack{\mathbf{s}_{t+1}\sim\rho(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)\\ \mathbf{a}_{t+1}\sim\pi}}{E}\left[r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\left(Q_{soft}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) + \tau H(\pi(\cdot|\mathbf{s}_{t+1}))\right)\right] \tag{C2}$$

τ is entropy weight. Given by Eq.(C2), the loss function of critic net is acquired.

$$J_Q(\omega) = \underset{\substack{(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})\sim D\\ \mathbf{a}_{t+1}\sim\pi}}{E}\left[\frac{1}{2}\left(Q_{soft}(\mathbf{s}_t, \mathbf{a}_t) - \left(r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\left(Q_{soft}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \tau\log\left(\pi\left(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}\right)\right)\right)\right)\right)^2\right] \tag{C3}$$

For updating actor net parameter ψ, the updating strategy is shown in Eq.(C4).

$$\pi_{new} = \arg\min_{\pi\in\prod} D_{KL}\left(\pi(\cdot|\mathbf{s}_t)\,\middle\|\,\frac{\exp\left(\frac{1}{\tau}Q_{soft}^{\pi_{old}}(\mathbf{s}_t, \cdot)\right)}{Z_{soft}^{\pi_{old}}(\mathbf{s}_t)}\right) \tag{C4}$$

where Π represents the set of strategy, Z is partition function, used to normalized distribution. D_{KL} is Kullback-Leibler (KL) divergence. Combining re-parameterized technique with Eq.(C4), the loss function of actor net is obtained.

$$J_\pi(\psi) = \underset{\mathbf{s}_t\sim D, \varepsilon_t\sim N}{E}[\tau\log\pi(f(\varepsilon_t; \mathbf{s}_t)|\mathbf{s}_t) - Q_{soft}(\mathbf{s}_t, f(\varepsilon_t; \mathbf{s}_t))] \tag{C5}$$

$\mathbf{a}_t = f(\varepsilon_t; \mathbf{s}_t)$, ε_t is the input noise, obeying the distribution N.

Algorithm C1 Avoidance guidance strategy is solved via SAC

Initialize the experience pool D and Total episodes M;
Build and initialize the actor network, critic network and target networks, get the initialized network parameters;
repeat
  Increase in training episodes
  Initialize target and NFZs positions with random deviation
  repeat
    Sample action from the policy $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    Sample state transition from the environment $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    Calculate process reward values $f_r = c_1\bar{v}_{fp} - c_2\bar{T}_{NFZ} - c_3\bar{\sigma}_{LOSF}$
    Store the experience data in the reply buffer ($\mathbf{s}_t$,$\mathbf{a}_t$,$f_r(t)$,$\mathbf{s}_{t+1}$)
  until End of avoidance guidance mission
  updating critic net:
  Actor net outputs the action probability
  calculate entropy $\mathcal{H}(\pi(\cdot|\mathbf{s}_t))$
  τ is added to $\mathcal{H}(\pi(\cdot|\mathbf{s}_t))$
  Calculate the value function $Q_{soft}(\mathbf{s}_t, \mathbf{a}_t)$
  Calculate the loss function of critic net $J_Q(\omega)$
  Critic net $\omega \leftarrow \nabla J_Q(\omega)$
  updating actor net
  Based on Kullback-Leibler (KL) divergence, the updating strategy is π_{new}
  The loss function of actor net $J_\pi(\psi)$
  Actor net $\psi \leftarrow \nabla J_\pi(\psi)$
until Training episodes are over M

Table D1 Simulation conditions and SAC training parameters

Category	Numerical value	Category	Numerical value
Initial velocity	6500 m/s	Learning episodes	1000
Initial velocity inclination	0°	Guidance period	0.1 s
Initial velocity azimuth	0°	Data sampling interval	50 Km
Initial position	(0°E, 0°N)	Discount factor	0.99
Initial altitude	65 km	Entropy weight	(-10,10)
Terminal altitude	30 km	Learning rate	0.005
Target position	(95°E, 30°N)	Sampling size for each train	128
Longitude of NFZ1	[30°E,31°E]	Net layers	2
Latitude of NFZ1	[20°N,21°N]	Net nodes	256
Longitude of NFZ2	[59°E,60°E]	Capacity of experience pool	20000
Latitude of NFZ2	[21°N,22°N]	Random seed	1
NFZ1 radius	500 Km	NFZ2 radius	300 Km

Appendix D Simulation analysis

Firstly, we simulate and analyze the training results of the SAC based avoidance guidance model, and verify whether the training efficiency has improved after adding process rewards. Secondly, we conduct online testing on the trained DNN and analyze adaptability by changing the position of NFZs and target position. Thirdly, on the basis of the original model, the number of NFZs is expanded, explore whether vehicle can achieve efficient penetration and high precise guidance on the basis of increasing the NFZs. Finally, we increase comparative experiments to verify the superiority of the proposed scheme in this study.

Appendix D.1 Simulation and SAC training conditions

Taking CAV-H to verify the avoidance guidance performance. The initial conditions, terminal altitude and SAC training parameters are given in Table D1. This study uses Python 3.8 and Pytorch 1.12 for simulation, and the processor used for training is Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz.

Appendix D.2 Avoidance guidance strategy simulation verification

The DNN training results are shown in Fig. D1. We conducted simulation analysis on the cases of increasing process rewards and only calculating terminal rewards, as shown in the Fig. D1(a). The training results corresponding to increasing process rewards are significantly better than those without process rewards. After adding process rewards, the training can quickly converge to a higher value. On the contrary, relying solely on the reward value corresponding to the terminal state requires more training episodes to achieve convergence. Fig. D1(b) depicts the training result of NFZs encounter time, in first 200 episodes, which is converged to zero, on the contrary, the training results of terminal position deviation and velocity have not converged to optimal value. In the last 100 episodes, there are more training results conforming the terminal multiple constraints. In Fig. D1(c), the terminal position deviation is converged after 250 episodes, in later exploring, the terminal position deviation is little effected. It can be seen from Fig. D1(d), the terminal velocity is changed intensely with exploring in the former 700 episodes, due to the lateral maneuvering affected directly velocity loss, in the last 200 episodes, the terminal velocity is converged to 2650 – 2750 m/s. In order to improve the adaptability of algorithm, the position of NFZs is not fixed when networks generate flight data, the changeable flight trajectories lead to a less fluctuation in terminal velocity. The training results of SAC network meet the vehicle terminal

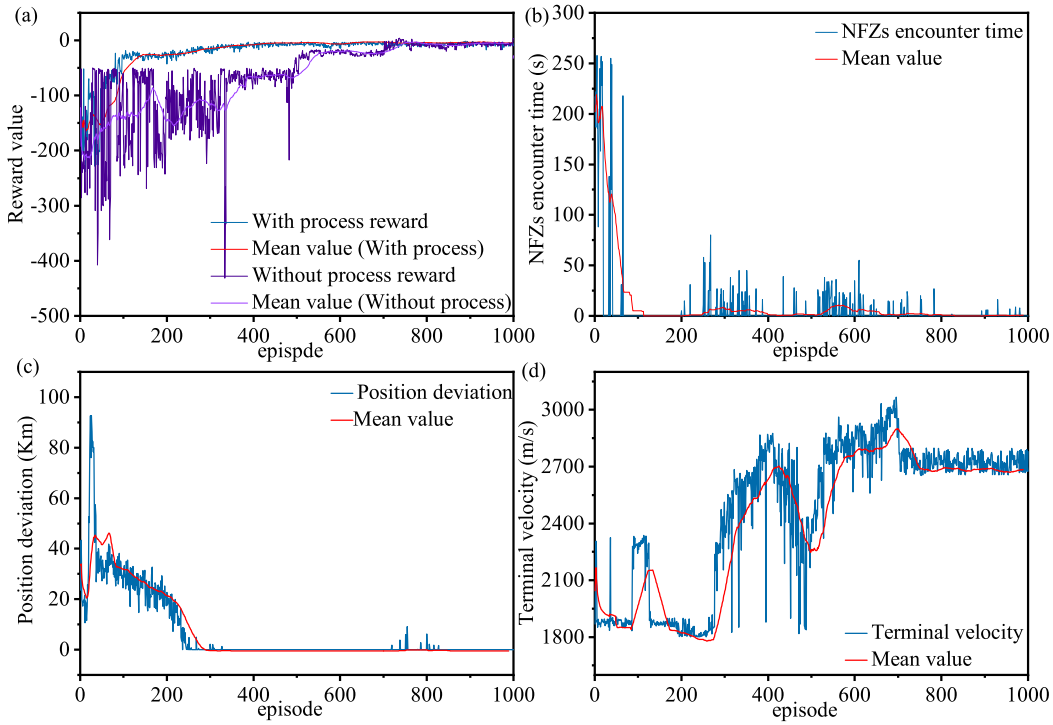

Figure D1 The DNN training results. (a) Reward value; (b) NFZs encounter time; (c) Position deviation; (d) Terminal velocity.

Table D2 The test results under different NFZs positions

NFZs position	NFZs encounter time (s)	Terminal position deviation (m)	Terminal velocity (m/s)
(29°E, 20°N) (60°E, 22°N)	0	2	2737.5
(29°E, 20°N) (60°E, 21°N)	0	0.16	2730.9
(28°E, 20°N) (60°E, 22°N)	0	0.25	2786
(28°E, 20°N) (62°E, 23°N)	0	0.4	2787.7

Table D3 The test results under different target positions

Target position	NFZs encounter time (s)	Terminal position deviation (m)	Terminal velocity (m/s)
(95°E, 28.5°N)	0	1.17	2846.8
(94°E, 29°N)	0	0.53	2922.7
(96°E, 29°N)	0	0.76	2709
(97°E, 30°N)	0	0.21	2519.2

constraints, indicating DRL is an effectiveness and feasibility method on solving the problem of avoidance. The off-line trained DNN parameters are stored and will be utilized to on-line generating avoidance command.

Appendix D.3 Adaptability simulation verification

The adaptability is mainly reflected in the adaptability of environment. The adaptability of the environment mainly considers whether the original evasive penetration strategy can adapt to the mission requirements of the new environment. The ability of vehicle to avoid NFZ depends on its own maneuverability. The advantage of DRL in solving avoidance guidance strategies lies in the ability to generate maneuver instructions online through offline training of neural networks. In order to enhance the adaptability of algorithm, the training process of DNN is improved by adding a random error to the position of NFZs within one unit of longitude or latitude when generating flight data. To verify whether the adaptability of algorithm is enhanced, we change the position of NFZs within two units of longitude or latitude, guaranteeing the position of NFZs is not existed in training sample.

The value of terminal constraints is shown in the Table D2, on the basis of completing avoidance mission, the terminal position deviation is within 2 m. These simulation results verify the DRL strategy has certain adaptability, under the threat of dynamic NFZs, the vehicle can achieve avoidance guidance mission with high precise and less energy loss.

As shown in the Table D3, by adjusting the position of the terminal target, the adaptability is verified based on the original network parameters.

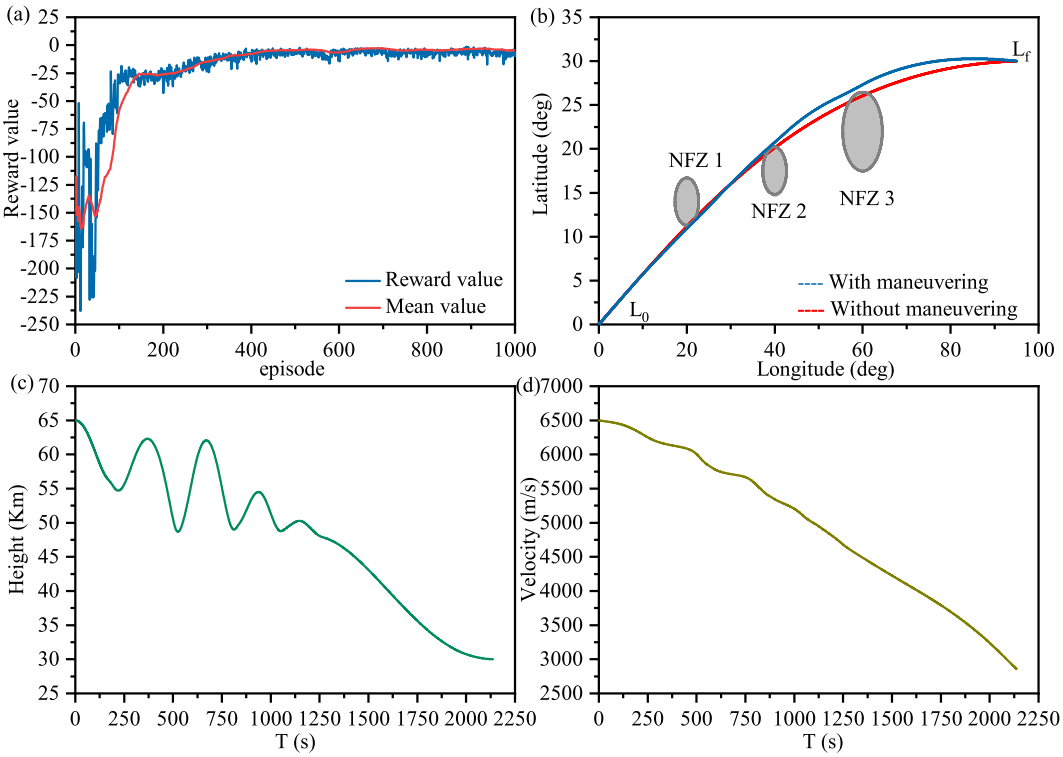

Figure D2 The training result. (a) Reward value; (b) The penetration trajectory; (c) The height; (d) The velocity.

Appendix D.4 Simulation results of increasing of NFZ

On the basis of the original model, the number of NFZs is expanded, as shown in the Fig. D2 (b). The model is solved via the algorithm proposed in this study, exploring whether vehicle can achieve efficient penetration and high precise guidance on the basis of increasing the NFZs.

Fig. D2(a) shows the terminal reward value, in the initial training phase, the strategies are continuously exploring by SAC networks, and the DNN parameters are not optimized. With the training of networks, the more optimal strategies are found by actor net. The terminal reward value converges to the maximum, while the training results are steady. Fig. D2(b) shows the penetration guidance trajectory, and compared with the trajectory without the evasion penetration maneuver strategy, we can see that the strategy trained in this study can achieve efficient penetration on the premise of meeting the high-precision guidance. As shown in Fig. D2(c), the flight height is satisfied the terminal constraint, which is close to 30 Km at the terminal position. As shown in Fig. D2(d), the terminal velocity is consumed by lateral maneuvering overload while avoiding the NFZs, the velocity loss is controlling to the minimum and close to 2800 m/s. To sum up, the intelligent avoidance penetration guidance strategy proposed in this study is also applicable to the scene of the expansion of the NFZ, and achieves efficient penetration under the premise of ensuring high-precision guidance.

Appendix D.5 Algorithm comparison experiment

We first validate the superiority of the SAC algorithm compared to other strategy algorithms based on training results. Secondly, by comparing with traditional avoidance methods, we verify whether the integrated guidance and avoidance solution proposed in this study can achieve guidance and avoidance missions with minimal energy loss.

The criteria for evaluating the superiority of DRL method mainly include the accuracy of training results and the efficiency of training. DDPG [2] and PPO [3] algorithms are representative traditional algorithms of DRL. Through numerical simulation of these two algorithms, we compared with the SAC method adopted in this study from the perspective of theory and simulation experiments, so as to verify the advantages of SAC method.

The training results of DDPG and SAC are shown in the Fig. D3(a). In the first 100 episodes, the network is in the stage of exploring strategies. Compared with SAC algorithm, DDPG algorithm has explored better strategies. With the training, DDPG failed to learn the better strategies into the network. Because DDPG adopts a deterministic strategy, and the purpose of learning is to maximize the Q value. However, this algorithm is sensitive to hyper-parameters, it is easy to overestimate the existing strategies and fall into the local optimal solution. The SAC algorithm gradually converges with the training, and the network learns the optimal strategy. Because the SAC algorithm uses a random strategy, which assigns equal probability to actions with similar Q values, and does not assign very high probability to any action within the action range, so as to avoid repeatedly selecting the same action and falling into the sub-optimal solution. The training results of PPO and SAC are shown in the Fig. D3(b). In the first 600 episodes, the PPO algorithm has been in the exploration stage and cannot converge. This is because the PPO algorithm uses an on-policy strategy, which has the problem of low sampling efficiency. It needs to learn from a large number of samples before it can converge. To sum up, SAC has stronger exploratory and efficient training and learning ability. Compared with DDPG and PPO algorithm, it is more suitable for the algorithm requirements of this study.

In Ref.[4], the trajectory points closed to NFZs are calculated by sliding analytical formula, and the avoidance scheme is adopted via the trajectory points and segmentation guidance. The avoidance strategy in Ref.[4] is described as G1, and the strategy in this

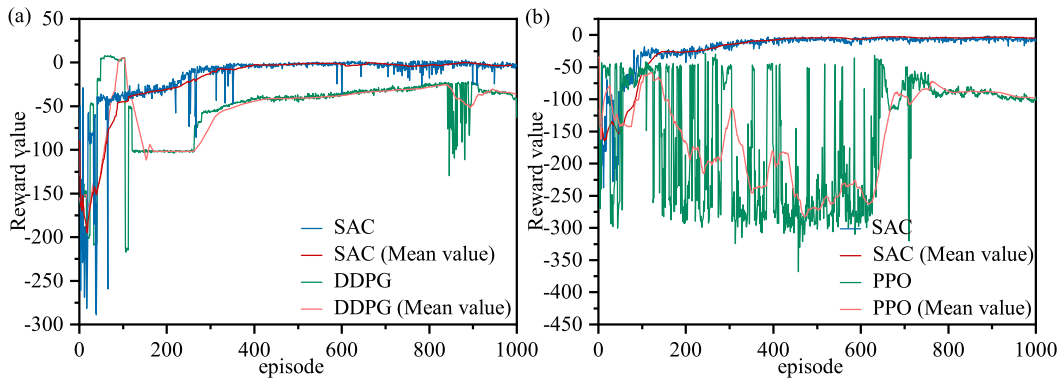

Figure D3 The training results. (a) Reward value between SAC and DDPG; (b) Reward value between SAC and PPO.

Table D4 The test results under different NFZs positions

NFZs position		NFZs encounter time (s)		Terminal position deviation (m)		Terminal velocity (m/s)	
NFZ1	NFZ2	G1	G2	G1	G2	G1	G2
(30°E, 20°N)	(60°E, 22°N)	0	0	3.4	2	1971.7	2737.5
(30°E, 20°N)	(59°E, 21°N)	0	0	0.2	0.8	2130.4	2730.9
(31°E, 21°N)	(60°E, 22°N)	0	0	0.1	0.9	2364.8	2786
(31°E, 21°N)	(59°E, 21°N)	0	0	0.38	3	2476.2	2787.7

study is described as G2.

Fig. D4 shows the trajectory based on G1 and G2 under the changeable NFZs positions. Both of these strategies can achieve the avoidance guidance mission, the amplitude of maneuvering with G2 is more smooth, on the contrary, there is a larger amplitude variation in maneuvering with G1. Table D4 is the terminal results of velocity, position, NFZs encountering time. Compared with G1, the terminal velocity of G2 is larger, which is attributed to the amplitude of maneuvering is less. Obviously, the energy loss of avoidance with G2 is more less than G1.

Compared with the traditional avoidance guidance method, the intelligent avoidance strategy proposed in this study describes NFZs avoidance as a dynamic planning problem, rather than a static planning problem of the optimal flight trajectory. In addition, the strategy aims to find the global optimal solution, minimizing the energy loss by maneuvering, while satisfying the terminal multiple constraints. The results of Fig. D4 and Table D4 plenty verify the superiority of the strategy on solving avoidance penetration problems, and the strategy has strong performance to guarantee the maximized terminal velocity.

In the process of gliding flight, the actual flight status is transferred into the DRL to generate the lateral maneuvering command for avoiding NFZs. Combined optimal guidance with DRL strategy, the vehicle state parameters under different NFZs positions are shown in Fig. D5. The terminal velocity is consumed by lateral maneuvering overload while avoiding the NFZs, the velocity loss is controlled to the minimum and close to 2750m/s, as shown in Fig. D5(a). The control of flight height is realized by longitudinal overload and attack angle, there is a smaller variation range in the attack angle and longitudinal overload, which is attributed to the QEG, as shown in Fig. D5(b, c), the flight height is satisfied the terminal constraint, which is close to 30 Km at the terminal position. as shown in Fig. D5(d). There is larger amplitude modification in the bank angle and lateral overload during the avoidance phase, as shown in Fig. D5(e, f), since the avoidance NFZs strategy is realized by lateral maneuvering. Through the analysis of flight parameters, the intelligent guidance strategy has less impact on QEG, and terminal constraints are satisfied by the vehicle.

References

1 Zhu J W, Zhang H, Zhao S B, et al. Multi-constrained intelligent gliding guidance via optimal control and DQN. Science China Information Sciences, 2023, 66: 132202
2 Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. arXiv preprint, 2015, arXiv:1509.02971
3 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint, 2017 arXiv:1707.06347
4 Yu W B, Chen W C, Jiang Z G, et al. Analytical entry guidance for no-fly-zone avoidance. Aerospace Science and Technology, 2018,72: 426-442

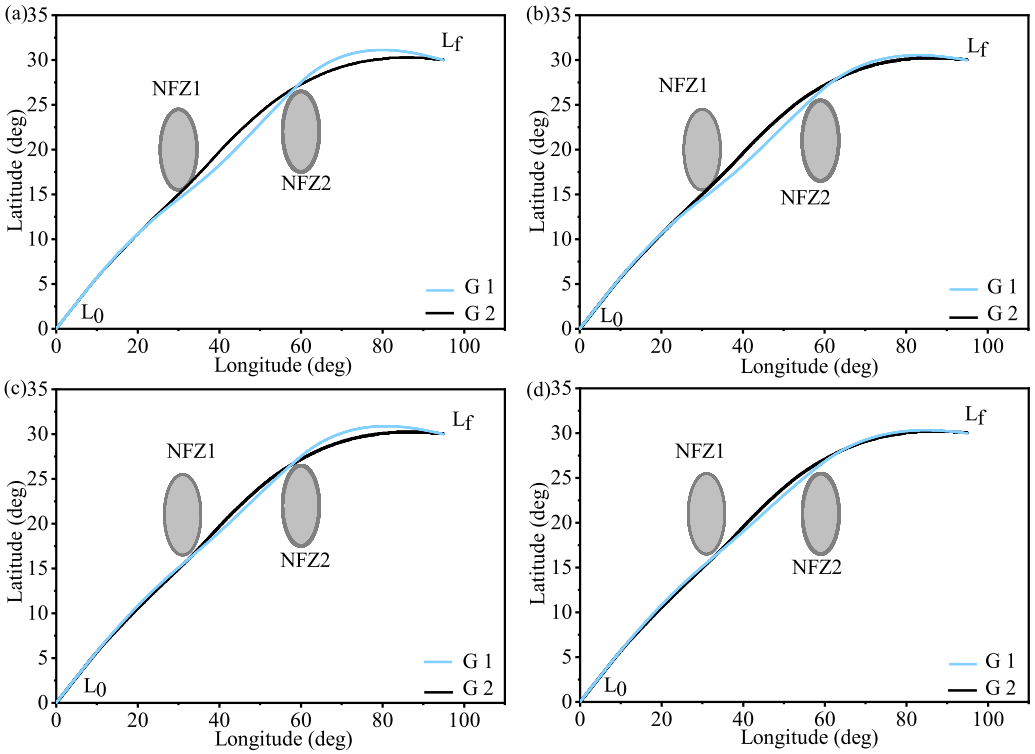

Figure D4 The test trajectory under changeable NFZs position. (a) NFZs position 1; (b) NFZs position 2; (c) NFZs position 3; (d) NFZs position 4.

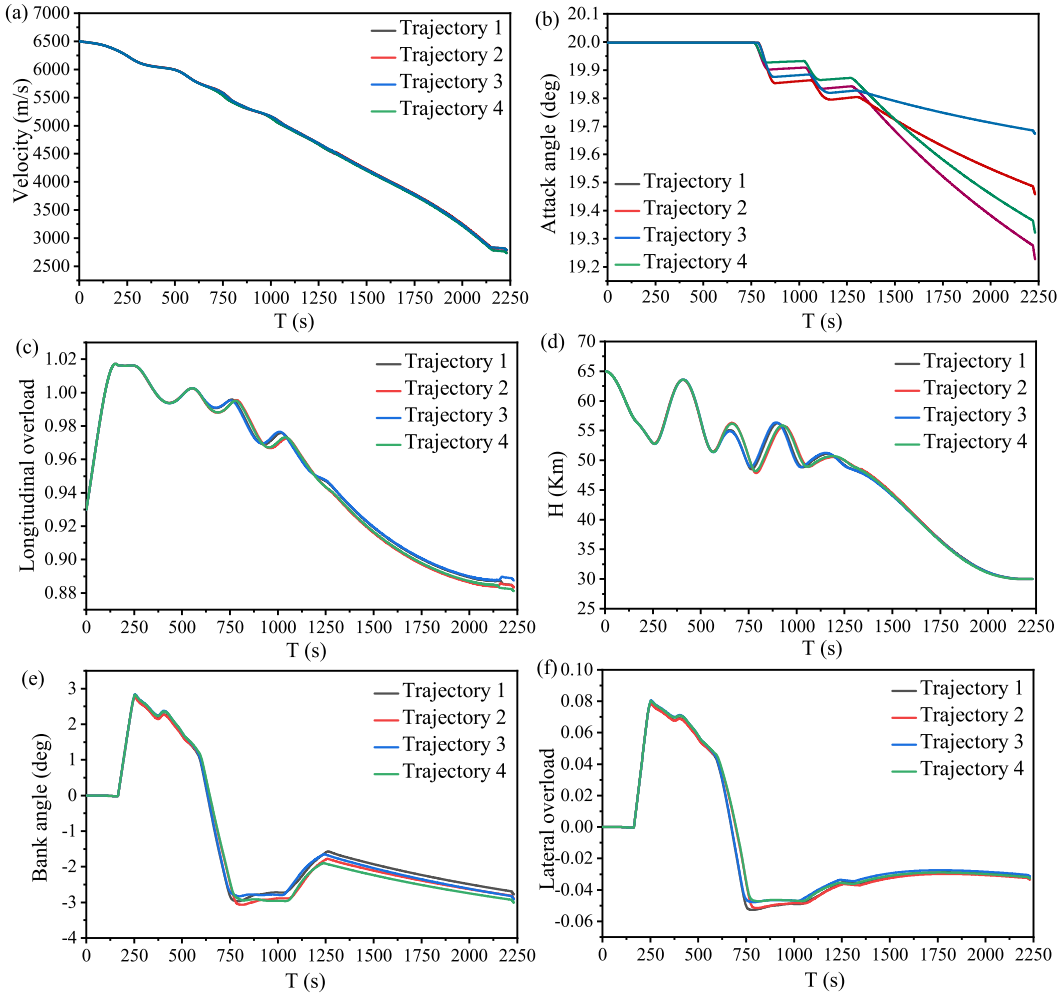

Figure D5 Flight parameters under changeable NFZs positions. (a) velocity; (b) attack angle; (c) longitudinal overload; (d) height; (e) bank angle; (f) lateral overload.