aerospace

MDPI

Article

Phase-Adaptive Reinforcement Learning for Self-Tuning PID Control of Cruise Missiles

Chang Tan, Jianfeng Wang, Hong Cai, Sen Hu, Bangchu Zhang * and Weiyu Zhu *

School of Aeronautics and Astronautics, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China
* Correspondence: zbc2020_sysu@163.com (B.Z.); zhuwy9@mail.sysu.edu.cn (W.Z.)

Abstract

Conventional fixed-gain PID controllers face inherent limitations in maintaining optimal performance across the diverse and dynamic flight phases of cruise missiles. To overcome these challenges, we propose Time-Fusion Proximal Policy Optimization (TF-PPO), a novel adaptive reinforcement learning framework designed specifically for cruise missile control. TF-PPO synergistically integrates Long Short-Term Memory (LSTM) networks for enhanced temporal state perception and phase-specific reward engineering enabling self-evolution of PID parameters. Extensive hardware-in-the-loop experiments tailored to cruise missile dynamics demonstrate that TF-PPO achieves a 36.3% improvement in control accuracy over conventional PID methods. The proposed framework provides a robust, high-precision adaptive control solution capable of enhancing the performance of cruise missile systems under varying operational.

Keywords: cruise missile; attitude control; reinforcement learning; PID; Proximal Policy Optimization

1. Introduction

Cruise missiles are missiles that fly within the atmosphere in a cruising state, and their flight process can be divided into multiple stages such as altitude adjustment, cruise, and dive, each with different aerodynamic characteristics and control requirements [1]. The control system must ensure that the missile maintains a stable attitude and precise trajectory tracking in complex and variable flight environments. Cruise missile controllers commonly use PID controllers for attitude control [2,3]. However, the traditional proportional–integral–derivative (PID) control method, due to its static design [4–6], exhibits significant limitations in addressing nonlinear systems and external disturbances [7]. Studies indicate that traditional PID control struggles to accurately model the dynamic characteristics of missiles, and parameter tuning requires iterative adjustments, which is a cumbersome process [8,9]. Furthermore, the multi-stage flight requirements of cruise missiles necessitate compromises in control parameters across different phases, making it impossible to achieve optimal control across the full speed range, full airspace, and under large maneuver conditions [10].

In order to improve the adaptability and tedious parameter adjustment problems caused by fixed-gain PID, real-time adaptive control schemes with dynamic parameter adjustment functions have received widespread attention. Adaptive PID control typically encompasses rule-based [11,12], and model-based approaches [13,14]. Rule-based methods, such as Ziegler–Nichols tuning [15] and fuzzy logic adaptive PID control, adaptively adjust gains by leveraging control signal characteristics like overshoot and PID error. Model-based techniques, like Model Reference Adaptive Control [16], neural adaptive sliding

Academic Editor: Antonios Tsourdos

Received: 23 August 2025
Revised: 17 September 2025
Accepted: 18 September 2025
Published: 20 September 2025

Citation: Tan, C.; Wang, J.; Cai, H.; Hu, S.; Zhang, B.; Zhu, W. Phase-Adaptive Reinforcement Learning for Self-Tuning PID Control of Cruise Missiles. *Aerospace* **2025**, *12*, 849. https://doi.org/10.3390/aerospace12090849

mode guidance for UAV piloting [17] and three-dimensional sliding pursuit guidance for surface-to-air missiles [18], utilize mathematical models to adjust PID parameters, accommodating changes in system dynamics or external disturbances, thereby improving tracking performance and stability. Although many methods [19,20] have made significant progress in their respective fields, they still have shortcomings for missile attitude control with strong nonlinearity. Rule-based methods rely heavily on predefined rules or expert knowledge, limiting their adaptability to unforeseen conditions in highly dynamic missile flight scenarios. Model-based approaches face challenges in highly nonlinear systems where design complexity increases significantly, and unmodeled dynamics or model errors lead to performance degradation. These challenges are particularly pronounced in the rapidly evolving field of cruise missile control. A novel approach is needed to enhance adaptability to dynamic systems and environments.

With advancements in artificial intelligence, reinforcement learning (RL) [21–23] offers a promising direction for missile attitude control [24–27]. Especially for unmanned aerial vehicle attitude control, RL performs better than traditional control methods. By leveraging deep reinforcement learning techniques combined with PID controllers [28] and designing a reinforcement learning framework [29], they achieved online optimization of network parameters, enhanced the control performance of the aircraft, and attained satisfactory tracking accuracy and robustness. Similarly, Zhao [30] designed a PPO based attitude control framework for unmanned aerial vehicles in a simulation environment demonstrating superior sample efficiency and stability in attitude control. The integration of RL with missiles has primarily focused on guidance law design for autonomous obstacle avoidance [31] and target tracking [32]. For missile attitude control, Zhang [33] developed a DDPG-based longitudinal controller for missile attitude control. Through longitudinal plane simulations, the study demonstrated the feasibility of this data-driven control method in stabilizing the nonlinear angle of attack dynamics in the missile's longitudinal motion. Lee [34] designed an RL-based PID control strategy for large angle-of-attack commands, achieving improved tracking performance.

Despite these developments, current research has certain limitations: existing methods often train in specific scenarios, limiting validation across diverse flight conditions. Simultaneously, they overlook the inherent non-Markovian nature of PID controllers—existing RL controller designs discard crucial historical state information. Additionally, as RL-based AI systems become more general and autonomous, designing reward mechanisms to elicit desired behaviors becomes increasingly important and challenging. Current researchers often focus solely on system design and straightforward objectives, thereby neglecting the design of reward engineering [35], which results in inadequate adaptability of algorithms in multi-stage tasks [32].

To address these limitations, we developed an adaptive PID tuning framework utilizing the Time-Fusion Proximal Policy Optimization (TF-PPO) algorithm. TF-PPO is an enhancement of the established Proximal Policy Optimization (PPO) [22] algorithm, whose stability and efficiency have been demonstrated in adaptive PID tuning applications [36]. Our approach integrates long short-term memory (LSTM) [37] network, robust training method, and adaptive reward function, which can effectively perceive historical states and adaptively adjust parameters to cope with changes in flight states in different scenarios. Specifically, we construct a PPO-based adaptive PID parameter tuning algorithm for missile attitude control, incorporating LSTM networks to capture historical state dependencies and achieve precise attitude regulation. This approach employs a full six-degree-of-freedom aerodynamic model, trained through interactions with a game-based adversarial simulation system. This approach utilizes a full six-degree-of-freedom aerodynamic model and trains through adversarial interactions within a simulation environment. During optimiza-

tion, PID controller gains function as policy parameters in the RL actor network, updated through online environment interactions. For characterizing flight-phase dynamics, we implement reward engineering with specialized functions for altitude adjustment, cruise, descent and terminal guidance phases. Through these designs, optimal parameters for each phase can be automatically identified without relying on initial values. Table 1 compares our method with existing approaches.

Table 1. Comparative Summary of Controller Performance.

Feature	Traditional PID	Rule-Based Adaptive PID	Model-Based Adaptive PID	RL-PID (e.g., DDPG PPO)	TF-PPO
Nonlinear System Handling	Moderate	Moderate	Limited	Good	Good
Parameter Tuning Complexity	High	Moderate	Moderate	Low	Low
Multi-Phase Adaptability	Poor	Moderate	Moderate	High	High
Historical State Utilization	None	Limited	Limited	None	Good

To experimentally assess the proposed framework, a rapid-prototyping missile control system was engineered. This verification infrastructure integrated step-response testing with full-trajectory simulation, with experimental data confirming robustness under divergent initial conditions and wind disturbances. We designed and constructed a reinforcement learning-based rapid prototyping system for missiles, establishing a verification framework that combines step response testing and full envelope simulation. By evaluating attitude tracking performance under varied initial conditions and different wind disturbances, we demonstrated the robustness and effectiveness of the proposed algorithm. This study advances reinforcement learning applications in missile attitude control, providing robust and adaptive solutions for nonlinear dynamics and multi-phase flight scenarios.

The rest of this paper is organized as follows: Section 2 presents system modeling and missile attitude control architecture. Section 3 details the TF-PPO algorithm, including reward engineering and implementation workflow. In Section 4, all experimental results are presented, including step response tests, offline training, and comparative experiments on the missile rapid prototyping system. Section 5 summarizes contributions and future work.

2. System Modeling and Control System Architecture

For high-fidelity missile dynamics modeling, aerodynamic parameter estimation of a cruise missile was performed. Parameters are assumed representative of a generic subsonic cruise missile, which utilizes aerodynamic controls, scaled for simulation purposes based on typical values from literature [38]. The following key parameters from Table 2 are utilized in the missile system model.

Table 2. Missile parameters.

Symbols	Parameter Explanation	Units	Values
m_{wet}/m_{dry}	Fully Loaded/Unloaded mass	kg	675/580
v_{max}	Maximum cruise speed (3000 m)	Ma	0.7
R	Range	km	270
$I_{y,wet}/I_{y,dry}$	Pitch moment of inertia (Fully Loaded/Unloaded)	-	1400/1350
T_{max}	Maximum thrust	N	2700
h_{max}	Service ceiling	m	8000

This paper conducts numerical modeling and analysis of missile aerodynamic characteristics using ANSYS FLUENT (v2022 R2). Combined with key parameters in Table 1

and six-degree-of-freedom rigid-body equations of motion [39–41], a nonlinear dynamic model of the pitch channel is established. Given the similarity in control logic across pitch, roll, and yaw channels, this study focuses exclusively on the self-evolutionary parameter analysis of the pitch loop. Under the assumption of small angles of attack ($\alpha < 15^{\circ}$) [42] and symmetric missile configuration, the coupling from yaw and roll channels can be reasonably neglected [43]. Neglecting coupling effects from yaw and roll channels, the nonlinear pitch dynamics can be described by

$$\begin{cases} \dot{\theta} = q \\ \dot{\alpha} = q + \frac{1}{mV}[\bar{q}S(C_{Z\alpha}\alpha + C_{Z\delta}\delta_e) - T\sin\alpha - mg\cos\theta] \\ \dot{q} = \frac{1}{I_y}\left[\bar{q}S\bar{d}(C_{m\alpha}\alpha + C_{m\delta}\delta_e + C_{mq}\frac{\bar{d}q}{2V})\right] \end{cases} \tag{1}$$

The symbols used in Equation (1) are defined in Table 3.

Table 3. Symbols used in pitch channel dynamics equations.

Symbols	Parameter Explanation	Units
θ	Pitch angle	rad
α	Angle of attack	rad
q	Pitch rate	Rad/s
m	Missile mass	kg
V	Missile velocity	m/s
$\bar{q}$	Dynamic pressure	Pa
S	Reference area	m^2
$\bar{d}$	Reference length	m
$C_{Z\alpha}$	Normal force coefficient due to angle of attack	-
$C_{Z\delta}$	Normal force coefficient due to elevator defection	-
$C_{m\alpha}$	Pitch moment coefficient due to angle of attack	-
$C_{m\delta}$	Pitch moment coefficient due to elevator defection	-
C_{mq}	Pitch moment damping coefficient	-
δ_e	Elevator defection angle	rad
T	Thrust	N
g	Gravitational acceleration	m/s^2
I_y	Pitch moment of inertia	$kg{\cdot}m^2$

The trajectory of a cruise missile consists of a planned trajectory and a guided trajectory. Before launch, the missile's flight plan is predetermined and cannot be altered once launched [1]. The input to the missile's attitude controller is the desired attitude angle generated by a fixed control law. The task of attitude control is to track the desired angle, minimizing the attitude angle error θ_e. The PID controller for the pitch loop attitude can be expressed as

$$\delta_{e,\mathrm{cmd}} = k_p\theta_e + k_i\int\theta_e dt - k_d\omega_z \tag{2}$$

In the equation, $\delta_{e,\mathrm{cmd}}$ represents the elevator command, and the steering gear model is represented by a second-order model [44]:

$$\frac{\delta_e}{\delta_{e,\mathrm{cmd}}} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{3}$$

To address the attitude control requirements of cruise missiles with complex motion characteristics, this paper designs an RL-PID intelligent control system that integrates deep reinforcement learning with classical PID control. Through an online parameter tuning mechanism, optimal control is achieved under different flight conditions.

The overall system architecture, as shown in Figure 1, consists of the RL-PID agent and the missile system model. The RL-PID agent utilizes a deep neural network to real-time parse missile state information s_t, calculates rewards based on a designed reward function, and dynamically outputs the optimal tuning values a_t for PID parameters. It can be expressed as

$$a_t = \pi_\theta(s_t) = [K_p^t, K_i^t, K_d^t]^T \tag{4}$$

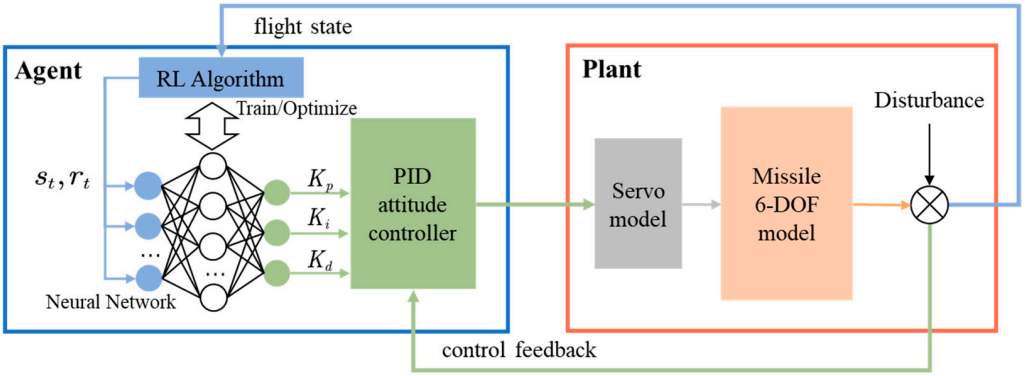


Figure 1. Architecture of the RL-PID Intelligent Control System. The agent adaptively adjusts PID parameters based on the current state s_t and environmental rewards r_t to optimize its policy.

The reinforcement learning algorithm is divided into a training phase and an online optimization phase. During the training phase, a maximum reward mechanism drives the agent to learn parameter tuning strategies. In online operation, the network parameters are fixed, and the K_p, K_i, and K_d coefficients are updated based on real-time flight data, enabling the PID controller to achieve condition awareness and parameter evolution capabilities.

3. Method

TF-PPO algorithm is an optimized framework for missile attitude control, developed on the basis of the Proximal Policy Optimization (PPO) algorithm. Its innovations involve reconstructing the state space, incorporating a temporal network to capture historical state dependencies, tailoring the algorithm workflow specifically for missile autopilot dynamics, and introducing a meticulously designed reward function that accounts for stage-specific characteristics of the cruise missile's flight.

The complete TF-PPO implementation necessitates comprehensive design specifications for states, rewards, actions, and network architecture. As specified in Section 2, the RL-PID controller's action outputs are formally defined. Subsequent sections will elaborate on the design methodology for each component.

3.1. State Space Reframing

Missile control system design relies on long-period observable parameters (e.g., Mach number M_a, angle of attack α, flight altitude H) as states tailored to specific flight phases. Given that the attitude controller's output consists of pitch angle commands, the commanded pitch angle θ_{des} and the current body pitch angle θ_{cur} must be incorporated. The state variables selected in this study are

$$s = [H \quad Ma \quad \alpha \quad \theta_{\text{cmd}} \quad \theta_{\text{body}}] \tag{5}$$

During missile flight, frequent PID adjustments may induce undesirable oscillations. We select $t = 5s$ as the control interval for the RL-PID system. However, since PID response constitutes a dynamic process, the state space design for missile attitude control must satisfy

dual constraints: dynamic response representational completeness and computational feasibility. Considering both missile performance characteristics and state information within control intervals, this paper proposes a sliding-window based temporal state space construction method. The complete 13-dimensional state vector is defined as follows:

$$s_t = \left[Ma_t, H_t, \alpha_t, \underbrace{\theta_{\text{cmd}}^{t-4}, \theta_{\text{cmd}}^{t-3}, \ldots, \theta_{\text{cmd}}^{t}}_{5-\text{stepdesiredpitch}}, \underbrace{\theta_{\text{body}}^{t-4}, \theta_{\text{body}}^{t-3}, \ldots, \theta_{\text{body}}^{t}}_{5-\text{stepactualpitch}} \right] \in \mathbb{R}^{13} \tag{6}$$

where $\theta_{\text{cmd}}^{t-k}$ and $\theta_{\text{body}}^{t-k}$ denote the commanded pitch angle and actual body pitch angle at time $t-k$, respectively, forming a raw temporal sequence spanning a 5 s time window.

3.2. Neural Architecture Design

The TF-PPO algorithm operates on an Actor–Critic [45] framework to collaboratively optimize control policies. Although the Actor–Critic architecture inherently assumes a Markov Decision Process (MDP), the missile PID control system exhibits non-Markovian characteristics due to its integral and derivative components requiring historical signal information prior to the current timestep. To address these non-Markovian properties, we embed a shared LSTM layer within the conventional Actor–Critic structure, thereby constructing a temporal feature fusion network. The detailed architecture is depicted in Figure 2.

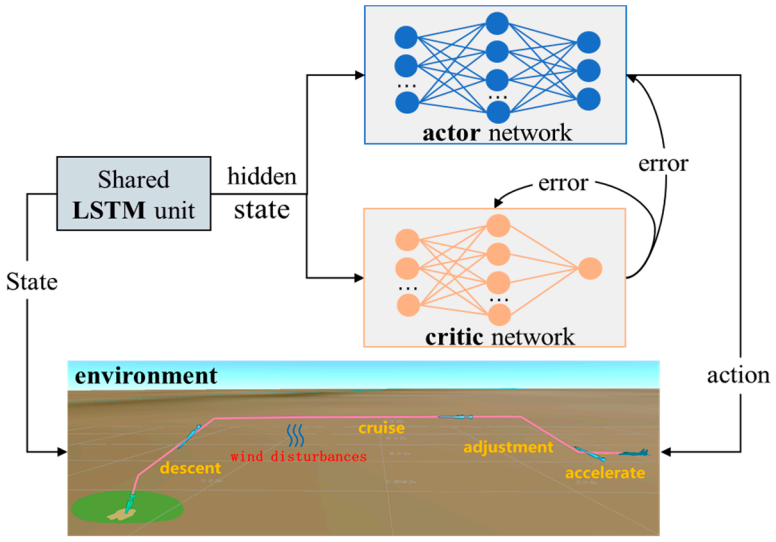


Figure 2. LSTM-Enhanced Actor–Critic Architecture.

The input layer receives a 13-dimensional state vector s_t, which undergoes Batch Normalization preprocessing before entering the shared feature extraction layer. The shared LSTM module employs a unidirectional structure. The temporal dependency between commands and responses is formulated as

$$(z_t, c_t) = \text{LSTM}(s, z_{t-1}, c_{t-1}) \tag{7}$$

where z_t denotes the hidden state output from the LSTM network, c_t represents the cell state in long-term memory. The hidden state is concurrently fed into both Actor and Critic networks, where two distinct fully connected layers generate the action a_t and state-value function $V_\phi(s_t)$, respectively, achieving historical feature reuse.

3.3. Reward Engineering

Reward engineering in reinforcement learning is the process of designing reward systems. Through deliberate reward structuring, it provides agents with discriminative signals that indicate behavioral correctness [35]. The missile control system represents a highly nonlinear precision engineering system. Prior to reinforcement learning deployment, the identification of reward-shaping metrics critical to optimizing control precision is essential for RL-driven controllers. Addressing distinct dynamic characteristics during altitude adjustment, cruise, and descent phases of cruise missiles, this section proposes an adaptive stage-strength quantified reward function. This framework reconstructs RL rewards through phase-strength quantification, enabling stable optimization trajectories for control policies.

3.3.1. Stage-Strength Quantification

The stage-strength $\Psi \in [0,1]$ parameter characterizes the dominance of the current flight phase, where a higher value indicates greater priority for phase-specific control objectives. This parameter depends solely on instantaneous flight states, eliminating inter-phase coupling complexities while ensuring global adaptability with seamless phase transitions. The quantification mechanism is defined as follows:

$$\Psi = \begin{cases} \Psi_{\text{adjust}}\left(\dot{h}_t\right) = \frac{1}{1+e^{-k_1(\hat{h}_t-\delta_1)}} \\ \Psi_{\text{descend}}\left(\dot{h}_t\right) = \frac{1}{1+e^{k_2(\hat{h}_t+\delta_2)}} \\ \Psi_{\text{cruise}}\left(\dot{h}_t\right) = 1-(\Psi_{\text{adjust}}+\Psi_{\text{descend}}) \end{cases} \tag{8}$$

where the shape parameter $k_1, k_2 > 0$ controls the function slope, $\delta_1, \delta_2 > 0$ denotes the phase offset and $\hat{h}$ represents the rate of change of altitude. To eliminate dimensional disparities and establish probabilistic weight allocation, we apply a softmax normalization, this guarantees $\sum \widetilde{\alpha}_i = 1$, thereby ensuring continuous transitions of phase dominance rights.

$$[\widetilde{\Psi}_c, \widetilde{\Psi}_r, \widetilde{\Psi}_d] = \text{softmax}([\Psi_{\text{adjust}}, \Psi_{\text{cruise}}, \Psi_{\text{descend}}]) \tag{9}$$

The schematic of the stage-strength evolution is shown in Figure 3.

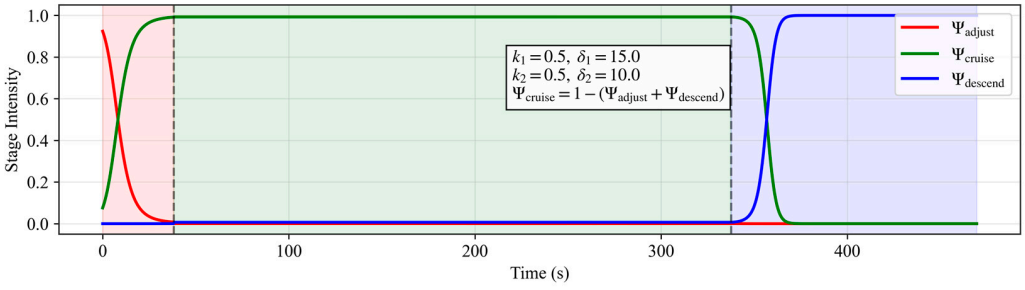


Figure 3. Schematic of Stage-Strength Evolution.

3.3.2. Adaptive Reward Function Formulation

The RL-PID system targets high-precision tracking of guidance-commanded pitch attitudes. Therefore, pitch angle error θ_e and pitch rate error q_e should be considered in the reward function. To mitigate unnecessary oscillations induced by frequent PID adjustments, the elevator deflection angle δ_e is incorporated into the reward function. The global reward function is formulated around three core components: pitch angle error θ_e,

pitch rate error q_e, and elevator deflection δ_e. Phase-adaptive optimization is achieved through the weighting matrix $\mathbf{A} \in \mathbb{R}^{3\times3}$. Its mathematical form is

$$R = \widetilde{\Psi}\mathbf{A}\phi^T \tag{10}$$

where $\phi = [\theta_e \quad q_e \quad \delta_e]$ denotes the parameters considered for the reward function. Tailored to distinct phase characteristics, the weighting matrix $\mathbf{A}$ is architected as

$$\mathbf{A} = \begin{bmatrix} \eta_{\theta,\text{adjust}} & \eta_{q,\text{adjust}} & 0 \\ \eta_{\theta,\text{cruise}} & 0 & \eta_{\delta,\text{cruise}} \\ \eta_{\theta,\text{descent}} & \eta_{q,\text{descent}} & \eta_{\delta,\text{descent}} \end{bmatrix} \tag{11}$$

Each column weight must meet the normalization constraint $\sum \eta_{i,j} = 1(\forall i \in \{\theta, q, \delta\})$, The values of the weight matrix are manually designed based on the characteristics of the flight phase. Through explicit functional coupling between the stage-strength Ψ parameter and weight matrix $\mathbf{A}$, adaptive prioritization switching of control objectives is achieved.

3.4. Training Process

The training process of the TF-PPO algorithm introduces an adaptive exploration mechanism to optimize stability within the canonical PPO framework. Parameters of the Actor and Critic networks are initialized orthogonally [46], ensuring the initial policy satisfies Gaussian distribution properties. The agent generates PID parameters K_p, K_i, K_d through the Actor network, with adaptive Gaussian noise superimposed to maintain exploratory behavior:

$$a_t \sim \mathcal{N}(\pi(\mu \Big| s_t), \sigma^2) \tag{12}$$

Traditional PPO algorithms typically employ fixed-decay exploration strategies, which cannot guarantee convergence. This paper proposes an adaptive exploration strategy based on advantage functions. When substantial policy advantage exists, it reduces randomness in action sampling to guarantee stable convergence to optimal policies. If the current strategy has strong advantages, reduce the randomness of exploration. This ensures that the algorithm converges to the optimal strategy.

$$\sigma_{e+1} = \sigma_e \cdot \left(1 - \zeta \cdot \frac{A(s_e, a_e)}{\max(A)}\right) \cdot \left(1 - \beta \cdot \frac{e}{E}\right) \tag{13}$$

In the formula, σ_e is the standard deviation of action sampling for the current episode, $A(s_e, a_e)$ is the dominant function value in state s_e.ζ, β is the relevant attenuation coefficient $(0 < \zeta, \beta < 1)$, used to control the influence of the number of training episodes on the standard deviation. E is the total number of training episodes or the preset upper limit of training times.

During this process, the interaction between the missile autopilot and the environment forms a closed-loop control loop. The PID parameter increment output by the intelligent agent is applied in real-time to the six degrees of freedom dynamic model of the missile, generating elevator deflection commands δ_z and driving flight trajectory adjustment. The observation includes the attitude angle error θ_e, altitude change rate $\hat{h}$, and other critical state variables at the next moment under environmental feedback. Next, the scalar reward signal will be calculated using the reward function designed in the previous section. After each training round, the system calculates discount returns $G_t = \sum_{k=0}^{\cdot} \gamma^k r_{t+k}$ and advantage functions $A_t = G_t - V_\phi(s_t)$ based on accumulated rewards, Where $V_\phi(s_t)$ is the state value estimated by the critic network, γ^k represents the discount factor and r_{t+k} Instant reward obtained from time step t.

For the process of strategy optimization, the objective function of TF-PPO algorithm is designed as:

$$\mathcal{L}_{\text{actor}} = \mathbb{E}_t \left[\min \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} A_t, \text{clip} \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}, 1-\varepsilon, 1+\varepsilon \right) A_t \right) \right] \tag{14}$$

The actor network employs a clipped policy ratio mechanism to constrain parameter updates and prevent abrupt policy shifts. Simultaneously, the critic network updates by minimizing value estimation errors:

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_t [(V_\phi(s_t) - G_t)^2] \tag{15}$$

During algorithm training, reinforcement learning agents exhibit high sensitivity to training data quality in their parameter optimization. When suboptimal control trajectories are significantly present in the experience replay buffer, agents risk acquiring detrimental control patterns through policy gradient updates, leading to convergence toward suboptimal solutions or even divergence. This study incorporates dynamic early termination (DET) into the training process, enhancing operational robustness to mitigate this vulnerability.

The early termination criterion is mathematically defined as

$$\mathcal{T}_{\text{stop}} = \begin{cases} 1, & if |\theta_e| > \theta_{th} \wedge \theta_e \cdot q_e > 0 \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

When $\mathcal{T}_{\text{stop}} = 1$, terminate the current training round and adjust the termination time reward to

$$R_{\text{term}} = R_{\text{base}} - \lambda R_{\text{fixed}} \tag{17}$$

where R_{base} denotes the raw reward value at termination, $\lambda > 1$ represents the penalty coefficient, and R_{fixed} is a predefined constant penalty term. This design ensures training stability through policy exploration constraints and bounded reward scaling, achieving a 62% reduction in reward variance, as shown in Figure 4.

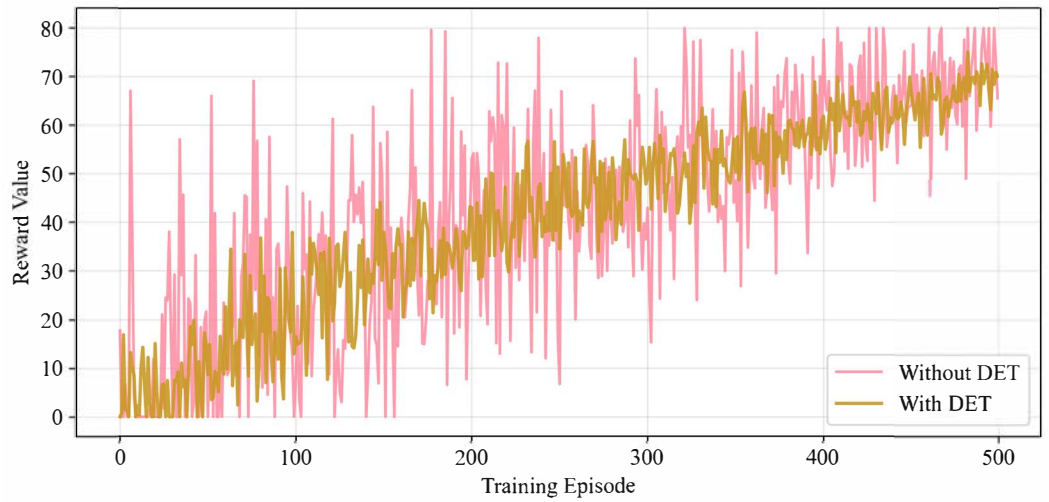


Figure 4. DET mechanism reduces training reward variance by 62%.

The pseudocode for the TF-PPO algorithm procedure is presented in Algorithm 1.

Algorithm 1: TF-PPO

Initialize policy network π and value network V
Set hyperparameters: learning rate α, discount factor γ, clip parameter ε, episodes E, update interval U
for $i \in \{1, \cdots, N\}$ **do**
  for $j \in \{1, \cdots, M\}$ **do**
  Run policy π_θ, collecting $\{s_t, a_t, r_t\}$
  Check termination condition via DET
  Estimate discountreturns G_t and advantages $A_t = G_t - V_\phi(s_t)$
  Decay action selection noise:
    $\sigma_{e+1} = \sigma_e \cdot \left(1 - \zeta \cdot \frac{A(s_e, a_e)}{\max(A)}\right) \cdot \left(1 - \beta \cdot \frac{e}{E}\right)$
  if $j\%T = 0$ **then**
   Update π_θ according to the objective function
    $$\mathcal{L}_{\text{actor}} = \mathbb{E}_t\left[\min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} A_t, \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}, 1-\varepsilon, 1+\varepsilon\right) A_t\right)\right]$$
  end if
 end for
end for

4. Experimentation and Evaluation

To validate the performance of the proposed TF-PPO algorithm, this section presents comprehensive mathematical simulations and hardware-in-the-loop (HIL) experiments. The validation is conducted in two parts: First, mathematical simulations using step response prove that the TF-PPO algorithm offers faster convergence and better performance than other RL algorithms under large step signal excitation. Second, a missile control rapid prototyping system is designed. The attitude tracking performance of TF-PPO and traditional PID methods is evaluated under different initial conditions and wind disturbances. This verifies the robustness and self-evolution capability of parameters of the proposed method.

4.1. Step Response Experiment

To evaluate the core contribution of the Temporal Feature Fusion mechanism, this part of the experiment uses a simplified reward function $r_t = -\theta_e$ to compare TF-PPO with other algorithms. First, a 6-DOF mathematical simulation model of the missile is constructed. Then, different RL algorithms are connected to this mathematical model. We designed a fixed scenario where a large step command is given to the control system every 15 s. Using this scenario, each algorithm is trained for 5000 episodes. The convergence effectiveness of the different algorithms is assessed.

This paper uses three algorithms, TF-PPO, PPO, and DDPG, for comparative step response experiments. Table 4 describes the relevant algorithm training parameters. The PPO algorithm uses the same training parameters. For DDPG, the parameter τ is set to 0.005, while other parameters remain consistent. The training results are shown in Figure 5, where a moving average is applied to the results. The shaded area represents the variance of the reward.

From the results, it can be seen that the TF-PPO and PPO algorithms, due to the existence of the clip method, avoid drastic changes in strategy and have relatively stable training processes. Compared to the classical PPO algorithm, TF-PPO exhibits faster convergence speed and learns better results. The trained models are reconnected to the simulation system to compare the effects of the different algorithms.

Table 4. TF-PPO hyperparameter settings.

Symbols	Parameter Explanation	Reference Range	Selected Value
ε	Clip ratio (limits policy update magnitude)	0.1~0.3	0.1
η_{action}	Actor learning rate	0.00001~0.001	0.00005
η_{critic}	Critic learning rate	0.00001~0.001	0.0001
γ	Discount factor	0.8~0.999	0.99
N	Sample size per training step	64~512	128
K	Optimization passes per batch	1~100	10
T	Update interval steps	10~10,000	500

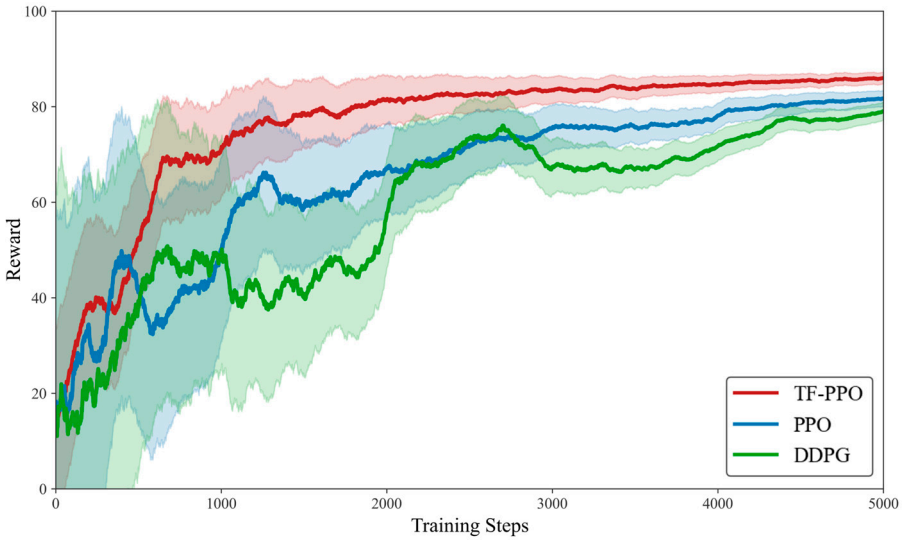


Figure 5. Comparison of Algorithm Convergence.

The specific effects are shown in Figure 6. The control accuracy of the TF-PPO algorithm is superior to the other algorithms. Compared with the traditional PPO algorithm, TF-PPO achieves an average control accuracy improvement of 14.3% with similar settling time, as summarized in Table 5. This demonstrates that the proposed algorithm has better training effectiveness for the missile guidance and control system.

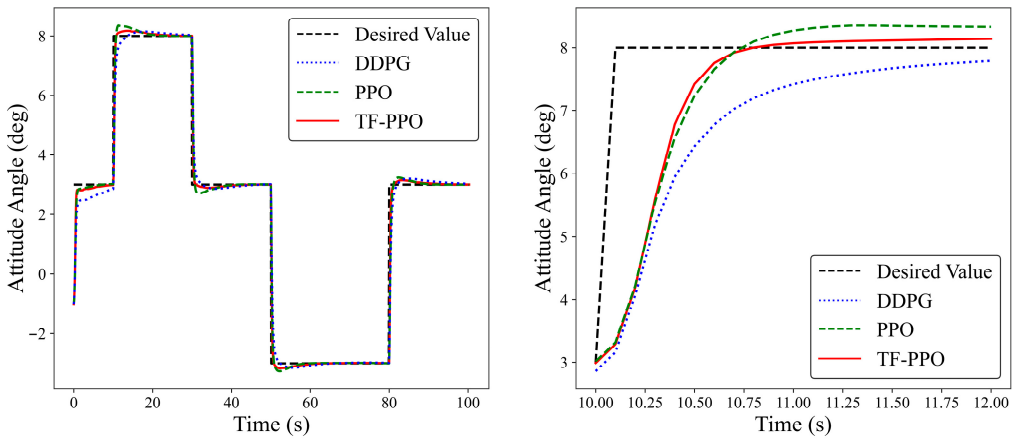


Figure 6. Comparison of response performance under step signal excitation: (**a**) attitude angles under different algorithms; (**b**) detailed drawing.

Table 5. Numerical simulation results.

Algorithm	$\bar{e}$	σ_e^2	$\bar{t}_s$
PPO	0.1487	0.5297	0.6627
DDPG	0.2091	0.5675	1.5323
TF-PPO	0.1274	0.4998	0.6250

4.2. Offline Training

The online optimization of the TF-PPO algorithm relies on high-quality offline training. Although TF-PPO is an on-policy algorithm, its missile-borne deployment on cruise missiles requires training the network parameters within the ground segment beforehand. Therefore, constructing a training environment that approximates real missile application conditions is essential. We use the AFSIM simulation platform as the training environment, deploy our six degrees of freedom missile model in AFSIM, and train the missile reinforcement learning control system by designing different initial conditions and flight plans.

This experiment constructed a cruise trajectory plan based on a simulation platform, which includes an acceleration phase, an altitude adjustment phase, a cruise phase, a descent phase, and a terminal guidance phase. The missile was launched by air launch, and Figure 7 shows a typical trajectory plan for missile flight. Based on typical trajectory schemes, the initial launch conditions and cruising altitude for each training round are determined through uniformly distributed random sampling, and the missile control system network is trained accordingly.

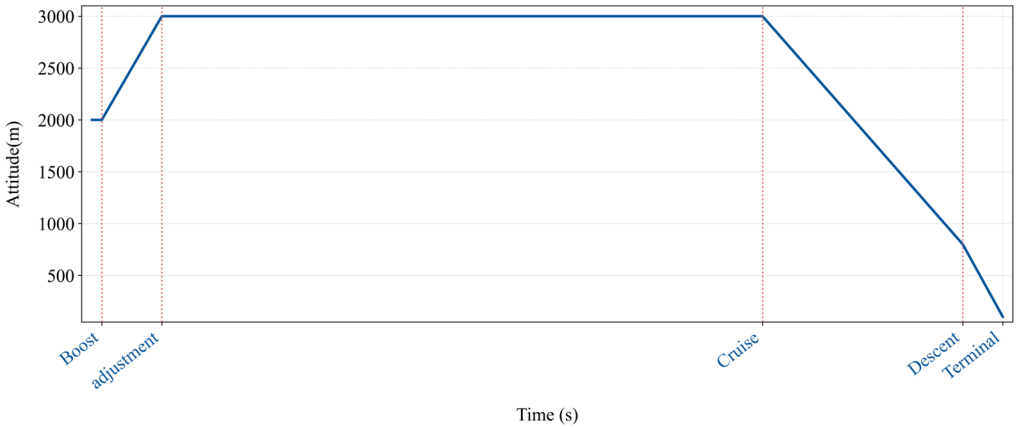


Figure 7. Cruise Missile Planned Trajectory.

Table 6 describes the specific parameter design of offline training parameters. The TF-PPO algorithm was trained using an adaptive reward function, and the model training hyperparameters were consistent with Table 4. After completing offline training of the algorithm, the trained algorithm was reconnected to a typical trajectory through a simulation system for full trajectory flight simulation. The simulation results are shown in Figure 8.

Table 6. Random Sampling Parameters for Offline Training.

Names	Values	Units
Training epochs	5000	-
Launch height	1000–5000	m
Launch elevation angle	−5~5	°
Initial Mach number	0.5~0.7	-
Cruising altitude	3000~5000	m

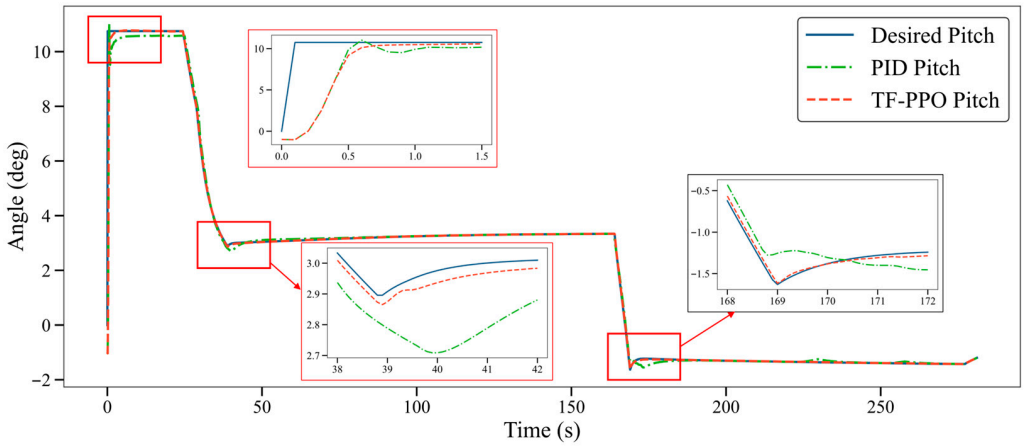


Figure 8. Performance Comparison of Trained TF-PPO Algorithm vs. PID Control.

The TF-PPO algorithm trained offline can automatically adjust the PID parameters of the pitch loop without prior knowledge. The results show that the missile pitch loop attitude control system under the TF-PPO algorithm exhibits good tracking performance, and in the stage transition stage, the algorithm outperforms the PID control system based on static design. Figure 9 illustrates the dynamic adjustment of PID parameters by RL-PID during missile flight. The PID parameters will be adaptively adjusted according to the stage characteristics.

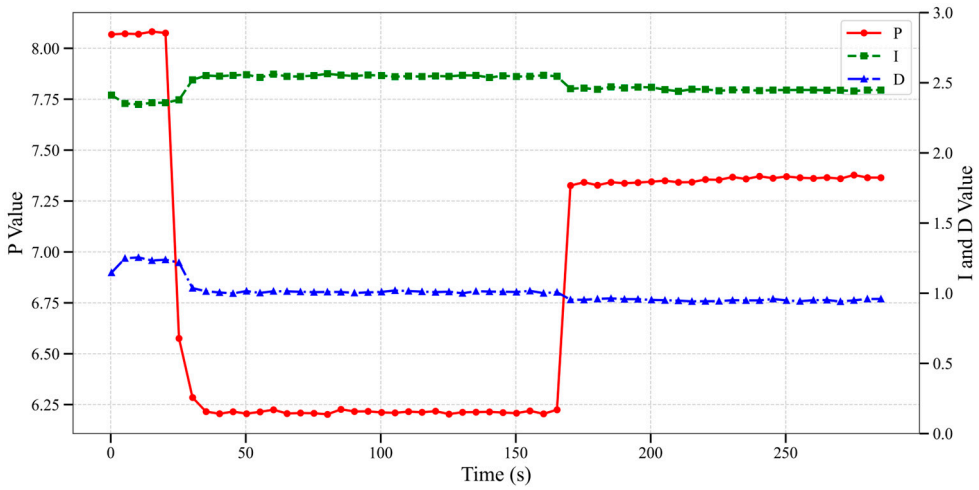


Figure 9. Response of PID Parameters Under TF-PPO Control.

4.3. Rapid Prototyping System

To validate the performance of the RL-PID algorithm in practical control, this study constructed a missile rapid prototyping system consisting of a simulation computer, missile-borne mission computer, and simulation platform, as shown in Figure 10. Hardware-in-the-loop (HIL) simulation was implemented to achieve closed-loop verification of multi-phase trajectories.

Rapid prototyping is a prototype development methodology that focuses on creating prototypes early in the development cycle [47]. This enables early feedback and analysis to support the development process. The rapid prototyping system described in this paper integrates three components: a Simulation Control Station for initial parameter configuration and mission programming that exchanges data bidirectionally with the core systems; a Missile-Borne Mission Processor (MBMP) implementing DSP + FPGA architecture to emulate physical missile control hardware, where this embedded processor executes

the full missile control algorithm suite by receiving real-time missile states, computing aerodynamic surface commands, and transmitting control signals; and a Virtual Combat Environment leveraging the U.S. Air Force Research Laboratory's AFSIM [48] platform, within which our six-degree-of-freedom missile model operates to execute received control commands, simulate adversarial engagements, and provide continuous state feedback. Table 7 describes the hardware configuration of MBMP.

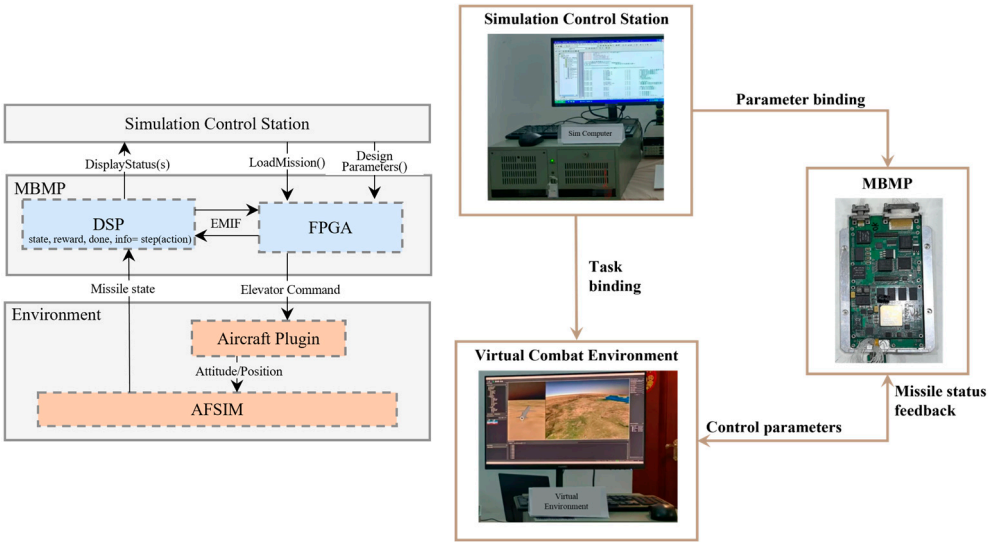


Figure 10. Rapid-prototyping missile control system: (**a**) Hardware Architecture; (**b**) Physical Implementation Diagram.

Table 7. Hardware configuration of MBMP.

Component	Values
DSP (TMS320C6678)	8-core, 1.25 GHz clock speed, 4 GB DDR3
FPGA (Xilinx Kintex-7 XC7K325T)	326 K logic cells, 1.4 Gb/s transceiver speeds
Gigabit Ethernet	100 Mb/s bidirectional data exchange
LVDS Serial Links	End-to-end latency < 0.5 ms

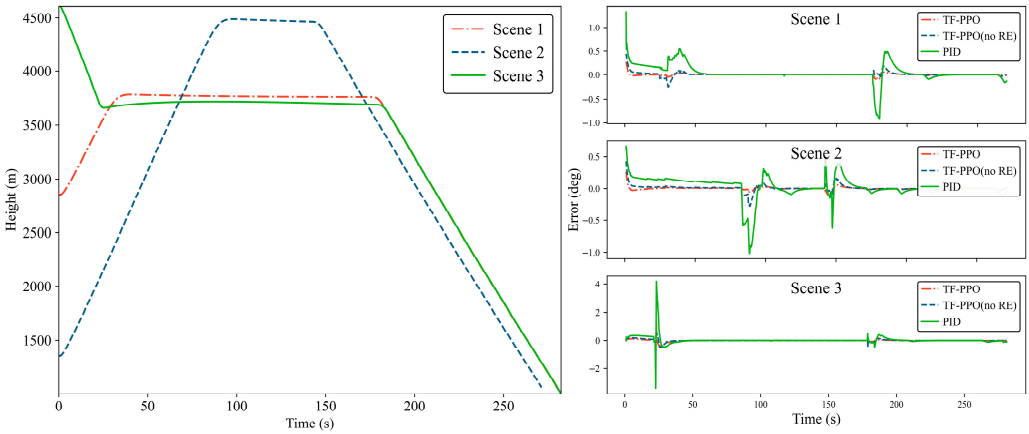
To validate the control algorithm's adaptability to various flight profiles, multiple distinct planned trajectories were randomly generated within the parameter boundaries specified in Table 6 for simulation. Concurrently implemented two control groups demonstrate algorithm and reward engineering advantages: the first group utilizes a manually tuned PID controller with fixed gains for standard flight profiles, and the second group employs a TF-PPO algorithm with $r_t = -\theta_e$ as its reward function (also trained offline for 5000 iterations under identical conditions; no reward engineering, labeled no RE in figures). Both groups were validated under identical randomized initial condition configurations as the experimental group.

The experimental framework was designed around a simulated test scenario featuring an air-launched cruise missile deployed by a fighter aircraft. The missile's objective is to strike a designated maritime target in a simulated maritime environment.

Due to different control strategies employed in the terminal guidance phase, data collection was limited to the launch-to-descent phase. Statistical results from all experiments are shown in Table 8. Representative experimental runs are depicted in Figure 11, which illustrates the missile flight altitude and attitude angle error.

Table 8. Pitch Angle Error Statistics.

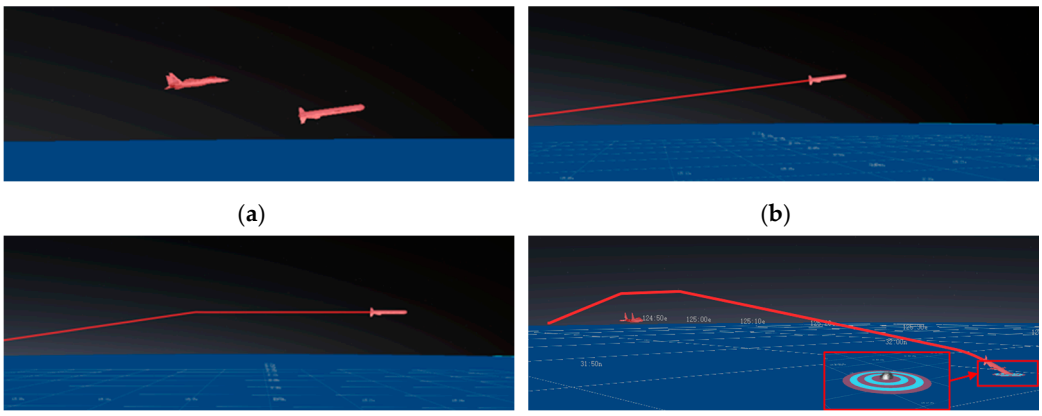
Evaluation Scope	Control Method	Mean (95% CI) (°)	Std (°)	Max (°)
Entire trajectory	PID	0.0212 (0.0186–0.0234)	0.1215	4.9703
	TF-PPO (no RE)	0.0144 (0.0131–0.0162)	0.0494	2.9721
	TF-PPO	0.0135 (0.0126–0.0151)	0.0235	1.4233
Transition phase	PID	0.1204 (0.0913–0.1495)	0.1701	4.9703
	TF-PPO (no RE)	0.0667 (0.0561–0.0783)	0.0853	2.9721
	TF-PPO	0.0528 (0.0421–0.0615)	0.0690	1.4233



(**a**) (**b**)

Figure 11. Representative Results: (**a**) Missile flight altitude; (**b**) Attitude angle error of experiments.

Figure 12 shows simulation snapshots capturing key moments of the engagement scenario within the AFSIM environment. The results indicate that the TF-PPO algorithm demonstrates significant improvement over the empirically tuned PID controller, achieving a 36.3% increase in pitch angle control accuracy across the entire trajectory. Analysis of maximum values reveals that control errors predominantly concentrate during transitional phases. During these transitional phases, compared to TF-PPO without DET, the TF-PPO algorithm reduced the maximum attitude tracking error by 52.1%.



(**a**) (**b**)

(**c**) (**d**)

Figure 12. Simulation snapshots capturing key moments of this engagement scenario: (**a**) Missile air launch; (**b**) Missile climb; (**c**) Missile enters fixed altitude cruise; (**d**) Missile hits target.

Furthermore, reward engineering constraints on elevator deflection prevent excessive PID adjustments, reducing control error variance by 54% throughout the flight. These

findings demonstrate the practical utility and robustness of the reward engineering-based TF-PPO algorithm.

To further validate the disturbance rejection capability of the TF-PPO-controlled longitudinal channel, wind disturbance profiles were applied during altitude adjustment, cruise, and descent phases in typical simulation scenarios, verifying the algorithm's attitude stabilization performance under external disturbances. Table 9 details the design matrix for wind disturbance variations.

Table 9. Wind Disturbance Variation Matrix.

Duration (s)	Steady-State Wind Bias (East, North, Up; m/s)	Gust Wind Bias (East, North; m/s)
20~40	(-, -, -)	(20, -)
20~60	(-, -, −6)	(-, -)
120~140	(-, -, -)	(20, 20)
150~200	(-, -, 10)	(-, -)
370~390	(-, -, -)	(-, 20)
390~430	(-, -, 8)	(-, -)

The results demonstrate that TF-PPO's adaptable reward mechanism effectively mitigates attitude fluctuations induced by external disturbances. Under wind disturbances, the TF-PPO algorithm exhibits significantly lower average error than conventional PID control, as quantitatively compared in Table 10 and Figure 13. Compared to the version without reward engineering, the complete TF-PPO algorithm achieves 31.6% higher comprehensive accuracy, confirming its practical value for engineering implementation in complex battlefield environments.

Table 10. Error Statistics Comparison Under Wind Disturbance Variation Tests.

Evaluation Scope	Control Method	Mean	Std	Max
Wind disturbance period	PID	0.2489	0.2202	0.8249
disturbance	TF-PPO (no RE)	0.0569	0.0474	0.5622
period	TF-PPO	0.0389	0.0359	0.2628

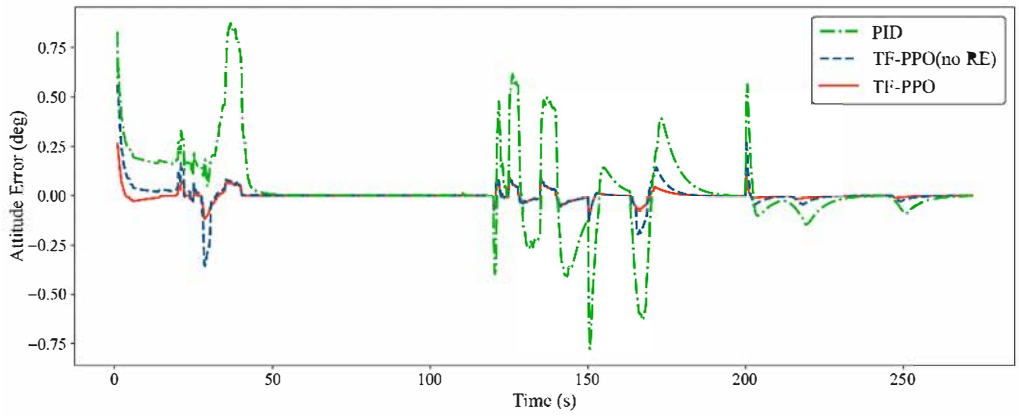


Figure 13. Error Comparison Under Wind Disturbance Variation Tests.

5. Conclusions

(1) An expert-free RL-PID architecture is proposed for missile control parameter self-evolution through autonomous online optimization, eliminating manual tuning dependency during complex flight maneuvers while ensuring real-time adaptation across all mission phases.

(2) The TF-PPO framework incorporating LSTM networks to enhance reinforcement learning adaptability for PID parameter tuning, where strategic reward engineering and adaptive exploration strategies effectively identify optimal parameters across distinct missile flight phases.
(3) A modular rapid-prototyping missile control platform integrates hardware-in-loop simulation with combat-realistic environments, enabling direct validation of controller performance under stochastic disturbances through rapid iteration capability.
(4) Established quantifiable verification methodology combining step-response analysis and rapid prototyping experiments objectively evaluates transient response improvement and disturbance rejection superiority over conventional methods.

6. Limitations and Future Work

Our study has several limitations: the system modeling overlooks the coupling effects between different attitude channels, potentially leading to deviations from actual missile dynamics; the application of reinforcement learning in missile control requires stronger interpretability to meet the high safety standards of the field; and the proposed method lacks comprehensive comparisons with state-of-the-art algorithms. To develop more reliable and practical solutions, our future research will prioritize creating a unified framework for the simultaneous and coordinated optimization of all three attitude channels.

Author Contributions: Conceptualization, C.T. and B.Z.; methodology, C.T. and B.Z.; software, C.T. and H.C.; validation, C.T., J.W. and H.C.; formal analysis, C.T. and S.H.; investigation, C.T. and J.W.; data curation, C.T., J.W. and B.Z.; writing—original draft preparation, C.T.; writing—review and editing, J.W. and W.Z.; visualization, C.T.; supervision, B.Z.; project administration, B.Z. and W.Z.; funding acquisition, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Grant No. 52202513) and Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515010023).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Krieger, R.J. Supersonic Missile Aerodynamic and Performance Relationships for Long Range Mission Profiles. *J. Spacecr. Rocket.* **1984**, *21*, 234–240. [CrossRef]
2. Hicks, S. Advanced cruise missile guidance system description. In Proceedings of the IEEE 1993 National Aerospace and Electronics Conference-NAECON 1993, Dayton, OH, USA, 24–28 May 1993.
3. Siouris, G.M. *Missile Guidance and Control Systems*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
4. Wu, F.; Packard, A.; Balas, G. Systematic gain-scheduling control design: A missile autopilot example. *Asian J. Control* **2002**, *4*, 341–347. [CrossRef]
5. Jang, J.; Alaniz, A.; Hall, R.; Bedrossian, N.; Hall, C.; Ryan, S.; Jackson, M. Ares i flight control system design. In Proceedings of the AIAA Guidance, Navigation, and Control Conference, Toronto, ON, Canada, 2–5 August 2010.
6. Golestani, M.; Mohammadzaman, I. PID guidance law design using short time stability approach. *Aerosp. Sci. Technol.* **2015**, *43*, 71–76. [CrossRef]
7. Zhao, C.; Guo, L. Towards a theoretical foundation of PID control for uncertain nonlinear systems. *Automatica* **2022**, *142*, 110360. [CrossRef]
8. Kada, B. A new methodology to design sliding-pid controllers: Application to missile flight control system. *IFAC Proc. Vol.* **2012**, *45*, 673–678. [CrossRef]
9. Tang, J.; Hu, Y.A.; Xiao, Z.; Li, J. Missile PID controller parameter tuning based on iterative learning control. In Proceedings of the 2010 2nd International Conference on Signal Processing Systems, Dalian, China, 5–7 October 2010.
10. Fawzy Ahmed, M.; Dorrah, H.T. Design of gain schedule fractional PID control for nonlinear thrust vector control missile with uncertainty. *Autom. Časopis Autom. Mjer. Elektron. Račun. Komun.* **2018**, *59*, 357–372.

11. Anusha, S.; Karpagam, G.; Bhuvaneswarri, E. Comparison of tuning methods of PID controller. *Int. J. Manag. Inf. Technol. Eng.* **2014**, *2*, 1–8.
12. Gonsalves, P.G.; Caglayan, A.K. Fuzzy logic PID controller for missile terminal guidance. In Proceedings of the Tenth International Symposium on Intelligent Control, Monterey, CA, USA, 27–29 August 1995.
13. Narendra, K.; Valavani, L. Stable adaptive controller design--Direct control. *IEEE Trans. Autom. Control* **1978**, *23*, 570–583. [CrossRef]
14. Landau, Y.D. Adaptive Control—The Model Reference Approach. *IEEE Trans. Syst. Man Cybern.* **1984**, *SMC-14*, 169–170. [CrossRef]
15. Hang, C.C.; Åström, K.J.; Ho, W.K. Refinements of the Ziegler–Nichols tuning formula. *IEE Proc. D (Control Theory Appl.)* **1991**, *138*. [CrossRef]
16. Mahdianfar, H.; Prempain, E. Adaptive augmenting control design for a generic longitudinal missile autopilot. In Proceedings of the 2016 American Control Conference (ACC), Boston, MA, USA, 6–8 July 2016.
17. Bekhiti, B.; Fragulis, G.F.; Hariche, K. A New 3D Sliding Pursuit Guidance Law for Fixed Wing Combat Drone Piloting: Application to El-Djazaïr 54. *Unmanned Syst.* **2025**, 1–27. [CrossRef]
18. Bekhiti, B.; Fragulis, G.F.; Rahmouni, M.; Hariche, K. A Novel Three-Dimensional Sliding Pursuit Guidance and Control of Surface-to-Air Missiles. *Technologies* **2025**, *13*, 171. [CrossRef]
19. Ferro, C.; Cafaro, M.; Maggiore, P. Optimizing Solid Rocket Missile Trajectories: A Hybrid Approach Using an Evolutionary Algorithm and Machine Learning. *Aerospace* **2024**, *11*, 912. [CrossRef]
20. Bai, Y.; Zhou, D.; He, Z. Optimal Pursuit Strategies in Missile Interception: Mean Field Game Approach. *Aerospace* **2025**, *12*, 302. [CrossRef]
21. Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; Riedmiller, M. Deterministic policy gradient algorithms. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
22. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347. [CrossRef]
23. Schulman, J. Trust Region Policy Optimization. *arXiv* **2015**, arXiv:1502.05477.
24. Borase, R.P.; Maghade, D.K.; Sondkar, S.Y.; Pawar, S.N. A review of PID control, tuning methods and applications. *Int. J. Dyn. Control* **2021**, *9*, 818–827. [CrossRef]
25. An, W.; Wang, H.; Sun, Q.; Xu, J.; Dai, Q.; Zhang, L. A PID controller approach for stochastic optimization of deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
26. Carlucho, I.; De Paula, M.; Acosta, G.G. An adaptive deep reinforcement learning approach for MIMO PID control of mobile robots. *ISA Trans.* **2020**, *102*, 280–294. [CrossRef] [PubMed]
27. Zhou, Z.; Lu, Y.; Kokubu, S.; Tortós, P.E.; Yu, W. A GAN based PID controller for highly adaptive control of a pneumatic-artificial-muscle driven antagonistic joint. *Complex Intell. Syst.* **2024**, *10*, 6231–6248. [CrossRef]
28. Li, S.; Liu, T.; Zhang, C.; Yeung, D.-Y.; Shen, S. Learning unmanned aerial vehicle control for autonomous target following. *arXiv* **2017**, arXiv:1709.08233. [CrossRef]
29. Lin, C. Adaptive critic autopilot design of bank-to-turn missiles using fuzzy basis function networks. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2005**, *35*, 197–207. [CrossRef] [PubMed]
30. Zhao, X.; He, L.; Liu, X.; Han, K.; Li, J. A Novel Reinforcement Learning Framework for Optimizing Fixed-Wing UAV Flight Control Strategies. *Aerosp. Sci. Technol.* **2025**, 110512. [CrossRef]
31. Hong, D.; Park, S. Avoiding obstacles via missile real-time inference by reinforcement learning. *Appl. Sci.* **2022**, *12*, 4142. [CrossRef]
32. Yan, M.; Yang, R.; Zhang, Y.; Yue, L.; Hu, D. A hierarchical reinforcement learning method for missile evasion and guidance. *Sci. Rep.* **2022**, *12*, 18888. [CrossRef]
33. Zhang, W.; Chen, G.; Ni, H.; Tong, T.; Zhou, Y. Design of missile longitudinal controller based on deep deterministic policy gradient learning algorithm. In Proceedings of the 2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 15–17 September 2023.
34. Lee, Y.; Park, J.; Kim, Y. High Angle of Attack Missile Control for Agile Turn Based on Reinforcement Learning. In Proceedings of the AIAA SCITECH 2024 Forum, Orlando, FL, USA, 8–12 January 2024.
35. Dewey, D. Reinforcement Learning and the Reward Engineering Principle. 2014. Available online: https://cdn.aaai.org/ocs/7704/7704-34364-1-PB.pdf (accessed on 20 October 2024).
36. Koch, W.; Mancuso, R.; West, R.; Bestavros, A. Reinforcement learning for UAV attitude control. *ACM Trans. Cyber-Phys. Syst.* **2019**, *3*, 1–21. [CrossRef]
37. Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.

38. Fleeman, E.L. Tactical Missile Design. 2006. Available online: https://www.callcenterweekawards.com/media/6571/3122.pdf (accessed on 17 September 2025).
39. Stevens, B.L.; Lewis, F.L.; Johnson, E.N. *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
40. Cook, M.V. *Flight Dynamics Principles: A Linear Systems Approach to Aircraft Stability and Control*; Butterworth-Heinemann: Oxford, UK, 2012.
41. Wise, K.A.; Sedwick, J.L.; Eberhardt, R.L. *Nonlinear Control of Missiles*; Defense Technical Information Center: Fort Belvoir, VA, USA, 1995.
42. Kang, S.; Kim, H.J.; Lee, J.-I.; Jun, B.-E.; Tahk, M.-J. Roll-Pitch-Yaw Integrated Robust Autopilot Design for a High Angle-of-Attack Missile. *J. Guid. Control Dyn.* **2009**, *32*, 1622–1628. [CrossRef]
43. Gustafson, D.E.; Baillieul, J.; Levi, M. Nonlinear Control Theory for Missile Autopilot Design. In Proceedings of the 1987 American Control Conference, Minneapolis, MN, USA, 10–12 June 1987; pp. 43–49.
44. Jelali, M.; Kroll, A. *Hydraulic Servo-Systems: Modelling, Identification and Control*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
45. Konda, V.; Tsitsiklis, J. Actor-critic algorithms. Advances in Neural Information Processing Systems, Denver, CO, USA, 1999. Available online: https://papers.nips.cc/paper/1786-actor-critic-algorithms (accessed on 17 September 2025).
46. Huang, W.; Du, W.; Da Xu, R.Y. On the neural tangent kernel of deep networks with orthogonal initialization. *arXiv* **2020**, arXiv:2004.05867.
47. Pan, B.; Tang, S.; Wie, B. Rapid Prototyping of a Guidance and Control System for Missiles. AIAA Guidance, Navigation and Control Conference and Exhibit, Honolulu, HI, USA, 18–21 August 2008. Available online: https://www.researchgate.net/publication/268571351_Rapid_Prototyping_of_a_Guidance_and_Control_System_for_Missiles (accessed on 17 September 2025).
48. Zhang, L.A.; Xu, J.; Gold, D.; Hagen, J.; Kochhar, A.; Lohn, A.; Osoba, O. Air Dominance Through Machine Learning: A Preliminary Exploration of Artificial Intelligence–Assisted Mission Planning. 2020. Available online: https://www.rand.org/pubs/research_reports/RR4311.html (accessed on 17 September 2025).