aerospace

MDPI

Article

Fault-Tolerant Attitude Control of Flexible Spacecraft via Reinforcement Learning

Zhuoyue Peng and Qiang Shen *

School of Aeronautics and Astronautics, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China; pengzhouyue@sjtu.edu.cn
* Correspondence: qiangshen@sjtu.edu.cn

Abstract

This paper proposes an integrated attitude control framework for flexible spacecraft subject to external disturbances, rigid–flexible dynamic coupling, and actuator faults. The control framework combines the Twin Delayed Deep Deterministic Policy Gradient (TD3) reinforcement learning algorithm with an adaptive fault-tolerant (AFT) compensator. First, a rigid–flexible coupling dynamic model is formulated using Modified Rodrigues Parameters. Second, an observer-based TD3 attitude controller is designed, where a hierarchical reward function incorporating the observer-estimated flexible modal displacement $\hat{\boldsymbol{\eta}}$ is constructed to train the agent for simultaneous attitude convergence and vibration suppression. Third, a composite fault-tolerant control structure is developed by integrating the trained TD3 policy with an adaptive sliding mode compensator that handles both partial loss-of-effectiveness faults and time-varying additive faults. The proposed framework is evaluated under a progressive five-scenario uncertainty evaluation framework encompassing measurement noise, parameter mismatch, external disturbances, and actuator faults. Simulation results demonstrate that (i) the $\hat{\boldsymbol{\eta}}$-augmented reward enables substantial improvements in vibration suppression over the baseline reward, achieving a better balance between pointing accuracy and vibration attenuation; (ii) under the most demanding fault scenario, the AFT compensator proves essential for precise convergence, and the composite TD3+AFT architecture achieves the best overall performance among the four compared control schemes.

Keywords: flexible spacecraft; fault-tolerant control; attitude control; reinforcement learning; TD3

Check for updates

Academic Editor: Vladimir S. Aslanov

Received: 7 June 2026
Revised: 19 June 2026
Accepted: 19 June 2026
Published: 24 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Flexible spacecraft attitude control represents a critical yet challenging problem in modern space missions, characterized by rigid–flexible coupling effects, multi-source uncertainties, and stringent performance requirements [1,2]. Unlike rigid spacecraft, flexible spacecraft feature large lightweight appendages (such as solar arrays and antenna reflectors) that introduce low-frequency elastic vibrations intrinsically coupled with rigid-body attitude motion, degrading pointing accuracy and risking persistent oscillations during maneuvers [2].

Recent advances have addressed flexible vibration suppression, disturbance rejection, actuator fault tolerance, and agile maneuvering through disturbance observers, sliding mode control (SMC), input shaping, and prescribed performance control. Observer-based and SMC approaches include Golestani et al.'s fixed-time disturbance observer with

prescribed performance control [1], Shahid et al.'s composite anti-unwinding finite-time adaptive sliding mode with a modal vibration observer [3], and Dian et al.'s scheduled input shaping with sinusoidal path planning and a nonlinear disturbance observer [4]. For synchronous maneuver and vibration suppression, He and Cao designed cooperative linear quadratic regulator (LQR) and proportional-derivative (PD) with input shaping [2], Zhong et al. employed the internal model principle to asymptotically reject sinusoidal disturbances without modal measurements [5], and Song and Agrawal experimentally validated pulse width pulse frequency (PWPF) modulation with positive position feedback to attenuate thruster-induced vibrations [6]. Robust and adaptive approaches addressing model uncertainties include Yu et al.'s gain-scheduled H_∞ control for large-angle maneuvers [7], Angeletti et al.'s end-to-end robust design via linear fractional transformation [8], and Khoroshylov et al.'s adaptive notch-filter strategies for space-based synthetic aperture radar (SAR) platforms [9].

Deep reinforcement learning (DRL) has emerged as a broader trend in spacecraft guidance, navigation, and control, as surveyed in recent reviews [10–12]. Among DRL algorithms, the Twin Delayed Deep Deterministic policy gradient (TD3) [13] is selected in this work for its deterministic policy formulation well-suited for continuous torque commands, off-policy sample efficiency critical for computationally expensive flexible-body simulations, and twin-critic architecture that mitigates value overestimation bias. TD3 has been extensively applied across aerospace scenarios including unmanned aerial vehicle (UAV) operations [14–18], spacecraft attitude control [19], hypersonic vehicle maneuvering [20], and aero-engine regulation [21], with algorithmic modifications such as proportional-integral-derivative (PID)-guided exploration [19], Steffensen value iteration (SVI) for faster convergence [22], frame stacking for energy-efficient UAV path planning [14], and 3D environment encoding for dynamic obstacle avoidance [16].

More recently, several works have explored the intersection of TD3/DRL with flexible spacecraft dynamics and fault-tolerant control. Shi et al. [23] applied TD3 to cooperative control of ultra-close flexible spacecraft formation flying. Li et al. [24] integrated a DRL-based disturbance observer with the fully-actuated system approach for dumbbell-shaped flexible spacecraft. In the direction of fault tolerance, El-Dalahmeh et al. [25] developed TD3-HD with Hindsight Experience Replay for attitude fault recovery, and Henna et al. [26] introduced expert-guided exploration for satellite fault-tolerant attitude control. Nevertheless, two gaps remain to be addressed. First, Shi et al. [23] and Li et al. [24] address flexible attitude control without considering actuator faults. Second, El-Dalahmeh et al. [25] and Henna et al. [26] address fault-tolerant control but treat the spacecraft as rigid, neglecting rigid–flexible coupling dynamics. To the best of the authors' knowledge, no existing work has simultaneously addressed flexible vibration suppression and actuator fault tolerance within a unified TD3-based framework.

Motivated by this gap, this paper develops an integrated attitude control framework for flexible spacecraft stabilization, capable of simultaneously addressing external disturbances, rigid–flexible coupling dynamics, and actuator faults. The framework employs the TD3 algorithm integrated with adaptive fault-tolerant control.The main contributions of this work are: (1) A hierarchical phased reward function for TD3-based flexible spacecraft attitude control. Unlike prior reward designs that rely solely on attitude error [27] or lack vibration-aware shaping [23,24], the proposed reward comprises four components (base, convergence, smoothness, and terminal) implementing a phased convergence strategy, with the convergence term augmented by observer-estimated flexible modal displacement $\hat{\boldsymbol{\eta}}$ to explicitly guide oscillation suppression. (2) A composite fault-tolerant control structure integrating the trained TD3 agent with an adaptive sliding mode fault-tolerant (AFT) compensator, extending the integral-type sliding mode framework of [28] from rigid to flexible

spacecraft. The sliding manifold and AFT compensator are reformulated with rigid–flexible coupling terms ($\boldsymbol{\delta}^T\ddot{\boldsymbol{\eta}}$ and $\boldsymbol{\delta}^T\dot{\boldsymbol{\eta}}$) using observer-estimated modal coordinates.

The source code and simulation scripts accompanying this work are openly available at https://github.com/EllencePeng/Aerospace-RL-Flexible-Spacecraft (accessed on 18 June 2026).

The organization of this paper is as follows: Section 1 presents the research background and motivation. Section 2 presents the flexible spacecraft dynamics, fault model, flexible state observer, and problem statement. Section 3 establishes the RL training framework. Section 4 designs the adaptive fault-tolerant controller. Section 5 verifies the proposed schemes via simulations. Section 6 concludes this paper by summarizing its main contributions and putting forward prospects for future research.

2. Preliminaries

2.1. Flexible Spacecraft Attitude Dynamics and Kinematics

Two reference frames are used throughout this paper: an inertial reference frame $\mathcal{I}$ and a body-fixed frame $\mathcal{B}$ attached to the spacecraft's main body. The angular velocity vector $\boldsymbol{\omega} \in \mathbb{R}^3$ represents the inertial angular velocity of the spacecraft with respect to $\mathcal{I}$ and expressed in $\mathcal{B}$. The attitude orientation of $\mathcal{B}$ with respect to $\mathcal{I}$ is described by the Modified Rodrigues Parameters (MRPs) [29], denoted by $\boldsymbol{p} \in \mathbb{R}^3$. The corresponding attitude kinematic model, derived by referring to the modeling method in [30], is given by

$$\dot{\boldsymbol{p}} = \frac{1}{4}\left[\left(1-\boldsymbol{p}^T\boldsymbol{p}\right)\boldsymbol{I}_3 + 2\boldsymbol{p}^\times + 2\boldsymbol{p}\boldsymbol{p}^T\right]\boldsymbol{\omega} = \boldsymbol{T}(\boldsymbol{p})\boldsymbol{\omega}, \tag{1}$$

where $\boldsymbol{p} \in \mathbb{R}^3$ represents the MRP vector, and $\boldsymbol{I}_3$ denotes the 3×3 identity matrix. The compact form $\boldsymbol{T}(\boldsymbol{p})\boldsymbol{\omega}$ is adopted for brevity in subsequent equations. The symbol $\boldsymbol{p}^\times$ denotes the skew-symmetric matrix of $\boldsymbol{p}$, defined as

$$\boldsymbol{p}^\times = \begin{bmatrix} 0 & -p_3 & p_2 \\ p_3 & 0 & -p_1 \\ -p_2 & p_1 & 0 \end{bmatrix}. \tag{2}$$

Compared with conventional attitude parameterizations, MRPs provide a computationally efficient and numerically robust attitude description [29], well-suited for precise spacecraft attitude control simulation. MRPs are globally valid for eigenaxis rotations up to 360°, but encounter a singularity where $\|\boldsymbol{p}\| \to \infty$ when the rotation angle reaches exactly 360°. This is readily handled by shadow-set switching. Since all maneuvers considered in this work involve rotation angles far below 360°, $\|\boldsymbol{p}\|$ remains well within the unit sphere and shadow-set switching is unnecessary here.

The attitude dynamics of flexible spacecraft is adopted from [31] in low Earth orbit (LEO), formulated as

$$\begin{cases} \boldsymbol{J}\dot{\boldsymbol{\omega}} + \boldsymbol{\delta}^T\ddot{\boldsymbol{\eta}} = -\boldsymbol{\omega}^\times\left(\boldsymbol{J}\boldsymbol{\omega} + \boldsymbol{\delta}^T\dot{\boldsymbol{\eta}}\right) + \boldsymbol{\tau} \\ \ddot{\boldsymbol{\eta}} + \boldsymbol{C}\dot{\boldsymbol{\eta}} + \boldsymbol{K}\boldsymbol{\eta} = -\boldsymbol{\delta}\dot{\boldsymbol{\omega}} \end{cases}, \tag{3}$$

where $\boldsymbol{J} \in \mathbb{R}^{3\times3}$ denotes the inertial matrix of the spacecraft, $\boldsymbol{\tau} \in \mathbb{R}^3$ is the control torque input, $\boldsymbol{\delta}$ represents the coupling matrix between the rigid main body and flexible appendages, and $\boldsymbol{\eta} \in \mathbb{R}^m$ denotes the flexible modal coordinates vector, treated as dimensionless. m is the total number of flexible vibration modes, $\boldsymbol{C} = \mathrm{diag}\{2\xi_i\omega_{n_i}, i = 1, ..., m\}$ and $\boldsymbol{K} = \mathrm{diag}\{\omega_{n_i}^2, i = 1, ..., m\}$ are the modal damping and stiffness matrices respectively, ω_{n_i} is the natural frequency of the i-th flexible mode, and ξ_i is the corresponding damping ratio.

2.2. Fault Model

Assume that the dynamic response characteristics of spacecraft actuators can be ignored. An actuator is defined as fault-free when its output signal is completely consistent with the input command. The typical actuator faults considered in this study are modeled based on the existing research framework as follows [28]:

$$\boldsymbol{\tau} = [\boldsymbol{I}_3 - \boldsymbol{E}(t)]\boldsymbol{u} + \bar{\boldsymbol{u}}, \tag{4}$$

where $\boldsymbol{E}(t) = \mathrm{diag}\{e_1(t), e_2(t), e_3(t)\} \in \mathbb{R}^{3\times 3}$ represents the actuator efficiency loss matrix of the spacecraft, satisfying $0 \leq e_i(t) < 1$ $(i = 1,2,3)$. Specifically, the condition $e_i(t) = 0$ means the i_{th} actuator operates in a normal state, while $0 < e_i(t) < 1$ indicates the i_{th} actuator suffers from partial efficiency loss without complete functional failure. Additionally, $\bar{\boldsymbol{u}} \in \mathbb{R}^3$ denotes the bounded time-varying additive fault occurring in the actuator. On this basis, the nonlinear attitude dynamic model of the spacecraft, which integrates the actuator fault model (4) and external disturbance $\boldsymbol{d}$, can be reconstructed as the following expression:

$$\boldsymbol{J}\dot{\boldsymbol{\omega}} + \boldsymbol{\delta}^T\ddot{\boldsymbol{\eta}} = -\boldsymbol{\omega}^{\times}\left(\boldsymbol{J}\boldsymbol{\omega} + \boldsymbol{\delta}^T\dot{\boldsymbol{\eta}}\right) + [\boldsymbol{I}_3 - \boldsymbol{E}(t)]\boldsymbol{u} + \bar{\boldsymbol{u}} + \boldsymbol{d}. \tag{5}$$

To lay a solid foundation for the design of the composite fault-tolerant controller, three reasonable assumptions are proposed for the actuator faults and external disturbances, as detailed below.

Assumption 1. *The external disturbance $\boldsymbol{d}$ is bounded, satisfying $\|\boldsymbol{d}\| \leq d_m$, where d_m is a positive scalar and $\|\cdot\|$ represents the Euclidean norm.*

Assumption 2. *For the actuator efficiency loss matrix $\boldsymbol{E}(t) = \mathrm{diag}\{e_1(t), e_2(t), e_3(t)\}$, the inequality $0 \leq \max\{e_1, e_2, e_3\} \leq e_m < 1$ holds, where e_m is a positive scalar.*

Assumption 3. *The time-varying additive actuator fault $\bar{\boldsymbol{u}}$ is bounded, satisfying $\|\bar{\boldsymbol{u}}\| \leq \bar{u}_m$, with $\bar{u}_m$ being a positive constant.*

Remark 1. *Assumptions 1–3 are consistent with the actual operating environment of most spacecraft systems. Assumption 1 states that the external disturbance $\boldsymbol{d}$ acting on the flexible spacecraft is norm-bounded, which is consistent with the fact that environmental perturbations (atmospheric drag, gravity gradient, and solar radiation pressure) are always limited in actual space missions. Assumption 2 describes the actuator partial loss of effectiveness fault, where the upper bound $0 \leq e_m < 1$ ensures that the actuator still retains a certain control ability rather than complete failure. Assumption 3 restricts the additive actuator fault to be bounded, reflecting that the deviation or bias caused by component aging, signal drift or mechanical wear is always finite in real physical systems. These physically meaningful assumptions provide a practical prerequisite for the design and analysis of the fault-tolerant control scheme.*

2.3. Flexible State Observer

In practice, the flexible modal variables $\boldsymbol{\eta}$ and $\dot{\boldsymbol{\eta}}$ are not directly measurable by onboard sensors. To address this, the nonlinear observer developed by Zou et al. [30] is adopted. The observer is of the high-gain type: it exploits the rigid–flexible coupling through the matrix $\boldsymbol{\delta}$ to infer the unmeasurable modal states from the available attitude measurement $\boldsymbol{p}$. The correction terms $\theta\theta_1\tilde{\boldsymbol{p}}$ and $\theta^2\theta_2\tilde{\boldsymbol{p}}$, driven by the attitude estimation error $\tilde{\boldsymbol{p}} = \boldsymbol{p} - \hat{\boldsymbol{p}}$, propagate through the coupled dynamics to update the estimates of $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{\eta}}$, and $\hat{\boldsymbol{\psi}}$. A single design parameter $\theta > 0$ governs the convergence rate; by choosing a

sufficiently large θ, the estimation errors $(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\psi}})$ are guaranteed to converge to zero exponentially. The complete observer dynamics are given by

$$\begin{cases} \dot{\hat{\boldsymbol{p}}} = \hat{\boldsymbol{v}} + \theta\theta_1\tilde{\boldsymbol{p}}, \\ \dot{\hat{\boldsymbol{v}}} = f(\boldsymbol{p}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\psi}}) + g\boldsymbol{\tau} + \theta^2\theta_2\tilde{\boldsymbol{p}}, \\ \dot{\hat{\boldsymbol{\eta}}} = \hat{\boldsymbol{\psi}} - \boldsymbol{\delta}\boldsymbol{P}\hat{\boldsymbol{v}}, \\ \dot{\hat{\boldsymbol{\psi}}} = -\boldsymbol{C}\hat{\boldsymbol{\psi}} - \boldsymbol{K}\hat{\boldsymbol{\eta}} + \boldsymbol{C}\boldsymbol{\delta}\boldsymbol{P}\hat{\boldsymbol{v}}, \end{cases} \tag{6}$$

where $\boldsymbol{v} = \dot{\boldsymbol{p}}$, $\boldsymbol{\psi} = \dot{\boldsymbol{\eta}} + \boldsymbol{\delta}\boldsymbol{\omega}$, $\boldsymbol{P} = \boldsymbol{T}^{-1}$, $g = \boldsymbol{T}\bar{\boldsymbol{J}}^{-1}$, $\bar{\boldsymbol{J}} = \boldsymbol{J} - \boldsymbol{\delta}^T\boldsymbol{\delta}$ and $f(\boldsymbol{p}, \boldsymbol{v}, \boldsymbol{\eta}, \boldsymbol{\psi}) = -\boldsymbol{T}\dot{\boldsymbol{P}}\boldsymbol{v} - \boldsymbol{T}\bar{\boldsymbol{J}}^{-1}(\boldsymbol{P}\boldsymbol{v})^{\times}\left(\bar{\boldsymbol{J}}\boldsymbol{P}\boldsymbol{v} + \boldsymbol{\delta}^T\boldsymbol{\psi}\right) + \boldsymbol{T}\bar{\boldsymbol{J}}^{-1}\boldsymbol{\delta}^T(\boldsymbol{C}\boldsymbol{\psi} + \boldsymbol{K}\boldsymbol{\eta} - \boldsymbol{C}\boldsymbol{\delta}\boldsymbol{P}\boldsymbol{v})$. θ, θ_1 and θ_2 are positive constants, $\hat{\boldsymbol{p}}$ with $\hat{\boldsymbol{p}}(0) = \boldsymbol{p}(0)$ and $\hat{\boldsymbol{v}}$ with $\hat{\boldsymbol{v}}(0) = 0$ are estimates of $\boldsymbol{p}$ and $\boldsymbol{v}$, respectively, $\tilde{\boldsymbol{p}} = \boldsymbol{p} - \hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{\psi}}$, $\hat{\boldsymbol{\eta}}$ are estimates of $\boldsymbol{\psi}$, $\boldsymbol{\eta}$. The observer gains in this work are set as $\theta = 5.0$, $\theta_1 = 2.0$, and $\theta_2 = 5.0$.

2.4. Problem Statement

Problem 1. *For a flexible spacecraft system modeled in as* (1) *and* (3)*, under concurrent external disturbances and actuator faults described by* (4)*, design a composite control scheme combining reinforcement learning and adaptive fault-tolerant control to achieve attitude stabilization and vibration suppression.*

3. TD3-Based Controller Design

This section formulates a baseline controller utilizing the TD3 algorithm for flexible spacecraft attitude stabilization under fault-free conditions. The TD3 algorithm is first briefly reviewed, followed by three design aspects that instantiate it for the problem at hand: the state and action representation, the reward function, and the network architectures.

3.1. TD3 Algorithm

TD3 is an off-policy actor–critic algorithm for continuous control, proposed by Fujimoto et al. [13] as an improvement over DDPG. It employs twin critic networks and delayed policy updates to mitigate value overestimation, and a deterministic policy that is well-suited for generating smooth continuous torque commands. The training procedure follows the standard TD3 protocol established in [13].

3.2. State and Action Representation

It is assumed that only the attitude and angular velocity of the flexible spacecraft are directly measurable, while the flexible modal displacement $\boldsymbol{\eta}$ is not directly measurable but can be estimated by the observer introduced in Section 2.3. Consequently, flexible modal coordinates are excluded from the state vector provided to the TD3 policy.

The state vector provided to the TD3 policy is defined as

$$\boldsymbol{s}_{RL} = \begin{bmatrix} \boldsymbol{p}^T & \boldsymbol{\omega}^T \end{bmatrix}^T \in \mathbb{R}^6, \tag{7}$$

The action generated by the policy corresponds to the continuous control torque applied to the spacecraft body. The action vector is defined as

$$\boldsymbol{a} = \boldsymbol{\tau} \in \mathbb{R}^3, \tag{8}$$

where $\boldsymbol{\tau}$ represents the control torque along the three principal body axes.

To ensure actuator feasibility, the control torque is subjected to saturation constraints,

$$\|\boldsymbol{\tau}\| \leq u_{\max}, \tag{9}$$

where $u_{\max} = 1\,\mathrm{N{\cdot}m}$ denotes the maximum allowable actuator torque, a value consistent with the torque capacity of reaction wheels typically employed on small-to-medium-class spacecraft.

3.3. Reward Function Design

The reward function implements a phased convergence strategy with four components:

$$R = R_{\text{base}} + R_{\text{convergence}} + R_{\text{smoothness}} + R_{\text{terminal}} \tag{10}$$

where the 1-norm of $\boldsymbol{p}$ is defined as $\|\boldsymbol{p}\|_1 = |p_1| + |p_2| + |p_3|$.

The base reward provides dense per-step signals during the large-error phase ($\|\boldsymbol{p}\|_1 \geq 0.5$):

$$R_{\text{base}} = \begin{cases} 1 - \|\boldsymbol{\tau}\|^2, & \text{if } \|\boldsymbol{p}\|_1 \geq 0.5 \text{ and } (\|\boldsymbol{p}\|_1 < \|\boldsymbol{p}_{\text{prev}}\|_1) \\ 1 - \|\boldsymbol{\tau}\|^2, & \text{if } \|\boldsymbol{p}\|_1 \geq 0.5 \text{ and } (\|\boldsymbol{\omega}\|_1 < \|\boldsymbol{\omega}_{\text{prev}}\|_1) \\ -\|\boldsymbol{p}\|_1 - 10\|\boldsymbol{\omega}\|_1 - \|\boldsymbol{\tau}\|^2, & \text{otherwise} \end{cases} \tag{11}$$

The piecewise monotonic-improvement condition rewards the agent with $1 - \|\boldsymbol{\tau}\|^2$ when the attitude or angular velocity error decreases, and penalizes all error sources otherwise. The coefficient 10 on $\|\boldsymbol{\omega}\|_1$ normalizes it to the scale of $\|\boldsymbol{p}\|_1$.

When $\|\boldsymbol{p}\|_1 < 0.5$, inverse-proportional rewards are activated to accelerate fine convergence:

$$R_{\text{convergence, 1}} = \begin{cases} \dfrac{1}{\|\boldsymbol{p}\|_1 + 0.01} + \dfrac{0.05}{\|\boldsymbol{\omega}\|_1 + 0.01}, & \|\boldsymbol{p}\|_1 < 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

The form $1/(\|\boldsymbol{p}\|_1 + 0.01)$ is chosen because its gradient steepens as the error shrinks, naturally accelerating fine convergence. The $+0.01$ offset prevents division by zero and caps the reward, avoiding destabilizing Q-value targets during critic training.

To explicitly suppress flexible vibrations, the observer-estimated modal displacement $\hat{\boldsymbol{\eta}}$ (Section 2.3) is incorporated into the convergence reward:

$$R_{\text{convergence, 2}} = \begin{cases} \dfrac{1}{\|\boldsymbol{p}\|_1 + 0.01} + \dfrac{0.05}{\|\boldsymbol{\omega}\|_1 + 0.01} + \dfrac{0.05}{\|\hat{\boldsymbol{\eta}}\|_2 + 0.01}, & \|\boldsymbol{p}\|_1 < 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

The additional term uses the same inverse-proportional form. The weight 0.05 (vs. 1.0 for attitude) encodes a 20:1 priority ratio, reflecting that vibration suppression is subordinate to attitude convergence. The agents trained with these two formulations—agent A_0 (without $\hat{\boldsymbol{\eta}}$) and agent A (with $\hat{\boldsymbol{\eta}}$)—are compared in Section 5.3.

Smoothness penalties discourage abrupt actuation:

$$R_{\text{smoothness}} = -10 \cdot \|\tau - \boldsymbol{\tau}_{\text{prev}}\|_1 + \begin{cases} -10 \cdot \|\boldsymbol{\tau}\|^2, & \|\boldsymbol{p}\|_1 < 0.01 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $\boldsymbol{\tau}_{\text{prev}}$ denotes the RL output at the previous timestep. The L1-norm penalizes torque chattering linearly without being dominated by occasional large transients; the quadratic penalty $-\|\boldsymbol{\tau}\|^2$ activates only near convergence to prevent wasteful energy expenditure.

Upon termination, a graduated reward is assigned based on final convergence precision:

$$R_{\text{terminal}} = \begin{cases} -100, & \|\boldsymbol{p}\|_\infty \geq 1 \text{ or } \|\boldsymbol{\omega}\|_\infty \geq 1 \text{ rad/s} \\ 1000, & \|\boldsymbol{p}\|_1 < 0.01 \\ 200, & 0.01 \leq \|\boldsymbol{p}\|_1 < 0.05 \\ 100, & 0.05 \leq \|\boldsymbol{p}\|_1 < 0.1 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where $\|\boldsymbol{p}\|_\infty = \max(|p_1|, |p_2|, |p_3|)$. The graduated tiers (+1000, +200, +100) provide a curriculum-like signal: partial credit during early training helps the agent discover the basin of attraction, while progressively tighter thresholds refine convergence as the policy improves. The −100 divergence penalty supplements the opportunity cost of forfeited future rewards.

All weights and thresholds are calibrated so that each component's magnitude remains comparable at its intended phase.

3.4. Network Structures

Both the actor and critic networks follow the standard TD3 architecture of Fujimoto et al. [13], employing two hidden layers with 400 and 300 units, respectively. The network structures are summarized in Table 1.

Table 1. Network architectures of the actor and critic.

Component	Architecture	Activation
Actor	$6 \to 400 \to 300 \to 3$	ReLU (hidden), tanh (output)
Critic	$9 \to 400 \to 300 \to 1$	ReLU (hidden), linear (output)

The actor output is passed through a tanh activation scaled by $u_{\max}$ to respect the saturation limits defined in Section 3.2. The critic concatenates the state and action at its input layer and outputs a scalar Q-value. All hidden layers use ReLU activations; no batch normalization or dropout is applied.

4. Fault-Tolerant Attitude Controller Design

While the reinforcement learning-based controller offers significant advantages in adaptability and reduces manual tuning, its limitations in handling specific actuator fault modes necessitate the development of a dedicated fault-tolerant control scheme. Building upon the integral-type sliding mode framework of [28], an adaptive fault-tolerant (AFT) controller is formulated in this section. The proposed composite structure extends the framework of [28] in two key aspects: (i) the nominal controller is replaced by the trained TD3 agent rather than the saturated PD law in [28], which was designed for rigid-body dynamics only; and (ii) the sliding manifold and AFT compensator are reformulated to accommodate rigid–flexible coupling terms ($\delta^T\ddot{\boldsymbol{\eta}}$ and $\delta^T\dot{\boldsymbol{\eta}}$) with observer-estimated modal coordinates, thereby enabling fault tolerance in the presence of flexible dynamics that are not addressed in [28].

4.1. Adaptive Fault-Tolerant Controller

The controller is structured as

$$\boldsymbol{u} = \boldsymbol{u}_{RL} + \boldsymbol{u}_{aN}, \tag{16}$$

where $\boldsymbol{u}_{RL}$ is the nominal controller generated by the trained TD3 agent, and $\boldsymbol{u}_{aN}$ is the fault-tolerant controller used for compensating actuator faults. The pretrained TD3 agent is integrated into the composite control framework, as shown in Figure 1.

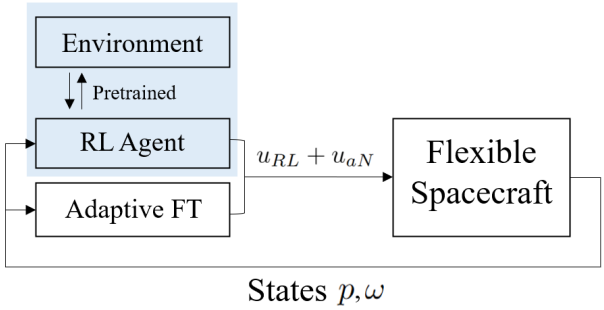


Figure 1. Composite controller.

Based on the standard sliding mode manifold design framework, the sliding manifold is constructed as follows:

$$\boldsymbol{s} = \boldsymbol{\Gamma}\left\{\boldsymbol{\omega}(t) - \boldsymbol{\omega}(t_0) - \int_{t_0}^{t} \boldsymbol{J}^{-1}\left[-\boldsymbol{\omega}(\sigma)^{\times}(\boldsymbol{J}\boldsymbol{\omega} + \boldsymbol{\delta}^T\dot{\boldsymbol{\eta}}) - \boldsymbol{\delta}^T\ddot{\boldsymbol{\eta}} + \boldsymbol{u}_{\mathrm{RL}}\right] d\sigma\right\}, \tag{17}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{3\times 3}$ denotes a constant matrix. The matrix $\boldsymbol{\Gamma}$ is selected such that the product $\boldsymbol{\Gamma}\boldsymbol{J}^{-1}$ is non-singular (invertible). The unmeasurable flexible modal states $\boldsymbol{\eta}$ is calculated by their observer estimates $\hat{\boldsymbol{\eta}}$ from Section 2.3. It should be noted that at time instant $t = t_0$, the sliding mode surface satisfies $\boldsymbol{s}(\boldsymbol{\omega}(t_0), t_0) = 0$, thereby eliminating the sliding mode reaching phase entirely.

The fault-tolerant compensator is designed as

$$\boldsymbol{u}_{aN} = \begin{cases} -\hat{\alpha}\dfrac{\left(\boldsymbol{\Gamma}\boldsymbol{J}^{-1}\right)^T\boldsymbol{s}}{\left\|\left(\boldsymbol{\Gamma}\boldsymbol{J}^{-1}\right)^T\boldsymbol{s}\right\|}, & \text{if } \hat{\alpha}\left\|\left(\boldsymbol{\Gamma}\boldsymbol{J}^{-1}\right)^T\boldsymbol{s}\right\| \geq \epsilon \\ -\hat{\alpha}^2\dfrac{\left(\boldsymbol{\Gamma}\boldsymbol{J}^{-1}\right)^T\boldsymbol{s}}{\epsilon}, & \text{if } \hat{\alpha}\left\|\left(\boldsymbol{\Gamma}\boldsymbol{J}^{-1}\right)^T\boldsymbol{s}\right\| < \epsilon \end{cases} \tag{18}$$

where ϵ is a small positive scalar alleviating undesirable chattering, and $\hat{\alpha} \geq 0$ is obtained by the following adaptive law:

$$\dot{\hat{\alpha}} = \gamma\left(\left\|\left(\boldsymbol{\Gamma}\boldsymbol{J}^{-1}\right)^T\boldsymbol{s}\right\| - \lambda\hat{\alpha}\right), \quad \text{with } \hat{\alpha}(0) \geq 0 \tag{19}$$

where γ and λ denote positive scalar constants, and the second term $-\lambda\hat{\alpha}$ serves to enhance robustness against external disturbances and unmodeled dynamic characteristics, while mitigating the excessive growth of the adaptive gain parameter. In this work, these parameters are set to $\gamma = 10.0$ and $\lambda = 2.5 \times 10^{-4}$, with the initial adaptive gain $\hat{\alpha}(0) = 1.0$ and the boundary layer thickness $\epsilon = 0.01$.

The fault-tolerant mechanism of the compensator (18) can be understood through the integral-type sliding mode (ISM) principle. By construction of the sliding manifold (17), the nominal rigid–flexible dynamics are embedded in the integral term; differentiating $\boldsymbol{s}$ cancels these nominal terms and yields (23), where only the fault, disturbance, and the deviation $\boldsymbol{u} - \boldsymbol{u}_{\mathrm{RL}}$ remain. When the compensator maintains $\boldsymbol{s} = 0$, the condition $\dot{\boldsymbol{s}} = 0$ enforces $(\boldsymbol{I}_3 - \boldsymbol{E})\boldsymbol{u} + \bar{\boldsymbol{u}} + \boldsymbol{d} = \boldsymbol{u}_{\mathrm{RL}}$, i.e., the equivalent control exactly counterbalances the fault and disturbance effects. Substituting this relation back into (3) recovers the nominal fault-

free dynamics driven solely by $\boldsymbol{u}_{\mathrm{RL}}$. Hence, the actuator effectiveness loss $\boldsymbol{E}(t)$, additive fault $\bar{\boldsymbol{u}}$, and disturbance $\boldsymbol{d}$ are all algebraically rejected on the sliding manifold—without requiring explicit fault identification. The compensator (18) enforces this sliding motion: its direction $(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}/\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\|$ is the steepest-descent direction for the Lyapunov function $\frac{1}{2}\boldsymbol{s}^T\boldsymbol{s}$, and the adaptive gain $\hat{\alpha}$ in (19) self-tunes to dominate the unknown combined fault–disturbance magnitude, thereby guaranteeing reachability without prior knowledge of the fault bounds.

4.2. Stability Analysis

Lemma 1. *Based on the switching gain update rule in* (19)*, $\hat{\alpha}$ is upper bounded. In other words, for all $t > 0$, there exists a positive scalar $\bar{\alpha}$ satisfying*

$$\hat{\alpha} \leq \bar{\alpha}, \quad \alpha \leq \bar{\alpha}, \quad \forall t > 0,$$

where α is expressed as $\alpha = \frac{\sqrt{3}e_m u_{\max} + \bar{u}_m + d_m + \varepsilon}{1 - e_m}$, with $u_{\max}$ denoting the upper bound of the RL output, e_m, d_m, and $\bar{u}_m$ defined in Assumptions 1–3, and ϵ being a bounded positive constant.

Proof. The proof follows the same lines as in the appendix of [28]. Therefore, it is omitted here. The essential idea is as follows. Consider the Lyapunov candidate $V_s = \frac{1}{2}\boldsymbol{s}^T\boldsymbol{s}$, whose derivative along the closed-loop trajectories satisfies $\dot{V}_s \leq -(1 - e_m)(\hat{\alpha} - \alpha)\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T s\|$, where α is the finite constant defined above. To establish boundedness of $\hat{\alpha}$, four mutually exclusive cases are examined based on whether $\hat{\alpha}$ exceeds α and whether $\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T s\|$ exceeds a small threshold related to the σ-modification parameter λ. In each case, either $\dot{V}_s \leq 0$ or the comparison principle is invoked, and in all cases a uniform upper bound for $\hat{\alpha}$ is established. □

Theorem 1. *Consider the flexible spacecraft attitude kinematics and dynamics described by* (1) *and* (3) *in the presence of fault described by* (4)*. Suppose that Assumptions 1–3 hold and only the boundedness of actuator faults/disturbances is known (their upper bounds are unknown). Under the action of the adaptive controller in* (16) *and the adaptive law in* (19)*, the system trajectories are able to enter a bounded region around the sliding manifold $\boldsymbol{s} = 0$ in finite time.*

Proof. First, construct the candidate Lyapunov function that fuses the sliding manifold error and the adaptive gain estimation error:

$$V = \frac{1}{2}\boldsymbol{s}^T\boldsymbol{s} + \frac{1 - e_m}{2\gamma}(\hat{\alpha} - \bar{\alpha})^2, \tag{20}$$

where $\bar{\alpha}$ is the upper bound of $\hat{\alpha}$ defined in Lemma 1, and $\gamma > 0$ is the adaptive gain. We analyze the time derivative $\dot{V}$ in two cases based on the boundary layer of the smoothed control $\boldsymbol{u}_{aN}$, as defined in Equation (18).

Case I: $\hat{\alpha}\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\| \geq \epsilon$

Taking the time derivative of Equation (17), the first-order derivative of the sliding surface is obtained as

$$\dot{\boldsymbol{s}} = \boldsymbol{\Gamma}\Big\{\dot{\boldsymbol{\omega}}(t) - \boldsymbol{J}^{-1}\Big[-\boldsymbol{\omega}(t)^{\times}(\boldsymbol{J}\boldsymbol{\omega}(t) + \boldsymbol{\delta}^T\dot{\boldsymbol{\eta}}(t)) - \boldsymbol{\delta}^T\ddot{\boldsymbol{\eta}}(t) + \boldsymbol{u}_{\mathrm{RL}}(t)\Big]\Big\}. \tag{21}$$

Multiplying both sides of Equation (3) by $\boldsymbol{J}^{-1}$ and rearranging terms, the angular velocity derivative $\dot{\boldsymbol{\omega}}$ is solved as

$$\dot{\boldsymbol{\omega}} = \boldsymbol{J}^{-1}\Big[-\boldsymbol{\omega}^{\times}\Big(\boldsymbol{J}\boldsymbol{\omega} + \boldsymbol{\delta}^T\dot{\boldsymbol{\eta}}\Big) - \boldsymbol{\delta}^T\ddot{\boldsymbol{\eta}} + (\boldsymbol{I}_3 - \boldsymbol{E}(t))\boldsymbol{u} + \bar{\boldsymbol{u}} + \boldsymbol{d}\Big]. \tag{22}$$

Substituting Equation (22) into Equation (21) yields

$$\dot{\boldsymbol{s}} = \boldsymbol{\Gamma}\boldsymbol{J}^{-1}[(\boldsymbol{I}_3 - \boldsymbol{E}(t))\boldsymbol{u} + \bar{\boldsymbol{u}} + \boldsymbol{d} - \boldsymbol{u}_{\mathrm{RL}}]. \tag{23}$$

Taking the time derivative of V, and substituting Equation (23), we obtain

$$\begin{aligned} \dot{V} &= \boldsymbol{s}^T\dot{\boldsymbol{s}} + \frac{1-e_m}{\gamma}(\hat{\alpha}-\bar{\alpha})\dot{\hat{\alpha}} \\ &\leq -(1-e_m)\left[\bar{\alpha} - \frac{\sqrt{3}e_m u_{\max} + \bar{u}_m + d_m}{1-e_m}\right]\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\| \\ &\quad - \lambda(1-e_m)(\hat{\alpha}-\bar{\alpha})\hat{\alpha}. \end{aligned} \tag{24}$$

From Lemma 1, we have the following key inequalities:

$$\bar{\alpha} - \frac{\sqrt{3}e_m u_{\max} + \bar{u}_m + d_m}{1-e_m} \geq \frac{\varepsilon}{1-e_m}, \quad |\hat{\alpha}-\bar{\alpha}| \leq \bar{\alpha}.$$

Substituting these into Equation (24), we further simplify $\dot{V}$ to

$$\begin{aligned} \dot{V} &\leq -\lambda(1-e_m)|\hat{\alpha}-\bar{\alpha}| - \varepsilon\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\| + \lambda(1-e_m)\left(|\hat{\alpha}-\bar{\alpha}| + \frac{1}{4}\bar{\alpha}^2\right) \\ &= -\delta_1|\hat{\alpha}-\bar{\alpha}| - \delta_2\|\boldsymbol{s}\| + \eta_1, \end{aligned} \tag{25}$$

where $\delta_1 = \lambda(1-e_m)$, $\delta_2 = \varepsilon\left\|\boldsymbol{\Gamma}\boldsymbol{J}^{-1}\right\|$, and $\eta_1 = \lambda(1-e_m)\left(\bar{\alpha} + \frac{1}{4}\bar{\alpha}^2\right)$ are positive scalars. Rewriting Equation (25) in terms of $\sqrt{V}$ (the square root of the Lyapunov function) for finite-time stability analysis:

$$\begin{aligned} \dot{V} &\leq -\sqrt{\frac{2\gamma\delta_1^2}{1-e_m}}\sqrt{\frac{1-e_m}{2\gamma}(\hat{\alpha}-\bar{\alpha})^2} - \sqrt{2}\delta_2\sqrt{\frac{1}{2}\boldsymbol{s}^T\boldsymbol{s}} + \eta_1 \\ &\leq -\delta_m V^{1/2} + \eta_1, \end{aligned}$$

where $\delta_m = \sqrt{2}\min\left\{\sqrt{\frac{\gamma\delta_1^2}{1-e_m}}, \delta_2\right\}$ is a positive constant.

Case II: $\hat{\alpha}\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\| < \epsilon$

For the case inside the boundary layer, the time derivative of V (Equation (20)) is derived as

$$\begin{aligned} \dot{V} = &-\frac{1-e_m}{\epsilon}\hat{\alpha}^2\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\|^2 + (1-e_m)\hat{\alpha}\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\| \\ &- (1-e_m)\left[\hat{\alpha} - \frac{\sqrt{3}e_m u_{\max} + \bar{u}_m + d_m}{1-e_m}\right]\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\| \\ &+ (1-e_m)(\hat{\alpha}-\bar{\alpha})\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\| - \lambda(1-e_m)\left(\hat{\alpha}^2 - \hat{\alpha}\bar{\alpha}\right). \end{aligned} \tag{26}$$

The quadratic term $-\frac{1-e_m}{\epsilon}\hat{\alpha}^2\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\|^2 + (1-e_m)\hat{\alpha}\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\|$ attains its maximum value $\frac{(1-e_m)\epsilon}{4}$ at $\hat{\alpha}\left\|(\boldsymbol{\Gamma}\boldsymbol{J}^{-1})^T\boldsymbol{s}\right\| = \frac{\epsilon}{2}$. Following the same derivation as Case I and using Lemma 1, we obtain

$$\dot{V} \leq -\delta_m V^{1/2} + \eta_2, \tag{27}$$

where $\eta_2 = \frac{1-e_m}{4}\left(\lambda\bar{\alpha}^2 + 4\lambda\bar{\alpha} + \epsilon\right)$ is a positive scalar.

Combining Case I and Case II, we obtain a unified inequality for $\dot{V}$ that holds for all $t > 0$:

$$\dot{V} \leq -\delta_m V^{1/2} + \eta_m, \tag{28}$$

where $\eta_m = \max\{\eta_1, \eta_2\} = \frac{1-e_m}{4}\left(\lambda\bar{\alpha}^2 + 4\lambda\bar{\alpha} + \epsilon\right)$. From the practical finite-time stability theory for nonlinear systems, the decrease of V drives the system trajectories into the set $V^{1/2} \leq \frac{\eta_m}{(1-\theta_0)\delta_m}$ in finite time, where $0 < \theta_0 < 1$ is a positive scalar. For the sliding manifold s, this implies the ultimate boundedness (the system converges to a small neighborhood of $\boldsymbol{s} = 0$):

$$\begin{aligned} \|\boldsymbol{s}\| &\leq \frac{\sqrt{2}\eta_m}{(1-\theta_0)\delta_m} \\ &= \frac{(1-e_m)\left(\lambda\bar{\alpha}^2 + 4\lambda\bar{\alpha} + \epsilon\right)}{4(1-\theta_0)\cdot\min\left\{\lambda\sqrt{\gamma(1-e_m)}, \varepsilon\left\|\boldsymbol{\Gamma}\boldsymbol{J}^{-1}\right\|\right\}}. \end{aligned} \tag{29}$$

Hence, the trajectory of the system remains bounded within a small neighborhood surrounding the sliding manifold $\boldsymbol{s} = 0$ over a finite time horizon. This concludes the proof. □

5. Simulation and Discussion

In this section, the training environment settings are introduced and hyperparameter configurations are summarized in Table 2. To systematically assess robustness, five evaluation scenarios (S0–S4) form a progressive uncertainty evaluation framework, as summarized in Table 3. Starting from a noise-free baseline (S0), each subsequent scenario introduces an additional source of non-ideality: measurement noise (S1), non-zero initial angular velocity (S2), parameter mismatch and external disturbance (S3), and finally concurrent actuator faults (S4). Detailed scenario configurations are provided in the notes accompanying Table 3.

This section validates the proposed control framework through numerical simulations, organized into three parts. First, the training setup, spacecraft physical parameters, and hyperparameter configurations are documented. Second, the trained TD3 agents are evaluated under fault-free conditions (Scenarios S0–S3), comparing the $\hat{\boldsymbol{\eta}}$-augmented agent A against the baseline agent A_0 to assess the effectiveness of vibration-aware reward shaping. Third, the composite TD3+AFT controller is tested under concurrent actuator faults (Scenario S4) and benchmarked against PD, PD+AFT, and the standalone TD3 agent.

5.1. Training Settings

During the training, each component of the initial attitude $\boldsymbol{p}_0$ is independently sampled from the uniform distribution over the interval $[-1, 1]$. No measurement noise, parameter perturbation or fault condition is introduced in the training phase. The environment is treated as ideal except for a sinusoidal disturbance $\boldsymbol{d}_{\text{train}}(t)$, as follows:

$$\begin{aligned} d_{\text{train},i}(t) &= A_i \sin\left(\frac{2\pi}{40}\, t + \phi_i\right) + 10^{-7}, \quad i = 1,2,3 \\ A_i &\sim \mathcal{U}(0,\ 0.05) \quad [\text{N·m}], \quad i = 1,2,3 \\ \phi_i &\sim \mathcal{U}(0,\ 2\pi) \quad [\text{rad}], \quad i = 1,2,3 \end{aligned} \tag{30}$$

The frequency $\frac{2\pi}{40}$ rad/s is chosen to be representative and to maintain the same order of magnitude as Equation (34). The amplitude A_i is intentionally set substantially larger than the actual disturbance magnitude. The exaggerated disturbance magnitude and fully randomized phases are designed to expose the agent to a diverse set of disturbance realizations during training, thereby encouraging the learning of robust control policies that generalize beyond the specific disturbance patterns encountered in evaluation.

The inertia matrix $\boldsymbol{J}$ is configured as

$$\boldsymbol{J} = \begin{bmatrix} 120 & 3 & 4 \\ 3 & 100 & 10 \\ 4 & 10 & 120 \end{bmatrix} \text{kg·m}^2. \tag{31}$$

Different from [32], where the off-diagonal components are neglected, this paper takes the inertial coupling into account to better match the practical engineering situation. The flexible parameters are selected with reference to [31]. The number of flexible modes m is set as 4, and the coupling matrix $\boldsymbol{\delta}$ is set as

$$\boldsymbol{\delta} = \begin{bmatrix} 6.45637 & 1.27814 & 2.15629 \\ -1.25819 & 0.91756 & -1.67264 \\ 1.11687 & 2.48901 & -0.83674 \\ 1.23637 & -2.65810 & -1.12530 \end{bmatrix}. \tag{32}$$

The flexible modal parameters are given by

$$\omega_n = \begin{bmatrix} 0.7681 & 1.1038 & 1.8733 & 2.5496 \end{bmatrix}, \xi = \begin{bmatrix} 0.0056 & 0.0086 & 0.0128 & 0.0252 \end{bmatrix}, \tag{33}$$

where ω_n denotes natural frequencies (rad/s) and ξ represents corresponding damping ratios.

The training procedure follows the standard TD3 protocol [13]. At each timestep, the agent observes the state $\boldsymbol{s}_{RL,t} = [\boldsymbol{p}^T, \boldsymbol{\omega}^T]^T$, selects an action $\boldsymbol{a}_t = \pi_\phi(\boldsymbol{s}_{RL,t}) + \epsilon_t$ with exploration noise $\epsilon_t \sim \mathcal{N}(0, \sigma^2_{\text{explore}} I)$, and receives a reward computed by the phased reward function (13). Transitions are stored in an experience replay buffer of capacity 10^6 and sampled uniformly in minibatches for network updates. Training begins after a warm-up phase of 1000 steps to populate the buffer. The complete hyperparameter configuration is listed in Table 2.

It is worth noting that the later TD3 policy network itself only receives the 6-dimensional state vector $[\boldsymbol{p}^T, \boldsymbol{\omega}^T]^T$ as defined in Section 3.2; the observer-estimated flexible states are used internally by the environment for reward computation and by the AFT control layer, and are never passed to the RL policy network.

The training is implemented in Python 3.10 using PyTorch 1.12.0. The flexible spacecraft environment is simulated with a fixed-step integration using SciPy's solve_ivp function with the RK45 method at a time step of $\Delta t = 1\,\text{s}$. All simulations are performed on a workstation with an Intel Core i7-13700 CPU and an NVIDIA RTX 3060 GPU; typical training duration is approximately 12 h for 20,000 episodes.

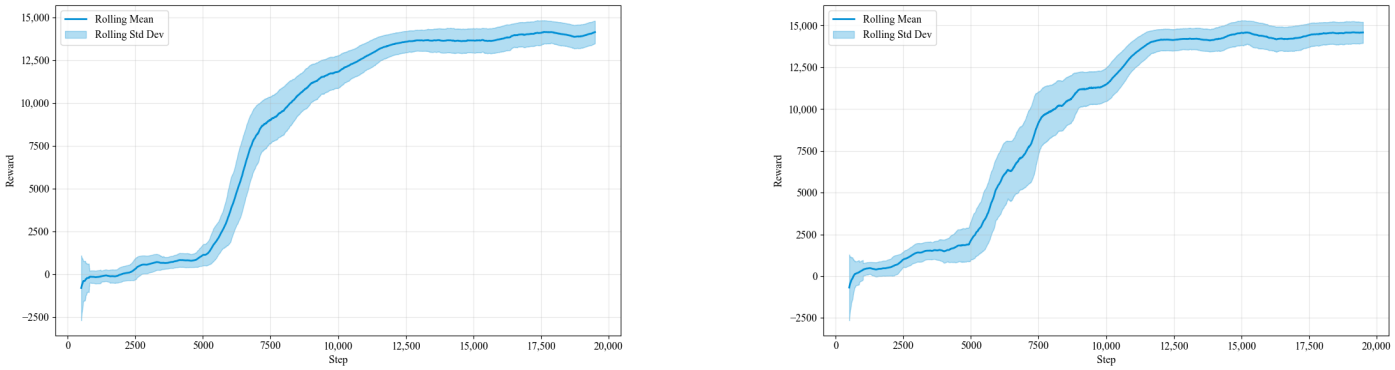
Table 2. Hyperparameter settings.

Parameter	Value
Learning Rates	
Actor learning rate (α_{actor})	3×10^{-4}
Critic learning rate (α_{critic})	3×10^{-4}
Network Architecture	
Hidden layer dimensions	400, 300
Number of hidden layers	2
RL Algorithm Parameters	
Discount factor (γ_{RL})	0.99
Soft update coefficient (τ)	0.01

Table 2. *Cont.*

Parameter	Value
Experience replay buffer size	1,000,000
Minimum buffer size for training	1000
Batch size	250
Policy update delay (d)	3
Exploration noise std (σ_{explore})	0.1
Target policy noise std (σ)	0.2
Target noise clip bound (c)	0.5
Optimizer and Reproducibility	
Optimizer	Adam
Random seeds (Python, NumPy, PyTorch)	2
Training Schedule	
Total training episodes (N_{episodes})	20,000
Episode length (T_{episode})	200 steps
Time step duration (Δt)	1 s

The effectiveness of incorporating the observer-estimated flexible state $\hat{\boldsymbol{\eta}}$ into the reward function is verified through an ablation experiment that compares the control performance of two agents trained by different reward functions (12) and (13). For clear distinction, the agent trained using the reward function (12) is designated as agent A_0, and the agent trained with the reward function (13) is denoted as agent A. The reward accumulation of the two agents is visualized in Figure 2. Both agents exhibit a rapid ascent after approximately 5000 episodes, indicating that the exploration phase transitions into policy exploitation around this point. Agent A_0 produces an overall smoother reward curve, whereas agent A exhibits slightly larger fluctuations but converges to a higher steady-state reward value. These differences stem from distinct reward structures: Agent A_0 optimizes solely with respect to attitude and angular velocity errors, yielding a simpler and more stationary reward signal; agent A additionally incorporates the observer-estimated flexible modal displacement $\hat{\boldsymbol{\eta}}$, which enriches the learning signal with vibration-suppression information but simultaneously introduces greater stochasticity due to the coupling between the flexible dynamics and the rigid-body motion. The richer feedback enables agent A to attain a higher asymptotic reward at the cost of moderately increased training variance.



(**a**) Reward growth of agent A0 (**b**) Reward growth of agent A

Figure 2. Reward growth of agents trained by different reward functions.

5.2. Test Scenario Settings

To systematically evaluate agent robustness under progressively challenging conditions, five evaluation scenarios (S0–S4) are constructed as a progressive uncertainty

evaluation framework. The complete scenario configurations are summarized in Table 3, and each scenario is described in detail below.

Table 3. Configurations of the five evaluation scenarios.

Setting	S0	S1	S2	S3	S4
Measurement noise	—	√	√	√	√
Non-zero initial ω_0	—	—	√	—	—
Parameter mismatch	—	—	—	√	√
External disturbance	—	—	—	√	√
Actuator faults	—	—	—	—	√

All five scenarios share a common initial attitude $\boldsymbol{p}_0 = [0.3,\ 0.1,\ -0.2]$, with a simulation duration of 200 s and a time step of 1 s. The target attitude is zero ($\boldsymbol{p} = 0$, $\boldsymbol{\omega} = 0$) in all cases.

Scenario S0 (pure baseline) represents the simplest evaluation condition: no measurement noise, no parameter mismatch, no external disturbance, and no actuator faults. The spacecraft starts from rest ($\boldsymbol{\omega}_0 = 0$) and the controller operates with perfect state information and a perfectly known plant model. This scenario establishes the nominal performance reference against which all subsequent, more challenging scenarios are compared.

Scenario S1 (measurement noise) introduces Gaussian white noise on both the MRP attitude and angular velocity measurements. The noise standard deviations are set to $\sigma_p = 8.0 \times 10^{-5}$ [33] and $\sigma_\omega = 5 \times 10^{-5}$ rad/s [34]. All other conditions remain identical to S0. This scenario evaluates the agents' tolerance to realistic sensor imperfections.

Scenario S2 (non-zero initial angular velocity) retains the measurement noise of S1 and additionally assigns a non-zero initial angular velocity $\boldsymbol{\omega}_0 = [0.02,\ -0.03,\ 0.04]$ rad/s. The spacecraft thus possesses initial kinetic energy at the start of the maneuver, making the attitude reorientation more demanding than the rest-to-rest cases of S0 and S1.

Scenario S3 (parameter mismatch and external disturbance) layers model-plant mismatch and persistent external disturbance on top of the measurement noise of S1. The parameter mismatch perturbs three off-diagonal elements of the moment of inertia matrix (from nominal values $J_{12} = J_{21} = 3$, $J_{13} = J_{31} = 4$ to $J_{12} = J_{21} = 6$, $J_{13} = J_{31} = 7\,\mathrm{kg\,m^2}$), while the flexible coupling matrix $\boldsymbol{\delta}$, natural frequencies ω_n, and damping ratios ξ are all increased by +10% relative to their nominal values. The external disturbance is a multi-sinusoidal torque adopted from [35]:

$$\boldsymbol{d}(t) = 0.2 \times 10^{-3} \cdot \begin{bmatrix} 3\cos(10\omega_d t) + 4\sin(3\omega_d t) - 10 \\ -1.5\sin(2\omega_d t) + 3\cos(5\omega_d t) + 15 \\ 3\sin(10\omega_d t) - 8\sin(4\omega_d t) + 10 \end{bmatrix} \mathrm{Nm}. \tag{34}$$

where $\omega_d = 0.1\,\mathrm{rad/s}$ in this work. The initial angular velocity is zero. Scenario S3 represents the most demanding fault-free condition in the evaluation suite.

Scenario S4 (actuator faults) incorporates all the non-idealities of S3—measurement noise, parameter mismatch, and external disturbance—and adds a time-varying actuator fault profile. Each actuator experiences a partial loss-of-effectiveness fault at $t = 10\,\mathrm{s}$. Subsequently, at $t = 50\,\mathrm{s}$, these actuators are also subjected to an additive time-varying fault, which is injected into the spacecraft dynamics in an additive manner. The parameters in (4) are selected as follows:

$$e_i(t) = \begin{cases} 0, & t < 10 \\ 0.5, & t \geq 10 \end{cases}$$
$$\bar{\boldsymbol{u}}(t) = \begin{cases} 0, & t < 50 \\ 0.3 + 0.05\sin(t), & t \geq 50 \end{cases} \tag{35}$$

The efficiency loss $e_i = 0.5$ and the additive bias $\bar{\boldsymbol{u}} = 0.3$ together represent a severe fault condition relative to the maximum actuator torque $u_{\max} = 1\,\mathrm{N{\cdot}m}$: the actuators lose half of their control effectiveness while being simultaneously subjected to a constant offset of $0.3\,\mathrm{N{\cdot}m}$, corresponding to static actuator anomalies such as sudden installation misalignment, mechanical wear, circuit zero-point drift, or long-term aging-induced output bias. The sinusoidal term $0.05\sin(t)$, introduces a time-varying component into the additive fault term, capturing the fact that actuator faults may not remain constant but can evolve over time—for instance, due to mechanical/electrical periodic perturbations such as reaction wheel bearing periodic friction, thruster jet pulsation, or electromagnetic circuit ripple.

This scenario represents the most challenging condition in the evaluation suite, testing the agents' resilience to concurrent sensor noise, modeling errors, environmental disturbances, and actuator degradation.

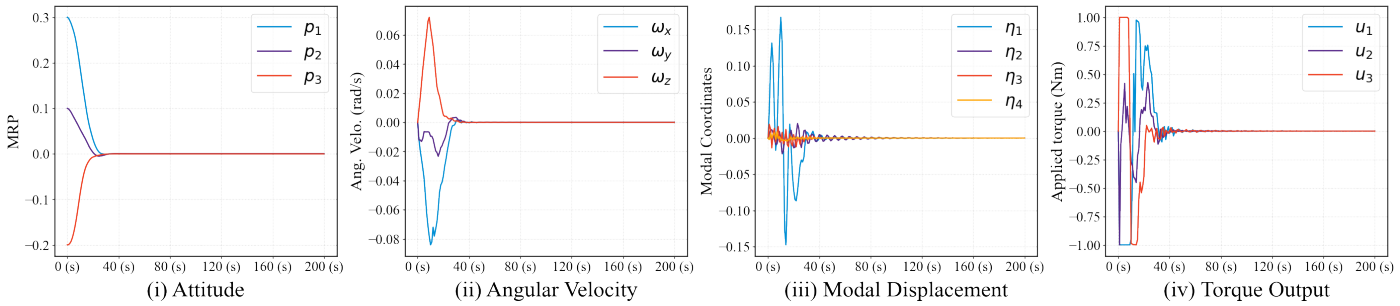
Scenarios S0–S3 are designed to compare the performance of agent A and agent A_0, demonstrating the benefit of incorporating the observer-estimated flexible modal state $\hat{\boldsymbol{\eta}}$ into the reward function. Scenario S4 serves as the validation platform for the adaptive fault-tolerant (AFT) control framework, where four control schemes are evaluated and compared under concurrent actuator faults.

5.3. Fault-Free Scenarios

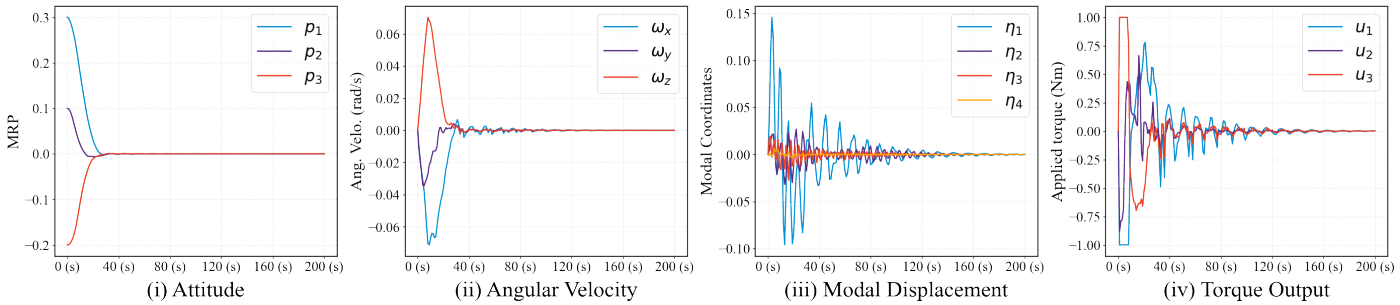
The performance of agent A and agent A_0 under S0–S3 is visualized in Figures 3–6. Each figure presents the time histories of the MRP attitude, angular velocity, control torque output, and flexible modal displacement for both agents in a side-by-side comparison.

Across all four scenarios, agent A consistently exhibits noticeably cleaner trajectories with reduced oscillations and smaller vibration peaks, whereas agent A_0 often produces persistent back-and-forth chattering in the control output accompanied by sustained flexible-mode excitation. This difference is a direct consequence of the $\hat{\boldsymbol{\eta}}$-augmented reward function (13): by rewarding the suppression of the observer-estimated modal displacement during training, agent A learns a policy that inherently avoids control actions that would excite the flexible modes. The resulting closed-loop behavior achieves simultaneous attitude regulation and vibration damping, yielding smoother state trajectories and a cleaner control effort. In contrast, agent A_0, trained without any vibration-related feedback (12), optimizes only for attitude errors and may inadvertently induce and sustain flexible oscillations, which manifest as noisier signals and persistent torque chattering.

To provide a quantitative assessment, steady-state performance metrics (computed over $t > 160\,\mathrm{s}$) of both agents across S0–S3 are summarized in Table 4. The metrics include the steady-state infinity norms of the MRP attitude ($\|\boldsymbol{p}\|_\infty$) and angular velocity ($\|\boldsymbol{\omega}\|_\infty$), the root mean square (RMS) of the flexible modal displacement vector norm (η RMS) as a measure of residual vibration energy, the settling time of the flexible modal displacement ($\eta_{0.01}$), and the RMS of the applied control torque (u RMS) as a measure of control effort. Here $\eta_{0.01}$ is defined as the earliest time after which the peak absolute modal displacement across all four flexible modes ($\max(|\eta_1|, |\eta_2|, |\eta_3|, |\eta_4|)$) remains permanently below 0.01, i.e., the vibration settling time. "—" denotes failure to converge within the 200 s evaluation horizon.

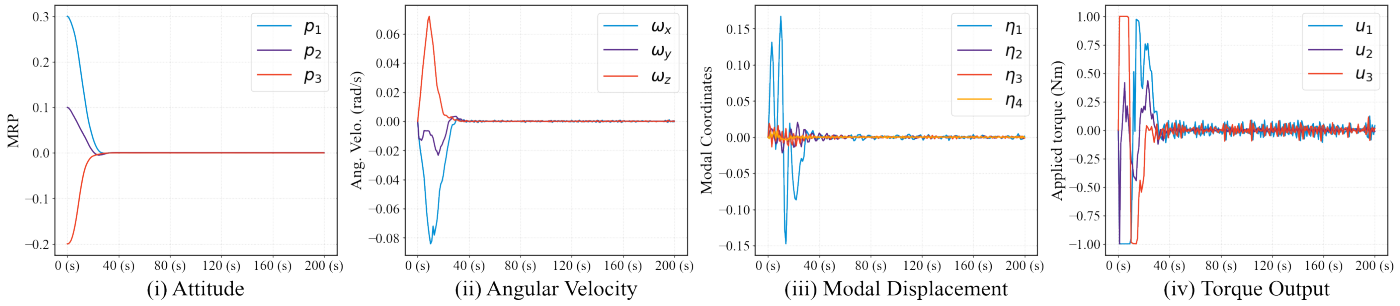


(**a**) Agent A performance in S0

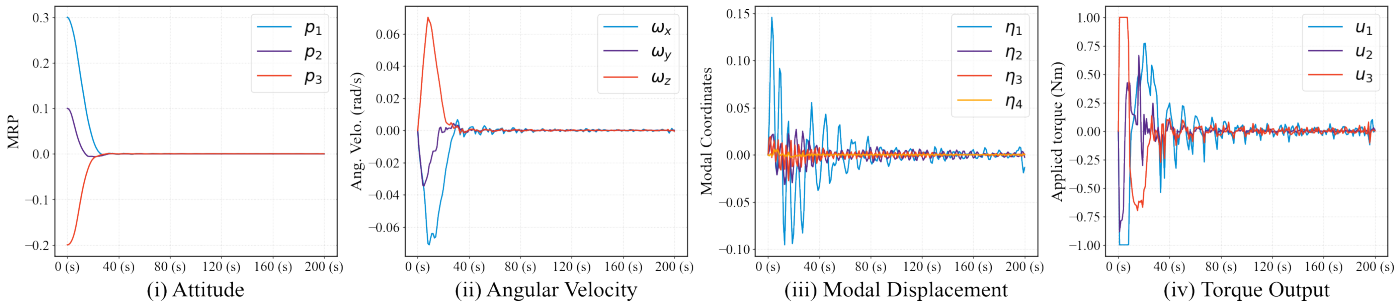


(**b**) Agent A_0 performance in S0

Figure 3. Agent performance comparison in Scenario S0 (pure baseline).



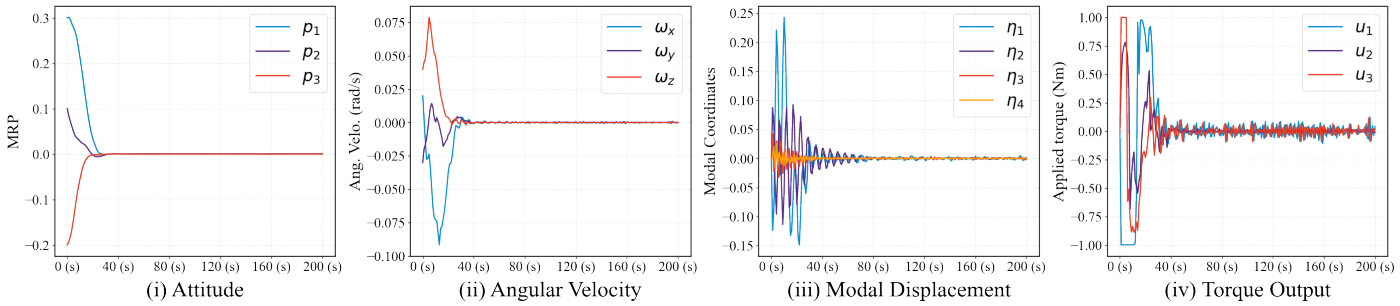
(**a**) Agent A performance in S1



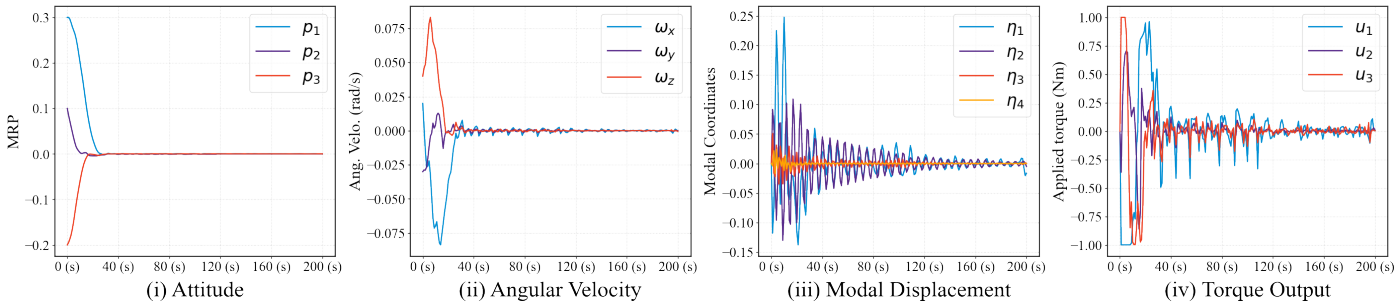
(**b**) Agent A_0 performance in S1

Figure 4. Agent performance comparison in Scenario S1 (measurement noise).

The quantitative results in Table 4 reveal several important findings. First, in terms of vibration suppression, agent A consistently and substantially outperforms agent A_0 across all four scenarios. Compared with agent A_0, agent A reduces the η RMS by approximately 92.7% (S0), 59.3% (S1), 67.2% (S2), and 88.0% (S3). Correspondingly, compared with agent A_0, agent A reduces $\eta_{0.01}$ (vibration settling time) by 68.4% (S0); in all other scenarios, agent A_0 fails to converge within the 200 s evaluation horizon. This demonstrates that the $\hat{\boldsymbol{\eta}}$-augmented reward function (13) provides a robust and generalizable vibration suppression capability that persists even as non-idealities accumulate.

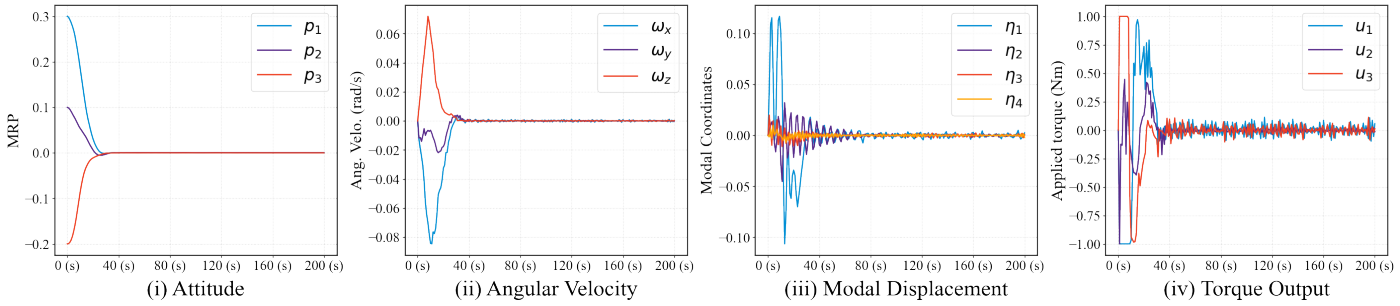


(**a**) Agent A performance in S2

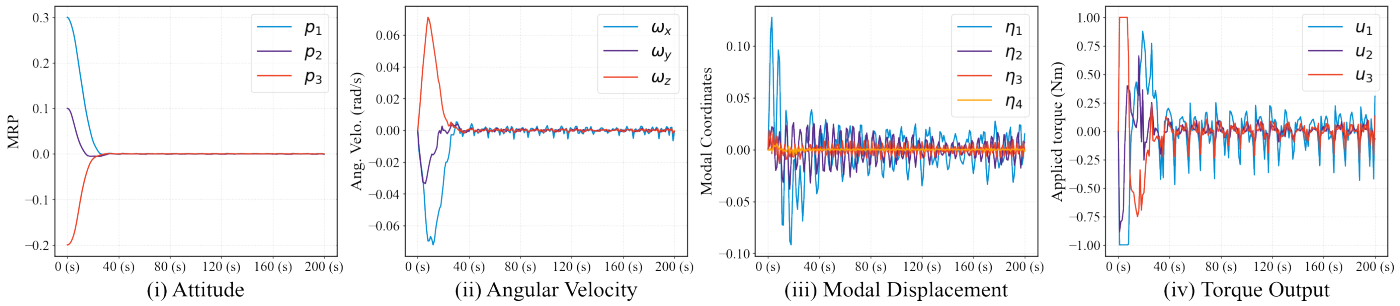


(**b**) Agent A_0 performance in S2

Figure 5. Agent performance comparison in Scenario S2 (measurement noise + non-zero $\boldsymbol{\omega}_0$).



(**a**) Agent A performance in S3



(**b**) Agent A_0 performance in S3

Figure 6. Agent performance comparison in Scenario S3 (measurement noise + parameter mismatch + external disturbance).

Second, a consistent trade-off between attitude pointing accuracy and vibration suppression is observed. Although agent A achieves better vibration damping across all scenarios, agent A_0 attains marginally better pointing accuracy in S0–S2, at the cost of severely degraded vibration damping. However, in S3, the most demanding fault-free scenario, agent A surpasses agent A_0 across all metrics including attitude pointing ($\|\boldsymbol{p}\|_\infty = 6.55 \times 10^{-4}$ vs. 8.91×10^{-4}), indicating that the $\hat{\boldsymbol{\eta}}$-guided policy learns a more robust control strategy that does not sacrifice pointing accuracy under compound uncertainties.

Table 4. Quantitative comparison of agent A vs A0 across Scenarios S0–S3 (steady-state, $t > 160$ s).

Scenario	Agent	$\|p\|_\infty$	$\|\omega\|_\infty$ (rad/s)	η RMS	$\eta_{0.01}$ (s)	u RMS (Nm)
S0	A	4.57×10^{-4}	7.00×10^{-6}	5.00×10^{-5}	30	6.16×10^{-4}
	A0	2.87×10^{-4}	1.17×10^{-4}	6.87×10^{-4}	95	4.91×10^{-3}
S1	A	5.44×10^{-4}	9.97×10^{-4}	2.24×10^{-3}	30	5.80×10^{-2}
	A0	3.63×10^{-4}	1.72×10^{-3}	5.50×10^{-3}	—	4.89×10^{-2}
S2	A	5.44×10^{-4}	9.95×10^{-4}	2.25×10^{-3}	56	5.79×10^{-2}
	A0	4.02×10^{-4}	1.83×10^{-3}	6.85×10^{-3}	—	4.95×10^{-2}
S3	A	6.55×10^{-4}	1.12×10^{-3}	2.32×10^{-3}	42	6.33×10^{-2}
	A0	8.91×10^{-4}	6.54×10^{-3}	1.94×10^{-2}	—	0.194

Third, the torque expenditure of agent A is moderately higher than that of agent A_0 across S1–S2 (e.g., 5.80×10^{-2} vs. 4.89×10^{-2} Nm in S1), reflecting the additional control effort required for active vibration suppression. In S3, however, agent A achieves substantially lower torque RMS (6.33×10^{-2} vs. 0.194 Nm), demonstrating that under compound uncertainties the $\hat{\boldsymbol{\eta}}$-guided policy simultaneously delivers superior vibration attenuation and higher control efficiency. In S0, agent A achieves significantly lower torque RMS (6.16×10^{-4} vs. 4.91×10^{-3} Nm), indicating that in the absence of perturbations, the learned policy achieves both superior vibration attenuation and higher control efficiency.

Overall, the ablation study across S0–S3 confirms that incorporating the observer-estimated flexible modal displacement $\hat{\boldsymbol{\eta}}$ into the reward function enables the agent to strike a better balance between attitude pointing accuracy and vibration suppression. The resulting policy generalizes effectively across diverse non-ideal conditions at a modest increase in control effort under perturbed scenarios.

5.4. Fault Scenario

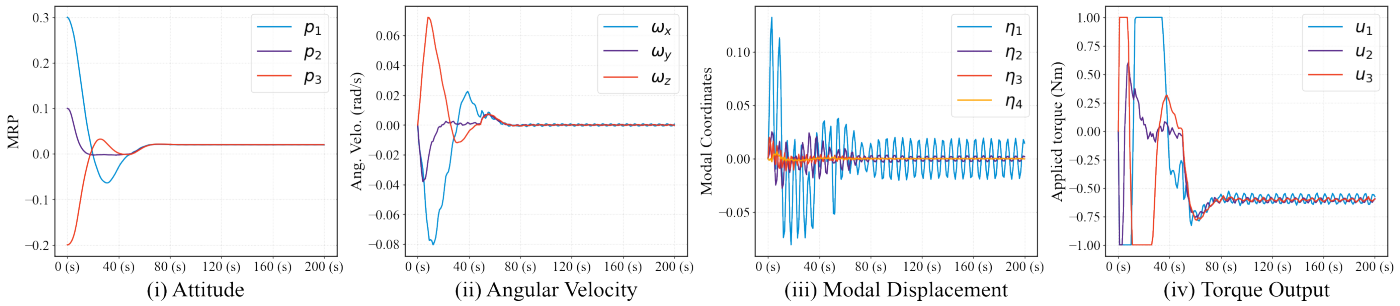
Given that agent A demonstrated superior vibration suppression and robust performance across S0–S3 in the ablation study, its trained policy weights are adopted as the TD3 nominal controller in this section. Four control architectures are compared under Scenario S4: (i) pure PD ($\boldsymbol{u}_{pd} = -k_p\boldsymbol{p} - k_d\boldsymbol{\omega}$ with $k_p = 30$, $k_d = 60$), (ii) pure TD3, (iii) PD augmented with the AFT compensator (PD+AFT), and (iv) TD3 augmented with the AFT compensator (TD3+AFT).

To evaluate the adaptability of the proposed control framework under actuator fault conditions, Scenario S4 is employed, whose configuration is detailed in Section 5.2. This simulation validates the efficacy of the composite control framework proposed in Section 4.

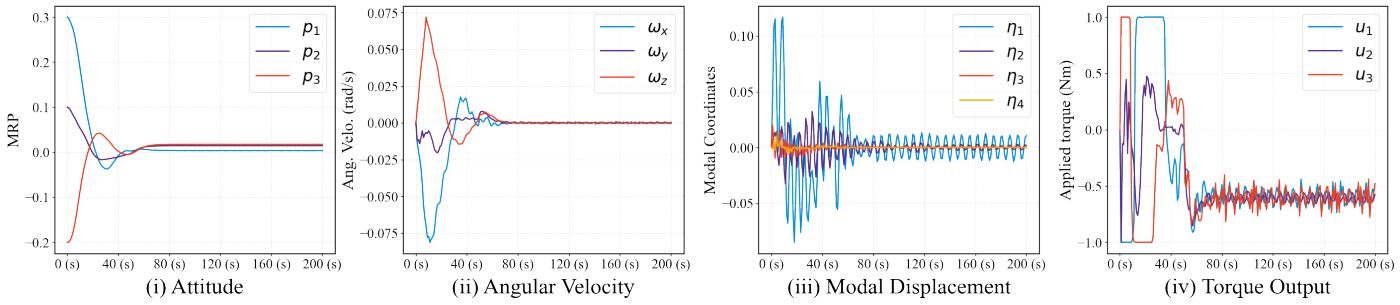
The simulation outcomes of the four control architectures under S4 are illustrated in Figure 7. To further provide a quantitative assessment, key performance metrics are summarized in Table 5.

In Table 5, all steady-state metrics are computed over $t > 160$ s. Integrated Torque RMS is computed over the full trajectory. Peak $\|\boldsymbol{\eta}\|_\infty$ and peak $\|\boldsymbol{\omega}\|_\infty$ capture the maximum values over the entire 200 s trajectory. Integrated η RMS is computed over the full trajectory as a proxy for total vibration energy. MRPs converging to 0.005 indicates the earliest time after which $\|\boldsymbol{p}\|_\infty$ remains below 0.005 ("—" denotes failure to converge within 200 s).

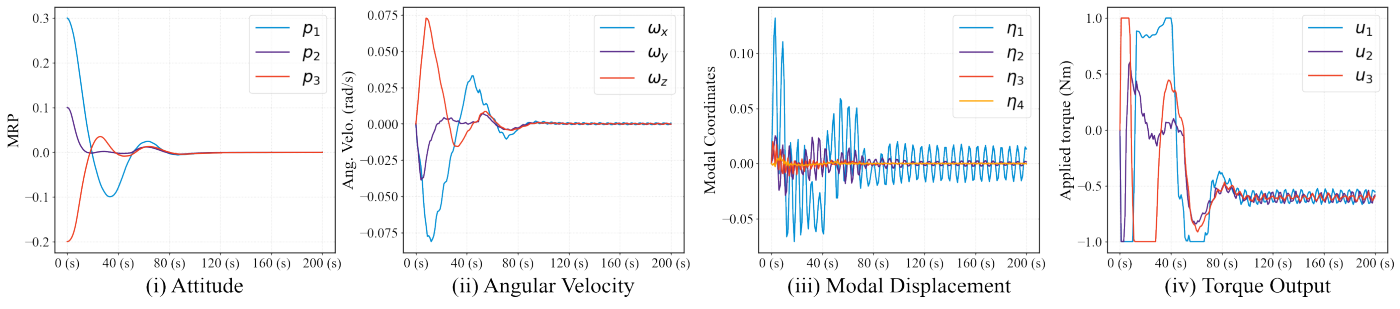
Several observations can be drawn from the quantitative comparison in Table 5. First, AFT augmentation is essential for precise convergence under actuator faults: pure PD and pure TD3 both fail to bring the MRP error below 0.005 within the 200 s evaluation horizon, whereas TD3+AFT achieves this threshold at 74 s and PD+AFT at 94 s. The proposed TD3+AFT architecture delivers the best overall pointing accuracy, reducing $\|\boldsymbol{p}\|_\infty$ by a factor of approximately 43.3 relative to pure TD3 (1.75×10^{-2} vs. 4.04×10^{-4}) and by a factor of approximately 50.0 relative to pure PD (2.02×10^{-2} vs. 4.04×10^{-4}).



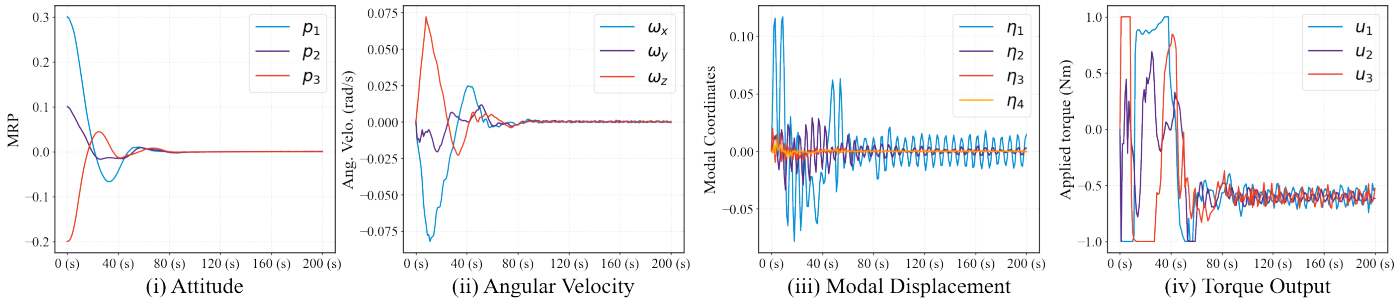
(**a**) PD controller performance in S4



(**b**) TD3 performance in S4



(**c**) PD + AFT controller performance in S4



(**d**) TD3 + AFT performance in S4

Figure 7. Performance comparison of four control architectures under Scenario S4 (actuator faults).

Second, TD3 consistently outperforms PD across all transient and steady-state vibration metrics. In steady state, TD3 achieves η RMS of 8.43×10^{-3} compared to 1.37×10^{-2} for PD, representing a 38.5% reduction. Among AFT-augmented architectures, TD3+AFT achieves lower η RMS (1.04×10^{-2} vs. 1.17×10^{-2} for PD+AFT). In the transient phase, TD3 and TD3+AFT tie for the lowest peak $\|\boldsymbol{\eta}\|_\infty$ (both 0.117) and integrated η RMS (both 2.91×10^{-2}), confirming that the learned TD3 policy inherently produces smoother maneuvers with reduced flexible mode excitation.

Third, a comparison between the two AFT-augmented methods shows that TD3+AFT achieves better pointing accuracy, convergence speed, and vibration suppression: 30.5% lower steady-state $\|\boldsymbol{p}\|_\infty$ (4.04×10^{-4} vs. 5.81×10^{-4}), 21.3% faster MRP convergence

(74 s vs. 94 s), and 11.1% lower steady-state η RMS (1.04×10^{-2} vs. 1.17×10^{-2}). PD+AFT retains marginal advantages in peak angular velocity ($\|\boldsymbol{\omega}\|_\infty$: 8.12×10^{-2} vs. 8.22×10^{-2} rad/s) and integrated torque RMS (1.112 vs. 1.118 N·m), though the differences are small. Overall, the learned TD3 nominal policy coordinates with the AFT compensator more effectively than the fixed-gain PD law, as the RL agent learns to generate control actions that complement the fault-tolerant layer rather than compete with it.

Table 5. Quantitative comparison of four control methods under Scenario 4 (actuator faults).

	PD	TD3	PD+AFT	TD3+AFT
Steady-state ($t > 160$ s)				
MRP $\|\boldsymbol{p}\|_\infty$	2.02×10^{-2}	1.75×10^{-2}	5.81×10^{-4}	4.04×10^{-4}
$\|\boldsymbol{\omega}\|_\infty$ (rad/s)	8.43×10^{-4}	8.54×10^{-4}	8.34×10^{-4}	7.54×10^{-4}
η RMS	1.37×10^{-2}	8.43×10^{-3}	1.17×10^{-2}	1.04×10^{-2}
Torque RMS (Nm)	1.042	1.045	1.044	1.044
Transient/full-trajectory				
Peak $\|\boldsymbol{\eta}\|_\infty$	0.132	0.117	0.132	0.117
Peak $\|\boldsymbol{\omega}\|_\infty$ (rad/s)	8.05×10^{-2}	8.12×10^{-2}	8.12×10^{-2}	8.22×10^{-2}
Integrated η RMS	2.98×10^{-2}	2.91×10^{-2}	3.03×10^{-2}	2.91×10^{-2}
MRPs converge to 0.005 (s)	—	—	94	74
Integrated Torque RMS (Nm)	1.082	1.086	1.112	1.118

Regarding control effort, all four controllers operate within the $u_{\max} = 1$ N·m actuator saturation limit stated in Section 3.2. The steady-state torque RMS values are tightly clustered (1.042–1.045 N·m), and the integrated torque RMS also remains within a narrow band (1.082–1.118 N·m). The fact that TD3+AFT achieves substantially superior pointing and vibration performance while expending comparable control effort indicates that both the RL policy and the AFT layer contribute to control efficiency in complementary ways: the TD3 agent provides smooth nominal commands that minimize unnecessary actuation, while the AFT compensator injects targeted corrections only when faults are active.

6. Conclusions

This study focuses on the attitude maneuver problem of flexible spacecraft under external disturbances, flexible dynamics, and actuator faults. An integrated control framework combining TD3-based RL agent with an adaptive fault-tolerant compensator is proposed. Simulation results demonstrate that incorporating the observer-estimated flexible modal displacement into the reward function enables the RL agent to learn active vibration suppression while maintaining competitive attitude accuracy. Under actuator faults, the AFT compensator significantly improves control precision, and the composite TD3+AFT architecture achieves the best pointing accuracy, convergence speed, and vibration suppression among all configurations, while maintaining comparable control effort. The RL agent and AFT compensator operate in a complementary manner: the RL agent provides smooth nominal commands that inherently suppress flexible mode excitation, while the AFT layer injects targeted fault-tolerant corrections when needed.

Future work includes experimental validation on physical testbeds to assess performance under real sensing and actuation imperfections beyond the Gaussian noise models used in the present simulations. Additionally, extending the RL policy to support on-orbit incremental learning would enable real-time adaptation to evolving spacecraft parameters and fault conditions, further improving the controller's generalization capability in complex operational environments.

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

Funding: This research was funded by the National Natural Science Foundation of China (No. U24B20157) and the Science and Technology Commission of Shanghai Municipality Key Technology R&D Plan (No. 25DZ3100703).

Data Availability Statement: The source code presented in this study is openly available at https://github.com/EllencePeng/Aerospace-RL-Flexible-Spacecraft (accessed on 18 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Golestani, M.; Zhang, W.; Yang, Y.; Xuan-Mung, N. Disturbance observer-based constrained attitude control for flexible spacecraft. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *59*, 963–972. [CrossRef]
2. He, G.; Cao, D. Dynamic Modeling and Attitude–Vibration Cooperative Control for a Large-Scale Flexible Spacecraft. *Actuators* **2023**, *12*, 167. [CrossRef]
3. Shahid, F.; Luo, H.; Jiang, Y.; Noman Hasan, M. Finite-Time Adaptive Sliding Mode Fault-Tolerant Attitude Control for Flexible Spacecraft. *IEEE Trans. Control Syst. Technol.* **2025**, *33*, 1700–1711. [CrossRef]
4. Dian, W.; Yunhua, W.; Hongyi, X.; Franco, B.Z.; Chen, X. Scheduled input shaping based attitude agile control for flexible spacecraft with vibration suppression. *Aerosp. Sci. Technol.* **2025**, *164*, 110355. [CrossRef]
5. Zhong, C.; Chen, Z.; Guo, Y. Attitude Control for Flexible Spacecraft with Disturbance Rejection. *IEEE Trans. Aerosp. Electron. Syst.* **2017**, *53*, 101–110. [CrossRef]
6. Song, G.; Agrawal, B.N. Vibration suppression of flexible spacecraft during attitude control. *Acta Astronaut.* **2001**, *49*, 73–83. [CrossRef]
7. Yu, Y.; Meng, X.; Li, K.; Xiong, F. Robust control of flexible spacecraft during large-angle attitude maneuver. *J. Guid. Control. Dyn.* **2014**, *37*, 1027–1033. [CrossRef]
8. Angeletti, F.; Iannelli, P.; Gasbarri, P.; Sabatini, M. End-to-end design of a robust attitude control and vibration suppression system for large space smart structures. *Acta Astronaut.* **2021**, *187*, 416–428.
9. Khoroshylov, S.; Martyniuk, S.; Sushko, O.; Vasyliev, V.; Medzmariashvili, E.; Woods, W. Dynamics and attitude control of space-based synthetic aperture radar. *Nonlinear Eng.* **2023**, *12*, 20220277. [CrossRef]
10. Izzo, D.; Märtens, M.; Pan, B. A survey on artificial intelligence trends in spacecraft guidance dynamics and control. *Astrodynamics* **2019**, *3*, 287–299. [CrossRef]
11. Khoroshylov, S.V.; Redka, M.O. Deep learning for space guidance, navigation, and control. *Space Sci. Technol.* **2021**, *27*, 38–52.
12. Silvestrini, S.; Lavagna, M. Deep learning and artificial neural networks for spacecraft dynamics, navigation and control. *Drones* **2022**, *6*, 270. [CrossRef]
13. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 1587–1596.
14. Hong, D.; Lee, S.; Cho, Y.H.; Baek, D.; Kim, J.; Chang, N. Energy-Efficient Online Path Planning of Multiple Drones Using Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2021**, *70*, 9725–9740. [CrossRef]
15. Abo Mosali, N.; Shamsudin, S.S.; Alfandi, O.; Omar, R.; Al-Fadhali, N. Twin Delayed Deep Deterministic Policy Gradient-Based Target Tracking for Unmanned Aerial Vehicle with Achievement Rewarding and Multistage Training. *IEEE Access* **2022**, *10*, 23545–23559. [CrossRef]
16. Zhang, D.; Li, X.; Ren, G.; Yao, J.; Chen, K.; Li, X. Three-Dimensional Path Planning of UAVs in a Complex Dynamic Environment Based on Environment Exploration Twin Delayed Deep Deterministic Policy Gradient. *Symmetry* **2023**, *15*, 1371. [CrossRef]
17. Yan, T.; Liu, C.; Gao, M.; Jiang, Z.; Li, T. A Deep Reinforcement Learning-Based Intelligent Maneuvering Strategy for the High-Speed UAV Pursuit-Evasion Game. *Drones* **2024**, *8*, 309. [CrossRef]
18. Zhou, T.; Liu, Z.; Jin, W.; Han, Z. Intelligent maneuver decision-making for UAVs using the TD3–LSTM reinforcement learning algorithm under uncertain information. *Front. Robot. AI* **2025**, *12*, 1645927. [CrossRef] [PubMed]
19. Zhang, Z.; Li, X.; An, J.; Man, W.; Zhang, G. Model-Free Attitude Control of Spacecraft Based on PID-Guide TD3 Algorithm. *Int. J. Aerosp. Eng.* **2020**, *2020*, 8874619. [CrossRef]

20. Gao, M.; Yan, T.; Li, Q.; Fu, W.; Zhang, J. Intelligent Pursuit–Evasion Game Based on Deep Reinforcement Learning for Hypersonic Vehicles. *Aerospace* **2023**, *10*, 86. [CrossRef]
21. Zhu, J.; Tang, W.; Dong, J. Design of Intelligent Controller for Aero-engine Based on TD3 Algorithm. *Inf. Technol. Control* **2023**, *52*, 1010–1024. [CrossRef]
22. Cheng, Y.; Chen, L.; Chen, C.L.P.; Wang, X. Off-Policy Deep Reinforcement Learning Based on Steffensen Value Iteration. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *13*, 1023–1032. [CrossRef]
23. Shi, J.; Liu, X.; Cai, G.; Liu, F.; Sun, J.; Zhu, D. Reinforcement Learning for Cooperative Control of the Ultra-Close Flexible Spacecraft Formation Flying. *IEEE Aerosp. Electron. Syst. Mag.* **2026**, *41*, 54–68. [CrossRef]
24. Li, Y.; Zhang, F.; Gao, Z. Attitude control for dumbbell-shaped spacecraft subject to parameter uncertainty via fully-actuated system approach and deep reinforcement learning. *Int. J. Syst. Sci.* **2026**, 1–18. [CrossRef]
25. El-Dalahmeh, G.; Jabbarpour, M.R.; Vo, B.Q.; Kowalczyk, R. Intelligent Spacecraft Attitude Fault Recovery Using Deep Reinforcement Learning. *IEEE Access* **2026**, *14*, 6238–6260.
26. Henna, H.; Toubakh, H.; Kafi, M.R.; Gürsoy, Ö.; Sayed-Mouchaweh, M.; Djemai, M. Satellite fault tolerant attitude control based on expert guided exploration of reinforcement learning agent. *J. Exp. Theor. Artif. Intell.* **2025**, *37*, 987–1011.
27. Mahfouz, A.; Valiullin, A.; Lukashevichus, A.; Pritykin, D. Reinforcement learning for attitude control of A spacecraft with flexible appendages. In Proceedings of the 73rd International Astronautical Congress. International Astronautical Federation, Paris, France, 18–22 September 2022.
28. Shen, Q.; Wang, D.; Zhu, S.; Poh, E.K. Integral-Type Sliding Mode Fault-Tolerant Control for Attitude Stabilization of Spacecraft. *IEEE Trans. Control Syst. Technol.* **2015**, *23*, 1131–1138. [CrossRef]
29. Terzakis, G.; Lourakis, M.; Ait-Boudaoud, D. Modified Rodrigues Parameters: An Efficient Representation of Orientation in 3D Vision and Graphics. *J. Math. Imaging Vis.* **2018**, *60*, 422–442. [CrossRef]
30. Zou, A.M.; De Ruiter, A.H.J.; Dev Kumar, K. Distributed attitude synchronization control for a group of flexible spacecraft using only attitude measurements. *Inf. Sci.* **2016**, *343–344*, 66–78. [CrossRef]
31. Di Gennaro, S. Output stabilization of flexible spacecraft with active vibration suppression. *IEEE Trans. Aerosp. Electron. Syst.* **2003**, *39*, 747–759. [CrossRef]
32. Peng, Z.; Shen, Q.; Song, C. Attitude Control of Spacecraft Based on Deep Reinforcement Learning TD3 Algorithm. In *Proceedings of the International Conference on Guidance, Navigation and Control*; Springer: Singapore, 2024; pp. 318–327. [CrossRef]
33. Zou, A.M.; Fan, Z. Fixed-time attitude tracking control for rigid spacecraft without angular velocity measurements. *IEEE Trans. Ind. Electron.* **2019**, *67*, 6795–6805. [CrossRef]
34. Javaid, U.; Basin, M.V.; Ijaz, S. Spacecraft Attitude Stabilization Control Under Actuator Faults and Input Saturation. In *Proceedings of the 2024 IEEE 18th International Conference on Advanced Motion Control (AMC), Kyoto, Japan*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [CrossRef]
35. Hu, Q.; Li, B.; Qi, J. Disturbance observer based finite-time attitude control for rigid spacecraft under input saturation. *Aerosp. Sci. Technol.* **2014**, *39*, 13–21. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.