Local CBF-QP Projection Credit Assignment for Reinforcement Learning in Terminal-Phase Spacecraft Rendezvous and Docking

Shiqi Ren[a,b,c], Pengcheng Wang[a,b,c,*]

[a] *Innovation Academy for Microsatellites, Chinese Academy of Sciences, Shanghai, 201203, China*
[b] *University of Chinese Academy of Sciences, Beijing, 100049, China*
[c] *Key Laboratory for Satellite Digitalization Technology, Chinese Academy of Sciences, Shanghai, China*

Abstract

External control-barrier-function quadratic programs (CBF-QPs) modify reinforcement-learning commands before execution, but projection obscures which action directions create constraint conflict. We propose PFCR-DQP, a post-filter-consistent residual-learning method that uses local QP geometry for policy credit in terminal-phase spacecraft rendezvous and docking. The active set and multipliers of the operational solve define a local affine projection model. Its auxiliary loss penalizes the current policy's residual along active constraint normals while retaining tangential freedom. Numerical regularity checks determine which reference samples contribute. In a five-seed comparison with the barrier-aligned PFCR-RC-AW reference, the selected DQP policies reduce the steps at which the safety layer seizes more than half of the per-axis thrust authority from 35.94% to 5.12% and the median correction magnitude from 0.3307 to 0.0312 N, while the rate of any numerically discernible correction falls from 67.30% to 54.84%. The nominal-command deficit falls from 0.3922 to 0.1102 N and the axis-summed actuator-output impulse from 205.9 to 77.2 N s. Protected docking success is 99.5% versus 97.6%; the paired success interval includes zero. Transfer without retraining yields 98.4% and 97.4% success in nonlinear Tudat-T0 and T1 propagation, respectively. Development curves show substantial late-training degradation, distinguishing the gains of selected policies from sustained training stability.

Keywords: autonomous rendezvous and docking, control barrier function, safe reinforcement learning, quadratic-program sensitivity, projection credit assignment, runtime safety layer

1. Introduction

Terminal-phase spacecraft rendezvous and docking requires precise relative-motion control under finite thrust. The admissible approach speed decreases with separation, so guidance must coordinate braking, control expenditure, and terminal capture. Reinforcement learning can learn feedback policies for this task across a distribution of initial conditions and model errors [1–4]. Its trajectory-level objective, however, leaves a separate problem at each control step: whether the proposed command satisfies the current constraints.

An external runtime safety layer addresses this problem by modifying the policy command before execution [5–10]. In a control-barrier-function quadratic program (CBF-QP), the filtered command is the Euclidean projection of the nominal command onto a state-dependent admissible set. With actuator mismatch, the complete action chain has three distinct quantities: the nominal command $\boldsymbol{u}_{\mathrm{des}}$, the supervisor output $\boldsymbol{u}_{\mathrm{act}}$, and the realized thrust $\boldsymbol{u}_{\mathrm{real}}$.

This projection also changes the information available for policy improvement [11–13]. Different nominal commands can project to the same filtered command and generate similar transitions and returns. PPO still evaluates the probability of the sampled nominal action, but task return alone does not directly identify which action component created the constraint conflict. The resulting ambiguity can sustain repeated braking corrections by the safety layer even after the policy learns to complete the task.

Intervention penalties and constraint budgets discourage such dependence by adding scalar costs to the training objective [14–18]. The QP solution contains more specific information. At a fixed active set, its Jacobian projects perturbations onto the tangent space of the active constraints. Components normal to that face are removed, while tangential components remain available for control. Using this geometry can direct a policy update toward the source of a conflict while retaining locally feasible action freedom.

We propose PFCR-DQP, a local projection-credit extension of post-filter-consistent residual learning. It constructs an affine projection model from the active set and multipliers of the QP that generated each sampled command. An auxiliary loss then penalizes the current policy's residual from that local projection. The reference actions and Jacobian remain fixed during differentiation, and numerical regularity checks determine which samples contribute. This uses the sensitivity of the operational solve [19–21] without introducing a second forward optimizer. The additional computation is confined to training.

The reference method, PFCR-RC-AW, combines barrier-normal costs at the command and actuator interfaces with a

*Corresponding author.
Email addresses: renshiqi@mail.ustc.edu.cn (Shiqi Ren), wangpc@microsate.com (Pengcheng Wang)

reversible, competence-gated anti-windup schedule. PFCR-DQP adds directional projection credit to that training objective. We evaluate the resulting policies under Clohessy–Wiltshire–Hill (CWH) relative dynamics and then transfer the selected controllers to nonlinear orbital propagation in TudatPy [22, 23]. The latter test examines how policy–filter coordination responds to environmental perturbations, mass depletion, and thrust errors beyond the training model.

The contributions are:

1. A construction of local projection sensitivity from the active constraints and multipliers of the operational CBF-QP, with explicit conditions for accepting a reference sample.
2. An affine projection-credit loss whose gradient targets the active constraint-normal residual while retaining tangential freedom in the local model, integrated with barrier-aligned dual-interface training.
3. A control study linking that credit to smaller safety-layer overrides, lower command deficits, and lower control impulse in terminal docking, supported by paired CWH evaluations and nonlinear Tudat transfer. Resolving the safety layer by correction magnitude, rather than by a single binary modification event, separates boundary-tracking trims from braking overrides. The learning curves also identify substantial late-training degradation.

On the selected-checkpoint CWH holdout, PFCR-DQP reduces the steps at which the safety layer seizes more than half of the per-axis thrust authority from 35.94% to 5.12%, mean control impulse from 205.9 to 77.2 N s, and nominal-command deficit from 0.3922 to 0.1102 N, with 99.5% protected docking success. These gains concern the selected policies: continued training does not preserve the early performance, a distinction examined directly in the results.

2. Related work

2.1. Safety filters and policy credit

Shielding, control barrier functions, and predictive safety filters combine learned task control with explicit runtime constraints [24–26]. Spacecraft applications have used runtime assurance for docking and inspection, where a learned policy operates alongside a model-based constraint handler [8, 10]. These architectures separate the proposed action from the executed control. Studies of safe reinforcement learning and action projection show that this separation also affects policy optimization [11–13].

Correction penalties and intervention costs encourage a policy to propose commands that need less filtering [14–16]. They summarize an intervention through its size or occurrence. A projection onto a polyhedron additionally identifies the constrained action directions. PFCR-DQP uses that directional information as an auxiliary policy gradient; the task advantage continues to determine which locally feasible directions are useful.

2.2. Differentiable quadratic programs

OptNet and subsequent sensitivity analyses provide tools for differentiating optimization solutions through their optimality conditions [20, 21]. For an identity-Hessian QP with a fixed, regular active set, the solution is locally affine in the nominal command. Its Jacobian is the orthogonal projector onto the null space of the active constraint matrix. Active-set switches and degeneracy delimit the region in which this classical derivative applies [13, 21].

PFCR-DQP uses this established sensitivity result to construct a training objective around sampled runtime projections. The active set comes from the Goldfarb–Idnani solve [19] that produced the filtered control. This makes the reference map consistent with the constraints and solution used in the rollout. The contribution lies in how the local map supplies policy credit and is combined with the two interface costs, rather than in a new QP differentiation formula.

2.3. Constraint budgets and actuator mismatch

Constrained policy optimization and dual-variable methods regulate expected costs alongside task return [17, 18]. Their weights balance performance against measured constraint demand during learning. This is complementary to runtime filtering: an average training budget shapes policy behavior across trajectories, while a CBF-QP acts on the current command.

Actuator mismatch introduces a further distinction between an admissible command and the force that drives the plant. Formal treatments of measurement and actuation uncertainty express safety conditions under specified error bounds [27, 28]. In the present setting, separate command and realized-force deficits provide training feedback at these two interfaces. Local QP credit resolves the command projection; realized-force feedback records the downstream effect of actuator mismatch.

3. Problem Formulation

3.1. Nominal CWH Relative Dynamics

The Hill frame is centered at the target spacecraft's center of mass. Let the relative position and velocity of the chaser be $\boldsymbol{r} = [r_x, r_y, r_z]^\top$ and $\boldsymbol{v} = [v_x, v_y, v_z]^\top$, respectively. The state is

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{r}^\top & \boldsymbol{v}^\top \end{bmatrix}^\top \in \mathbb{R}^6. \tag{1}$$

Neglecting mass variation, higher-order perturbations, and attitude–orbit coupling, the translational phase is modeled by the Clohessy–Wiltshire–Hill (CWH) equations [22],

$$\dot{\boldsymbol{x}} = A_{\mathrm{CWH}}\boldsymbol{x} + B_{\mathrm{CWH}}\boldsymbol{u}_{\mathrm{real}}, \tag{2}$$

$$A_{\mathrm{CWH}} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 3n^2 & 0 & 0 & 0 & 2n & 0 \\ 0 & 0 & 0 & -2n & 0 & 0 \\ 0 & 0 & -n^2 & 0 & 0 & 0 \end{bmatrix}, \qquad B_{\mathrm{CWH}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/m & 0 & 0 \\ 0 & 1/m & 0 \\ 0 & 0 & 1/m \end{bmatrix}. \tag{3}$$

The mean motion is $n = 0.001027\ \mathrm{s}^{-1}$, and the chaser mass is $m = 12$ kg. The nominal CWH plant uses forward-Euler integration. Both its control period and integration step are $\Delta t = 1$ s. The axiswise thrust bounds are

$$\mathcal{U} = \left\{ \boldsymbol{u} \in \mathbb{R}^3 : -u_{\max} \le u_i \le u_{\max} \right\}, \qquad u_{\max} = 1\ \mathrm{N}. \tag{4}$$

In the controller and CWH actuator model, this command set applies componentwise to the policy command, the primary-QP candidate, the supervisor output, and the saturated actuator output. The Tudat-T1 truth-force perturbation introduced below is applied after this command-domain saturation and can therefore move the final truth-model force outside $\mathcal{U}$.

3.2. Nonlinear Tudat Truth Dynamics

To implement the dual-model validation design described above, we use TudatPy 1.0.0 to construct a representative nonlinear orbital-simulation truth model [23]. Tudat organizes the body environment, acceleration models, and numerical integrators as composable modules, allowing spacecraft states to be propagated in a common inertial frame while explicitly including environmental forces and mass variation omitted by the CWH model. Here, the truth model generates the state transitions during evaluation.

The truth layer replaces only the state propagation in Eq. (2); it does not change the controller-side computations. CWH dynamics continue to support policy training, the nominal CBF-QP, the front-plane recoverability assessment, and the DQP active-set projection used during training. At each control period, the Hill-frame relative state is reconstructed from the Tudat states and supplied to the frozen policy and safety layer. The actuator output $\boldsymbol{u}_{\mathrm{real}}$ then serves as the baseline input to the Tudat thrust model. The target and chaser are propagated separately in the Earth-centered J2000 frame, and the Hill basis is reconstructed from the target's inertial position and velocity at every truth-model integration time. If $\boldsymbol{C}_{I \leftarrow H}(t) = [\boldsymbol{e}_r, \boldsymbol{e}_t, \boldsymbol{e}_h]$ denotes the rotation from Hill coordinates to J2000, the truth dynamics of the controlled chaser can be summarized as

$$\dot{\boldsymbol{r}}_k^I = \boldsymbol{v}_k^I, \qquad \dot{\boldsymbol{v}}_k^I = \boldsymbol{a}_{\mathrm{env},k}^I + \mathbf{1}_{\{k=\mathrm{cha}\}} \frac{\boldsymbol{C}_{I \leftarrow H}(t) \boldsymbol{F}_{\mathrm{truth}}^H}{m_c(t)}.$$

Here, $k \in \{\mathrm{tar}, \mathrm{cha}\}$ denotes the target and chaser, respectively. The controller produces a Hill-frame realized thrust once every 1 s. Its three Hill-frame components are held constant over that control interval, but at each Tudat derivative evaluation they are rotated continuously into J2000 with the propagated target's local orbital frame. The truth model is integrated with fixed-step fourth-order Runge–Kutta (RK4) integration at 0.25 s, explicitly separating the control zero-order hold from the integration substeps.

We use two truth-model levels. Tudat-T0 applies Earth point-mass gravity to the target and Earth point-mass gravity plus the direct mapping $\boldsymbol{F}_{\mathrm{truth}}^H = \boldsymbol{u}_{\mathrm{real}}$ to the chaser, whose mass is fixed at 12 kg. The representative Tudat-T1 model adds, for both spacecraft, Earth's degree-two, order-zero spherical-harmonic gravity (the point-mass term and J_2), NRLMSISE-00 atmospheric drag [29], cannonball solar-radiation pressure with Earth occultation, and point-mass gravity from the Sun and Moon. It models the chaser truth force and mass depletion as

$$\boldsymbol{F}_{\mathrm{truth}}^H = \boldsymbol{R}_{\mathrm{mis}} \operatorname{diag}(\boldsymbol{\kappa}) \boldsymbol{u}_{\mathrm{real}} + \boldsymbol{\epsilon}_t, \qquad \dot{m}_c = -\frac{\|\boldsymbol{F}_{\mathrm{truth}}^H\|_2}{I_{\mathrm{sp}} g_0}.$$

We set $I_{\mathrm{sp}} = 220$ s. Standard gravity is $g_0 = 9.806\,65\ \mathrm{m/s^2}$. The per-axis gains satisfy $\kappa_i \sim \mathrm{Unif}[0.97, 1.03]$; a fixed misalignment angle for each episode is sampled from $\mathrm{Unif}[0, 0.5^\circ]$, with its rotation axis obtained by normalizing a three-component standard-normal vector; and the stepwise additive thrust error is sampled independently across axes and control steps from a normal distribution with standard deviation 0.005 N, then clipped to $[-0.015, 0.015]$ N. The gain and misalignment draws are held fixed over an episode, whereas the additive-error sequence is sampled at the control-step rate. The baseline drag and solar-radiation-pressure parameters are as follows. The target has mass 500 kg and an area of 2.0 m^2 for both drag and radiation pressure; the chaser has an initial mass of 12 kg and an area of 0.1 m^2 for both effects. Both spacecraft use a drag coefficient of 2.2 and a radiation-pressure coefficient of 1.3. Tudat-T1 therefore propagates 12 translational states for the two spacecraft and one chaser-mass state. The reference target orbit is circular, with a semimajor axis determined by $n = 0.001027\ \mathrm{s}^{-1}$ and an inclination of 51.6°; the epoch, right ascension of the ascending node, and true anomaly are sampled in paired fashion according to the frozen protocol. The T1 parameters define the representative perturbation case evaluated here.

3.3. Mission, Actuator Mismatch, and Action Chain

The initial separation is sampled uniformly from 50–100 m. The initial bearing lies inside a cone centered on the $+x$ docking axis with a half-angle of 35°, and the initial relative velocity is zero. Capture requires

$$\|\boldsymbol{r}\|_2 \le 1\ \mathrm{m}, \qquad \|\boldsymbol{v}\|_2 \le 0.2\ \mathrm{m\,s^{-1}}. \tag{5}$$

The maximum mission duration is 1200 s, and an episode terminates if the relative separation exceeds 200 m.

The policy, runtime-assurance (RTA) supervisor, and actuator form a three-stage action chain,

$$\begin{aligned} \boldsymbol{u}_{\mathrm{act},t} &= \mathcal{F}_{\mathrm{RTA}}(\boldsymbol{x}_t, \boldsymbol{u}_{\mathrm{real},t-1}, \boldsymbol{u}_{\mathrm{des},t}), \\ (\boldsymbol{u}_{\mathrm{real},t}, \boldsymbol{z}_{t+1}) &= \mathcal{A}_{\boldsymbol{\xi}}(\boldsymbol{u}_{\mathrm{act},t}, \boldsymbol{z}_t). \end{aligned} \tag{6}$$

The map $\mathcal{F}_{\mathrm{RTA}}$ comprises the primary CBF-QP, the front-plane recoverability check, and the backup path defined below; when the primary candidate is accepted, it returns that QP solution. The actuator map $\mathcal{A}_{\boldsymbol{\xi}}$ contains a first-order lag state $\boldsymbol{z}_t$ and per-episode fixed hidden parameters $\boldsymbol{\xi}$. Independently for each episode, the mismatch domain samples a lag time constant $\tau \in [0, 4]$ s, three-axis thrust scales $\sigma_i \in [0.7, 1]$, and a shared dead-zone half-width $d_{\mathrm{dz}} \in [0, 0.05]$ N; thrust saturation is still applied at the output. The standard deviations of the position and velocity observation noise are 0.5 m and 0.03 m s^{-1}, respectively. The discrete actuator update and the ordered 36-component policy observation are given in Appendix Appendix A.1.

3.4. CBF-QP and Front-Plane Safety Supervisor

Let $R = \|\boldsymbol{r}\|_2$ and $V = \|\boldsymbol{v}\|_2$ denote the position and velocity magnitudes. The terminal speed envelope is

$$h_s(\boldsymbol{x}) = \nu_0 + \nu_1 R - V \geq 0, \qquad \nu_0 = 0.2 \text{ m s}^{-1}, \quad \nu_1 = 2.15n. \tag{7}$$

The QP rows use the following squared function:

$$\phi_s(\boldsymbol{x}) = (\nu_0 + \nu_1 R)^2 - V^2, \tag{8}$$

Because $V \geq 0$ and $\nu_0 + \nu_1 R > 0$, the inequalities $\phi_s \geq 0$ and $h_s \geq 0$ are equivalent. The implementation imposes the mixed condition $\dot{\phi}_s + 0.05 h_s + 0.1 h_s^3 \geq 0$; which combines the derivative of the squared barrier with a restoring term in the speed margin. Each axis additionally uses

$$h_{v,i}(\boldsymbol{x}) = v_{\max}^2 - v_i^2 \geq 0, \qquad v_{\max} = 10 \text{ m s}^{-1}, \tag{9}$$

together with $\dot{h}_{v,i} + 0.0005 h_{v,i} + 0.001 h_{v,i}^3 \geq 0$. All gains use SI units. Appendix Appendix A.4 gives the row coefficients and the range regularization. At control step t, these conditions and the thrust-box constraints can all be written as

$$\boldsymbol{G}_t \boldsymbol{u} \geq \boldsymbol{b}_t, \tag{10}$$

where $\boldsymbol{G}_t \in \mathbb{R}^{q \times 3}$ and $\boldsymbol{b}_t \in \mathbb{R}^q$ also depend on the previous realized thrust and the declared actuator domain when the recoverability row is present. The operational safety layer uses the standard CBF-QP construction [5, 6] to form the primary candidate

$$\begin{aligned} \boldsymbol{u}_{\mathrm{qp},t} := \arg\min_{\boldsymbol{u} \in \mathbb{R}^3} \quad & \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{u}_{\mathrm{des},t}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{G}_t \boldsymbol{u} \geq \boldsymbol{b}_t. \end{aligned} \tag{11}$$

The base constraint matrix contains four CBF rows and six thrust-box bounds. For C/D/E, the front-plane supervisor appends one axial-command row, so $q = 11$. The row imposes $u_x \geq b_{\mathrm{cmd}}$, where the bound is obtained by requiring a certified braking continuation after the current action. Appendix Appendix A.2 defines the 512-step force, velocity, and position lower bounds; Appendix Appendix A.3 derives the physical force floor and its robust preimage under actuator uncertainty. The row participates in the QP active set and therefore in the local DQP geometry. The objective Hessian is I_3, so the primary candidate is unique whenever the problem is feasible. All candidate QP projections for training, development, and hold-out evaluation are generated by the same active-set solver. On a projection-valid primary step, $\boldsymbol{u}_{\mathrm{act},t} = \boldsymbol{u}_{\mathrm{qp},t}$; a backup step instead returns the backup command through $\mathcal{F}_{\mathrm{RTA}}$.

The supervisor protects a front-plane buffer $x_{\mathrm{buf}} = 0.25$ m. If the QP fails or its candidate does not preserve the braking conditions, backup control takes over and the episode continues. Certification is checked for the returned command. Configurations C and E use the braking/hold law and release conditions in Appendix Appendix A.5; D uses the saturated LQR of Section 4.4.

At a post-integration sample, task success requires Eq. (5), terminal front-plane acceptance $x \geq x_{\mathrm{buf}} - 10^{-8}$ m, and, for the recovery supervisors used by C/D/E, no earlier front-plane-buffer violation. A buffer violation is recorded when $x_{t+1} - x_{\mathrm{buf}} < -10^{-8}$ m. Because there is no rigid-body contact model, a separate segment-based collision proxy detects a fast passage through the capture sphere between samples. With $\Delta\boldsymbol{r}_t = \boldsymbol{r}_{t+1} - \boldsymbol{r}_t$, it evaluates $\tau^* = \mathrm{clip}[-\boldsymbol{r}_t^\top \Delta\boldsymbol{r}_t / \|\Delta\boldsymbol{r}_t\|_2^2, 0, 1]$ (and sets $\tau^* = 0$ when the denominator is below 10^{-12}), then linearly interpolates $\boldsymbol{r}^* = \boldsymbol{r}_t + \tau^* \Delta\boldsymbol{r}_t$ and $\boldsymbol{v}^* = \boldsymbol{v}_t + \tau^*(\boldsymbol{v}_{t+1} - \boldsymbol{v}_t)$. The proxy is raised when $\|\boldsymbol{r}^*\|_2 \leq R_{\mathrm{cap}}$ and $\|\boldsymbol{v}^*\|_2 > V_{\mathrm{cap}}$. An episode is counted as a protected docking success when it is a task success and contains no capture-sphere collision proxy, front-plane-buffer violation, or uncertified supervisor sample; the last event denotes a sampled command that fails the supervisor's declared recoverability certification. This metric describes translational terminal capture. Speed-envelope excursions are reported separately and do not disqualify this metric. The scope of the sampled safety conditions is discussed in Section 7.3.

4. PFCR-DQP

Figure 1 provides an overview of the deployment loop and the training-only credit paths before the individual components are defined.

4.1. Barrier-aligned dual-interface training

The baseline PFCR-RC-AW method retains the three physical actions associated with each sampled transition and defines residuals at both the safety-filter and actuator interfaces:

$$\boldsymbol{\delta}_{\mathrm{F},t} = \boldsymbol{u}_{\mathrm{act},t} - \boldsymbol{u}_{\mathrm{des},t}, \qquad \boldsymbol{\delta}_{\mathrm{A},t} = \boldsymbol{u}_{\mathrm{real},t} - \boldsymbol{u}_{\mathrm{act},t}. \tag{12}$$

A scalar action-correction penalty does not resolve the geometry of a safety half-space. We define two row families $\mathcal{R}_t^{\mathrm{F}}$ and $\mathcal{R}_t^{\mathrm{A}}$, each containing the four CBF rows and one front-plane row. Their shared CBF rows are evaluated at $\boldsymbol{x}_t$; the front-plane rows depend additionally on $\boldsymbol{u}_{\mathrm{real},t-1}$. The F row uses the robust command bound b_{cmd}, while the A row uses the physical force

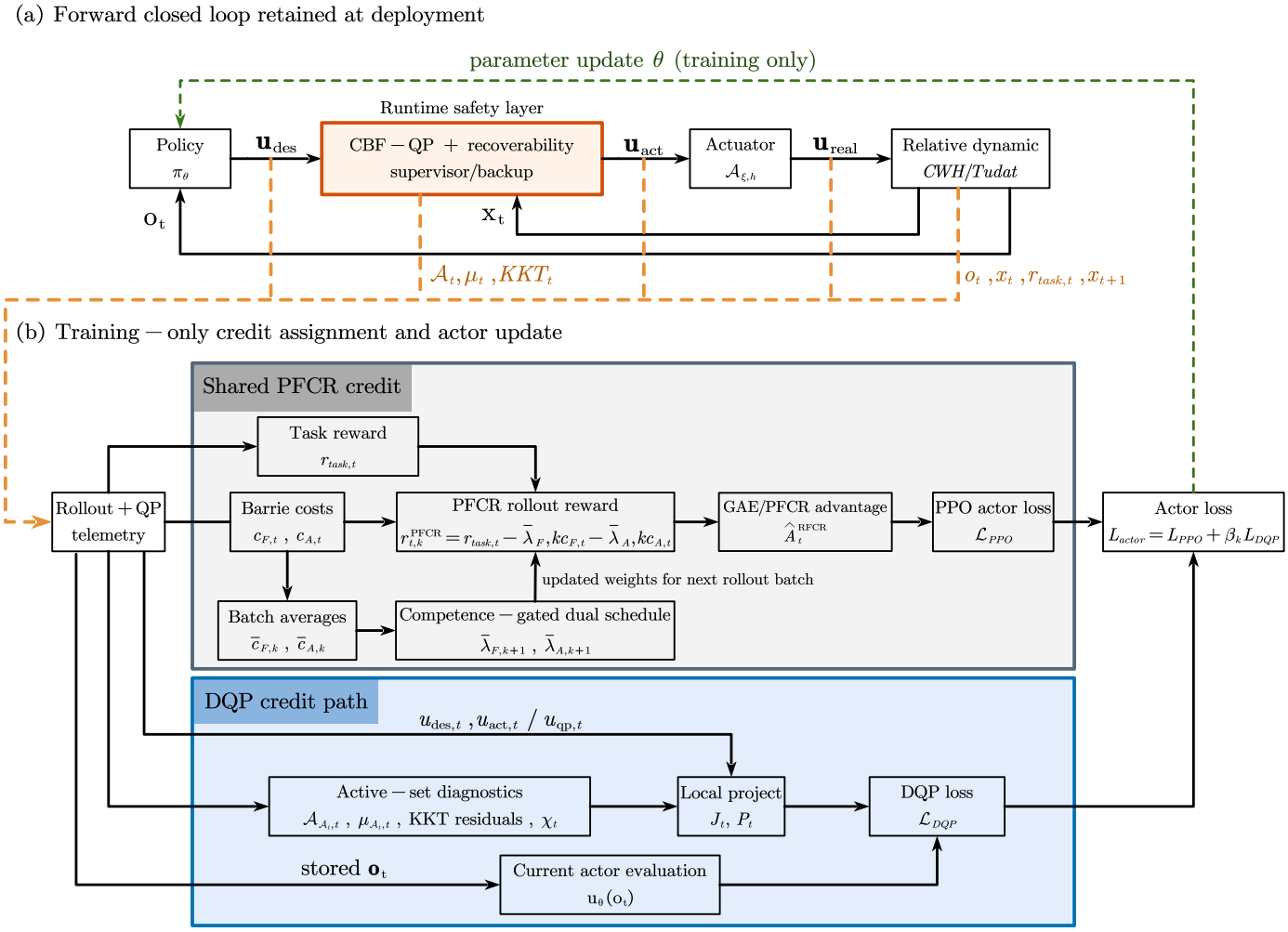


Figure 1: PFCR-DQP architecture. (a) The forward closed loop retained at deployment. The policy produces the nominal command $\boldsymbol{u}_{\text{des}}$, the runtime safety layer returns $\boldsymbol{u}_{\text{act}}$, and the actuator produces $\boldsymbol{u}_{\text{real}}$. This output drives CWH and Tudat-T0 directly; Tudat-T1 first maps it to $\boldsymbol{F}_{\text{truth}}^{H}$ through the frozen truth-force model. (b) The training-only credit paths. Task advantage, barrier-aligned costs, and dual scheduling are shared by PFCR-RC-AW and PFCR-DQP. The proposed DQP branch reads the actual active set and multipliers from the operational QP solver and, after KKT verification, constructs J_t, P_t, and $\mathcal{L}_{\text{DQP}}$. This branch does not backpropagate end-to-end through the QP, actuator, or dynamics, and all training-only modules are removed at deployment.

floor F_{req}. The six thrust-box bounds enter the QP but not these cost sets. Appendix Appendix A.4 gives their coefficients.

$$s_{j,t}^{\ell}(\boldsymbol{u}) = \frac{(g_{j,t}^{\ell})^{\top}\boldsymbol{u} - b_{j,t}^{\ell}}{\|g_{j,t}^{\ell}\|_2}, \qquad \ell \in \{\mathrm{F}, \mathrm{A}\}. \tag{13}$$

Let $\mathcal{I}_t^{\ell}$ index the finite, nondegenerate rows in the corresponding family. At zero margin, the costs are

$$\begin{aligned} c_{\mathrm{F},t} &= \max_{j \in \mathcal{I}_t^{\mathrm{F}}} [-s_{j,t}^{\mathrm{F}}(\boldsymbol{u}_{\text{des},t})]_+, \\ c_{\mathrm{A},t} &= \max_{j \in \mathcal{I}_t^{\mathrm{A}}} [-s_{j,t}^{\mathrm{A}}(\boldsymbol{u}_{\text{real},t})]_+. \end{aligned} \tag{14}$$

Both costs are normal-direction deficits in newtons, evaluated at the true pre-integration state. The command deficit measures conflict with the actuator-domain preimage; the realized-force deficit measures conflict with the physical rows.

Training imposes two average budgets:

$$\mathbb{E}[c_{\mathrm{F}}] \le \epsilon_{\mathrm{F}}, \qquad \mathbb{E}[c_{\mathrm{A}}] \le \epsilon_{\mathrm{A}}, \tag{15}$$

where $\epsilon_{\mathrm{F}} = 0.03$ N and $\epsilon_{\mathrm{A}} = 0.0015$ N. Within rollout batch k, the training reward is

$$r_{t,k}^{\text{PFCR}} = r_{\text{task},t} - \bar{\lambda}_{\mathrm{F},k} c_{\mathrm{F},t} - \bar{\lambda}_{\mathrm{A},k} c_{\mathrm{A},t}, \tag{16}$$

where each $\bar{\lambda}_{i,k}$ is held fixed throughout the rollout batch. After sampling, the dual state is updated using $\bar{c}_{i,k}$, the mean over all valid transitions.

We use a reversible, competence-gated anti-windup schedule for the dual variables. The schedule cannot be enabled before 0.8 million transitions have been collected. Thereafter, enabling requires an exponential moving average, with coefficient 0.9, of the per-rollout task-success fraction among completed episodes to reach 0.85 and remain at or above that threshold for five consecutive rollout updates. The schedule is disabled if the moving average falls below 0.75 for three consecutive updates. After each enabling event, the effective multiplier in the reward is ramped linearly from zero to its current value over 1.5 million transitions. Let $r_k \in [0, 1]$ denote this linear ramp factor and $z_{i,k}$

the integral state. During an enabled phase, the update is

$$z_{i,k+1} = \Pi_{[0,\lambda_{i,\max}]}\left[(1-\ell)z_{i,k} + r_k\eta_i[\bar{c}_{i,k} - \epsilon_i]_+ - \rho\eta_i[\epsilon_i - \bar{c}_{i,k}]_+\right]. \tag{17}$$

The parameters are $\eta_i = 0.002$, $\ell = 0.001$, $\rho = 5$, and $\lambda_{i,\max} = 1.5$, with $\bar{\lambda}_{i,k} = r_k z_{i,k}$. During a disabled phase, the integral state is multiplied by 0.98 after every rollout batch. This schedule avoids applying constraint pressure before the policy has acquired basic docking competence and releases accumulated dual pressure after a performance degradation. The schedule and DQP are separate components; configurations C and E use exactly the same schedule.

4.2. Local sensitivity of the operational QP active set

For compactness, at a fixed control step all rows of the QP in Eq. (11) are written as

$$A\boldsymbol{u} \geq \boldsymbol{b}, \tag{18}$$

where $A = \boldsymbol{G}_t$, $\boldsymbol{b} = \boldsymbol{b}_t$, and the time index is suppressed locally. Let the operational `quadprog` implementation of the Goldfarb–Idnani dual active-set method [19] return the optimum $\boldsymbol{u}_\star$, multipliers $\boldsymbol{\mu}$, and a basis active set $\mathcal{A}$. The corresponding Karush–Kuhn–Tucker (KKT) conditions are

$$\begin{aligned} \boldsymbol{u}_\star - \boldsymbol{u}_{\text{des}} - A^\top\boldsymbol{\mu} &= \boldsymbol{0}, \\ A\boldsymbol{u}_\star - \boldsymbol{b} &\geq \boldsymbol{0}, \\ \boldsymbol{\mu} &\geq \boldsymbol{0}, \\ \mu_j(A_j\boldsymbol{u}_\star - b_j) &= 0. \end{aligned} \tag{19}$$

Let $A_{\mathcal{A}}$ collect the active rows. If the active set and its rank remain locally unchanged, the projection can be written as

$$\Pi(\boldsymbol{u}) = \boldsymbol{u} + A_{\mathcal{A}}^\top\left(A_{\mathcal{A}}A_{\mathcal{A}}^\top\right)^\dagger(\boldsymbol{b}_{\mathcal{A}} - A_{\mathcal{A}}\boldsymbol{u}), \tag{20}$$

and hence

$$J_{\mathcal{A}} = \frac{\partial\Pi}{\partial\boldsymbol{u}} = I - A_{\mathcal{A}}^\top\left(A_{\mathcal{A}}A_{\mathcal{A}}^\top\right)^\dagger A_{\mathcal{A}}. \tag{21}$$

Proposition 1 (Fixed-active-set projection). *Suppose that the QP in Eq.* (11) *is feasible at a reference point, the rank of the active rows is constant in a neighborhood of that point, every active multiplier is strictly positive, and every inactive row has strictly positive slack. Then, within that neighborhood, the Euclidean projection is the affine map in Eq.* (20)*, with the Jacobian in Eq.* (21)*. Moreover, $J_{\mathcal{A}}$ is the symmetric idempotent projector onto the tangent space of the active constraints.*

Proof. With the active set fixed, the optimum satisfies $A_{\mathcal{A}}\boldsymbol{u}_\star = \boldsymbol{b}_{\mathcal{A}}$ together with KKT stationarity. Eliminating the multipliers gives Eq. (20), and differentiating with respect to $\boldsymbol{u}$ gives Eq. (21). The matrix $A_{\mathcal{A}}^\top(A_{\mathcal{A}}A_{\mathcal{A}}^\top)^\dagger A_{\mathcal{A}}$ is the orthogonal projector onto the active-row space; its complementary projector $J_{\mathcal{A}}$ is therefore symmetric and idempotent. Strict complementarity and positive inactive-row slack ensure that the active set is unchanged in a sufficiently small neighborhood. □

The numerical implementation recovers $J_{\mathcal{A}}$ from the active set and multipliers returned by the operational solver. It then performs the following checks for every sample:

- primal feasibility and active-row residuals use a tolerance of 10^{-7};
- the infinity norm of the stationarity residual must not exceed 10^{-6};
- every active multiplier must exceed 10^{-9};
- the minimum inactive-row slack must exceed 10^{-7}; and
- the symmetry, idempotence, and normal-annihilation errors of $J_{\mathcal{A}}$ must satisfy their numerical thresholds.

These checks implement the local regularity boundary required for KKT-based sensitivity rather than asserting differentiability at switches or degenerate solutions [21]. Samples that pass the KKT checks but do not satisfy strict separation are conservatively classified as *boundary or near-degenerate* samples. Samples that fail a KKT, finiteness, or projection-consistency check are marked as *invalid*. Neither category enters the DQP loss. When the QP applies no action correction, the active set is empty, $J = I_3$, and the projection residual is zero.

4.3. Fixed-active-set local projection-credit surrogate

During rollout collection, backup steps are marked projection-invalid and cannot enter the DQP loss. For each projection-valid primary-QP step, $\boldsymbol{u}_{\text{act},t} = \boldsymbol{u}_{\text{qp},t}$, and the reference-point tuple

$$(\boldsymbol{u}_{\text{des},t}, \boldsymbol{u}_{\text{act},t}, J_t) \tag{22}$$

is stored for every sample. During each proximal policy optimization (PPO) update, the clipped and scaled physical nominal thrust $\boldsymbol{u}_\theta(\boldsymbol{o}_t)$ is recomputed from the mean of the current policy distribution. Within the reference active-set cell, the local projection is

$$P_t(\boldsymbol{u}) = \boldsymbol{u}_{\text{act},t} + J_t\left(\boldsymbol{u} - \boldsymbol{u}_{\text{des},t}\right). \tag{23}$$

For the set $\mathcal{V}$ of samples numerically accepted as locally regular under the stated tolerances, the DQP auxiliary loss is

$$\mathcal{L}_{\text{DQP}}(\theta) = \frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}}\frac{1}{2u_{\max}^2}\left\|\boldsymbol{u}_\theta(\boldsymbol{o}_t) - P_t(\boldsymbol{u}_\theta(\boldsymbol{o}_t))\right\|_2^2. \tag{24}$$

Within the reference fixed-active-set cell, J_t is a tangent-space projector, so Eq. (24) retains only the active normal component removed by the QP and does not contract actions along the current feasible face. During a PPO update, the recomputed $\boldsymbol{u}_\theta$ is not guaranteed to remain in that cell; Eq. (24) is therefore used as an anchored local affine surrogate, not claimed as the global derivative of the piecewise-affine projection. The total policy-network objective retains the PPO policy loss:

$$\mathcal{L}_{\text{actor}} = \mathcal{L}_{\text{PPO}} + \beta_k\mathcal{L}_{\text{DQP}}, \tag{25}$$

where the final coefficient is 0.1, with a linear warm-up over the first ten PPO updates. During the first five updates, the implementation additionally records the ratio between the DQP and PPO gradient norms with respect to the policy parameters.

These diagnostics affect neither parameter updates nor checkpoint selection.

Gradients from Eq. (24) reach the policy parameters through the current command. The reference actions and Jacobian are constants; the actuator, dynamics, and backup law are outside this differentiation path.

Algorithm 1 summarizes the training loop. The validity mask applies only to the DQP auxiliary loss: backup steps and samples with rejected sensitivities remain available to the ordinary PPO update. Cached reference actions and Jacobians are held fixed during differentiation, whereas the current policy command is recomputed for each minibatch. An empty set of accepted samples gives a zero DQP loss.

4.4. Backup-controller system variant

Configurations C and E use the braking/hold backup in Appendix Appendix A.5. Configuration D leaves the learning algorithm, reward, physical domain, PPO procedure, dual schedule, and training budget unchanged, but replaces that backup controller with a robust saturated linear–quadratic regulator (LQR):

$$\boldsymbol{u}_{\mathrm{bk}} = \mathrm{sat}_{\mathcal{U}}\left[\boldsymbol{u}_{\mathrm{eq}} - K\left(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{hold}}\right)\right]. \tag{26}$$

In Eq. (26), $\boldsymbol{u}_{\mathrm{eq}}$ is the nominal equilibrium force that maintains zero velocity at the frozen CWH hold state. The hold state $\boldsymbol{x}_{\mathrm{hold}}$ freezes the current lateral position, sets the axial hold point to $\max(x, 3\ \mathrm{m})$, and assigns zero target-relative velocity. No online optimizer runs while the backup controller is active. The nominal LQR command consists of the equilibrium feedforward term and one fixed matrix–vector product; after componentwise saturation, the supervisor enforces the closed-form axial recovery floor, and unrecoverable geometry triggers the maximum positive axial-thrust command. Configuration D uses the same 512-step braking geometry and an eight-step minimum dwell, but activates proactively when the recovery margin is at most 0.15 m or the geometry is uncertifiable. It releases at a margin of at least 0.30 m and $v_x \geq -0.02\ \mathrm{m\,s^{-1}}$, provided geometry is certified. No primary QP runs during backup; one final LQR command is applied on the release step, and the QP resumes on the following step. The fixed gain is the discrete-time LQR gain for $A_d = I + \Delta t A_{\mathrm{CWH}}$, $B_d = \Delta t B_{\mathrm{CWH}}$, $Q = \mathrm{diag}(0.04, 0.04, 0.04, 4, 4, 4)$, and $R = 0.25 I_3$.

5. Experimental Design

5.1. Common CWH scenario and training configuration

Policy training, development-set selection, and the main CWH holdout comparison are conducted in the same actuator-mismatch scenario. All methods use identical CWH dynamics, initial-state distributions, actuator randomization, observation noise, CBF-QP, and task criteria. The base task reward is

$$\begin{aligned} r_{\mathrm{base},t} \coloneqq\ & 2\left(R_{t-1} - R_t\right) - 0.12\left\|\mathbf{u}_{\mathrm{real},t}\right\|_1 \Delta t - 0.03 \\ & + 200\,\mathbb{I}_{\mathrm{dock}} - 50\,\mathbb{I}_{\mathrm{constraint}} - 50\,\mathbb{I}_{\mathrm{leave}}. \end{aligned} \tag{27}$$

Here, $\mathbb{I}_{\mathrm{dock}} = 1$ only on a transition that terminates in task success, $\mathbb{I}_{\mathrm{constraint}} = 1$ whenever the minimum modeled post-step CBF margin is negative, and $\mathbb{I}_{\mathrm{leave}} = 1$ when the post-step range exceeds the 200 m leave radius. Docking and leaving terminate the episode, whereas a negative CBF margin alone does not; reaching the 1,200 s time limit truncates the episode separately. During training, a terminal-proximity state potential is added:

$$\Phi(\mathbf{x}) = 20 \exp\left[-\frac{R}{10} - \frac{1}{2}\left(\frac{V}{0.2}\right)^2\right]. \tag{28}$$

Combining the base reward in Eq. (27) with the potential in Eq. (28) gives the training-time task reward

$$r_{\mathrm{task},t} \coloneqq r_{\mathrm{base},t} + \gamma\Phi(\mathbf{x}_{t+1}) - \Phi(\mathbf{x}_t), \qquad \gamma = 0.999. \tag{29}$$

Evaluation does not use Eq. (29) either to select episodes or to compute the primary metrics.

The policy is trained with feedforward proximal policy optimization (PPO) [30]. The policy and value networks each contain two hidden layers with 256 units per layer. Four parallel environments each advance for 512 steps per rollout. The batch size is 2,048 transitions, with 10 optimization epochs per batch. The learning rate is 3×10^{-4}, the discount factor is $\gamma = 0.999$, the generalized-advantage-estimation parameter is 0.95, the clipping range is 0.2, the value-loss coefficient is 0.5, the maximum gradient norm is 0.5, and the entropy coefficient is zero. Observations are normalized as in Appendix Appendix A.1. Each configuration is trained with seeds 40–44; every seed receives 20.48×10^6 sampled transitions, and a complete checkpoint bundle is saved every 512×10^3 transitions.

5.2. Compared configurations

Table 1 defines the comparison. C/E differ only in DQP credit and form the primary method comparison; D measures the system-level effect of a different backup controller.

In total, the three configurations comprise 15 independent training runs and 307.2×10^6 sampled transitions. PFCR-DQP accounts for 102.4×10^6 of those transitions.

5.3. Checkpoint selection and evaluation

Eight candidate horizons from 1.024 to 20.480 million transitions are prescribed for each training seed. A candidate becomes eligible 1.5 million transitions after its most recent dual activation, leaving six evaluated horizons from 3.072 million onward. The development set contains 50 deterministic episodes per eligible candidate. Selection first requires zero recorded front-plane, certification, collision-proxy, and action-chain-consistency failures, then ranks candidates by protected docking success, velocity-envelope violation, mean impulse, and earlier horizon, in that order.

One checkpoint per training seed is selected before holdout evaluation. Each selected policy is evaluated for 200 episodes with initial states, actuator parameters, and observation-noise sequences paired across methods. The resulting CWH comparison contains 3,000 episodes. Supplementary Section S1 gives the full candidate grid, seed identifiers, and selected checkpoints.

Algorithm 1 PFCR-DQP training with local projection credit

Require: Policy π_θ, value function V_ϕ, interface budgets (ϵ_F, ϵ_A), and the fixed safety chain.

Initialize dual integral states $z_F = z_A = 0$ and the competence schedule (Section 4.1).
for each rollout update $k = 1, 2, \ldots$ **do**
  Clear the rollout buffer $\mathcal{B}$.
  for each control step t in the parallel training environments **do**
    Sample $a_t \sim \pi_\theta(\cdot \mid \boldsymbol{o}_t)$ and map the clipped action to physical thrust $\boldsymbol{u}_{\mathrm{des},t}$.
    Apply the nominal CWH CBF-QP and supervisor/backup to obtain $\boldsymbol{u}_{\mathrm{act},t}$; realize $\boldsymbol{u}_{\mathrm{real},t}$ through the actuator.
    Evaluate $c_{F,t}, c_{A,t}$ at the true pre-integration state using Eq. (14); propagate the CWH environment.
    Form $r_{t,k}^{\mathrm{PFCR}}$ using Eq. (16); store the transition and ordinary PPO statistics in $\mathcal{B}$.
    Set the DQP validity marker $m_t \leftarrow 0$.
    if the primary QP was updated and passes the projection-validity and local-regularity checks of Section 4.2 **then**
      Cache $(\boldsymbol{u}_{\mathrm{des},t}, \boldsymbol{u}_{\mathrm{act},t}, J_t)$ from the sampled QP solution; set $m_t \leftarrow 1$.
    end if
  end for
  Update the competence gate and dual states from completed episodes and valid interface costs (Section 4.1).
  Compute PPO advantages and returns from the stored rewards; set β_k using the ten-update warm-up in Section 4.3.
  for each PPO epoch and minibatch $\mathcal{M} \subset \mathcal{B}$ **do**
    Recompute $\boldsymbol{u}_\theta(\boldsymbol{o}_t)$ from the clipped and scaled current policy mean; keep cached reference quantities fixed.
    Let $\mathcal{V}_\mathcal{M}$ contain marked samples with finite required quantities; initialize $\mathcal{L}_{\mathrm{DQP}} \leftarrow 0$.
    if $\mathcal{V}_\mathcal{M} \neq \varnothing$ **then**
      $P_t(\boldsymbol{u}_\theta) \leftarrow \boldsymbol{u}_{\mathrm{act},t} + J_t(\boldsymbol{u}_\theta(\boldsymbol{o}_t) - \boldsymbol{u}_{\mathrm{des},t})$ for $t \in \mathcal{V}_\mathcal{M}$.
      $\mathcal{L}_{\mathrm{DQP}} \leftarrow \dfrac{1}{|\mathcal{V}_\mathcal{M}|} \displaystyle\sum_{t \in \mathcal{V}_\mathcal{M}} \dfrac{\|\boldsymbol{u}_\theta(\boldsymbol{o}_t) - P_t(\boldsymbol{u}_\theta)\|_2^2}{2u_{\max}^2}$.
    end if
    Update the policy with $\mathcal{L}_{\mathrm{PPO}} + \beta_k \mathcal{L}_{\mathrm{DQP}}$ and the value function with the standard PPO value loss.
  end for
end for
Deployment: use the policy mean and the same safety/actuator chain; omit the DQP Jacobian and auxiliary loss.

Table 1: Controller configurations.

Code	Method	Learning change	Backup control	Relation to E
C	PFCR-RC-AW	Barrier-aligned dual-interface costs and reversible anti-windup duals	Recoverable braking/hold	Direct reference; identical except for DQP
D	PFCR-RC-AW + robust saturated LQR	Same as C	Robust saturated LQR	System variant; the safety supervisor also changes
E	PFCR-DQP	C plus fixed-active-set local projection-credit surrogate	Same as C	Proposed method

5.4. Nonlinear orbital transfer

The same 15 selected policies and saved observation normalizers are evaluated under Tudat-T0 and T1 without additional training or selection. The controller continues to use CWH for its QP and recovery calculations, while Tudat propagates the plant as defined in Section 3.2. Each method, training seed, and truth level has 100 episodes, giving 3,000 runs. Initial conditions and actuator draws are paired across methods; T1 additionally pairs its truth-force error sequence. Supplementary Section S1 records the pairing and numerical-validation details.

5.5. Metrics and statistical units

The five training seeds are the statistical replicates. Episode quantities are first averaged within each seed, and transition quantities are pooled within each seed, before computing summaries across seeds. For Tudat, this aggregation is performed separately at each truth level.

The primary metrics are as follows:

- Task-success rate: the episode reaches the environment's terminal capture flag under the post-integration capture and terminal front-plane conditions defined in Section 3.
- Protected docking success rate: the episode is a task success and contains no segment-based capture-sphere collision proxy, front-plane-buffer violation, or uncertified supervisor sample, using the operational definitions in Section 3. Speed-envelope excursions are reported separately and do not disqualify it.
- Velocity-envelope violation rate: the fraction of valid transitions satisfying $V > \nu_0 + \nu_1 R$.
- Runtime-safety-layer modification rate: the fraction of valid transitions satisfying $\|\boldsymbol{\delta}_F\|_2 > 10^{-5}$ N. This threshold records a numerically discernible change between the

policy and supervisor commands. The same event defines the state-conditioned probabilities and local time-window fractions. It equals the numerical primary-QP correction rate when no backup command is used, as in the C/E CWH holdout, but includes backup-command modifications when a backup controller is active.

- Safety-layer override profile: the exceedance function $\Pr[\|\boldsymbol{\delta}_{\mathrm{F}}\|_2 > \tau]$ evaluated over the whole per-axis authority range $\tau \in [0, u_{\max}]$. The preceding rate is its $\tau \to 0$ limit and counts a 10^{-5} N trim and a 0.9 N override as the same event, so the profile is reported as a curve rather than at a single chosen threshold. Two threshold-free summaries accompany it: the median correction magnitude, and the composition of the corrected steps themselves into magnitudes below 0.05 N, between 0.05 and 0.5 N, and above 0.5 N. Where a single reading is quoted, the half-authority anchor $\tau = 0.5\,\mathrm{N} = u_{\max}/2$ is used.

- Barrier-interface cost: the transition-weighted mean of Eq. (14).

- Axis-summed actuator-output control impulse (reported as total impulse in the figures and tables):

$$J^{\mathrm{imp}} = \sum_t \left\|\mathbf{u}_{\mathrm{real},t}\right\|_1 \Delta t. \tag{30}$$

  This control-use metric is computed from the actuator-model output in every evaluation. In Tudat-T1 it is therefore neither the impulse of $\boldsymbol{F}^{H}_{\mathrm{truth}}$ nor a propellant-consumption impulse.

CWH paired changes are computed within training seed. Descriptive 95% intervals enumerate the $5^5 = 3125$ possible resampled five-seed means. Tudat uses Student-t intervals with four degrees of freedom, including paired T1–T0 changes. Section 7.4 discusses the statistical scope.

6. Results

6.1. Learning behavior and late-training degradation

Figure 2 follows the matched C/E comparison across all six eligible development checkpoints. At 4.096 million transitions, PFCR-DQP reaches 98.4% mean protected docking success with 54.78% safety-layer modification and 86.4 N s impulse; C reaches 96.0%, 65.96%, and 184.4 N s. This early region combines task completion with lower conflict and control use.

Performance subsequently deteriorates. At 20.480 million transitions, the mean success rates are 36.8%, 41.6%, and 9.6% for C, D, and E. PFCR-DQP's mean impulse rises to 576.3 N s, although its modification rate falls to 51.12%. Reduced intervention alone therefore does not indicate successful learning. Every selected checkpoint lies between 3.072 and 5.120 million transitions; the holdout results below describe these selected policies. The complete three-method selection surface and checkpoint table appear in Supplementary Section S1.

6.2. Availability of local projection credit

Across 102.4 million DQP training transitions, 66.78% pass the numerical local-regularity checks and contribute to the auxiliary loss. Of these, 69.45% have at least one active constraint. The mean active rank is 0.814 and the mean Jacobian trace is 2.186. Thus the training signal frequently removes a constrained normal component while leaving tangential degrees of freedom.

All operational QP solutions and projection diagnostics are valid. Over the first five updates of each seed, the DQP-to-PPO gradient-norm ratio ranges from 0.00406 to 0.04095. Supplementary Section S2 reports the full regularity, KKT, and gradient diagnostics.

6.3. Control performance of the selected policies

Table 2 summarizes the holdout results across the five training seeds. PFCR-DQP attained a mean protected docking success rate of 99.5%, with 98.5% for its worst seed; the corresponding means for PFCR-RC-AW and the LQR variant were 97.6% and 96.3%, respectively. PFCR-DQP's mean runtime-safety-layer modification rate, counting every correction above the 10^{-5} N discernibility threshold, was 54.84%, approximately 12.5 percentage points below those of the two reference configurations; Section 6.4 resolves that count by override magnitude. Its mean total impulse was 77.2 N s, equal to 37.5% and 27.7% of the values for configurations C and D, respectively.

The paired changes for PFCR-DQP relative to configuration C are listed in Table 3. Runtime-safety-layer modification, c_{F}, and c_{A} improved for all five seeds, whereas total impulse improved for four of the five seeds. The mean protected docking success gain was 1.9 percentage points, but the median paired change was zero, the descriptive interval crossed zero, and only two of the five seeds improved strictly. This pattern is consistent with a reference configuration whose seed-level success rates already cluster near the upper bound. The mean velocity-envelope violation rate was slightly lower, but only two of the five seeds improved; this result likewise does not support a claim of a consistent reduction.

Across the 1,000 fully paired episodes, configuration E recorded 24 protected docking success gains, five losses, and 971 unchanged outcomes relative to C. Relative to D, it recorded 35 gains, three losses, and 962 unchanged outcomes. These counts describe which paired initial conditions change outcome. Figure 3 presents the full holdout outcome and paired-change evidence.

6.4. Magnitude of safety-layer overrides

The 10^{-5} N event used in Table 2 records any discernible correction, so it assigns the same weight to a numerical trim and to an override of a full axis command. Figure 4 resolves the same frozen transitions by correction magnitude across the entire per-axis authority range $u_{\max} = 1$ N.

PFCR-DQP's exceedance curve lies below both references at every threshold, and its $\tau \to 0$ limit reproduces the rates already reported in Table 2. At the half-authority anchor $\tau = 0.5$ N, the safety layer seized more than half of a single axis's thrust authority on 5.12% of PFCR-DQP steps, against 35.94% for

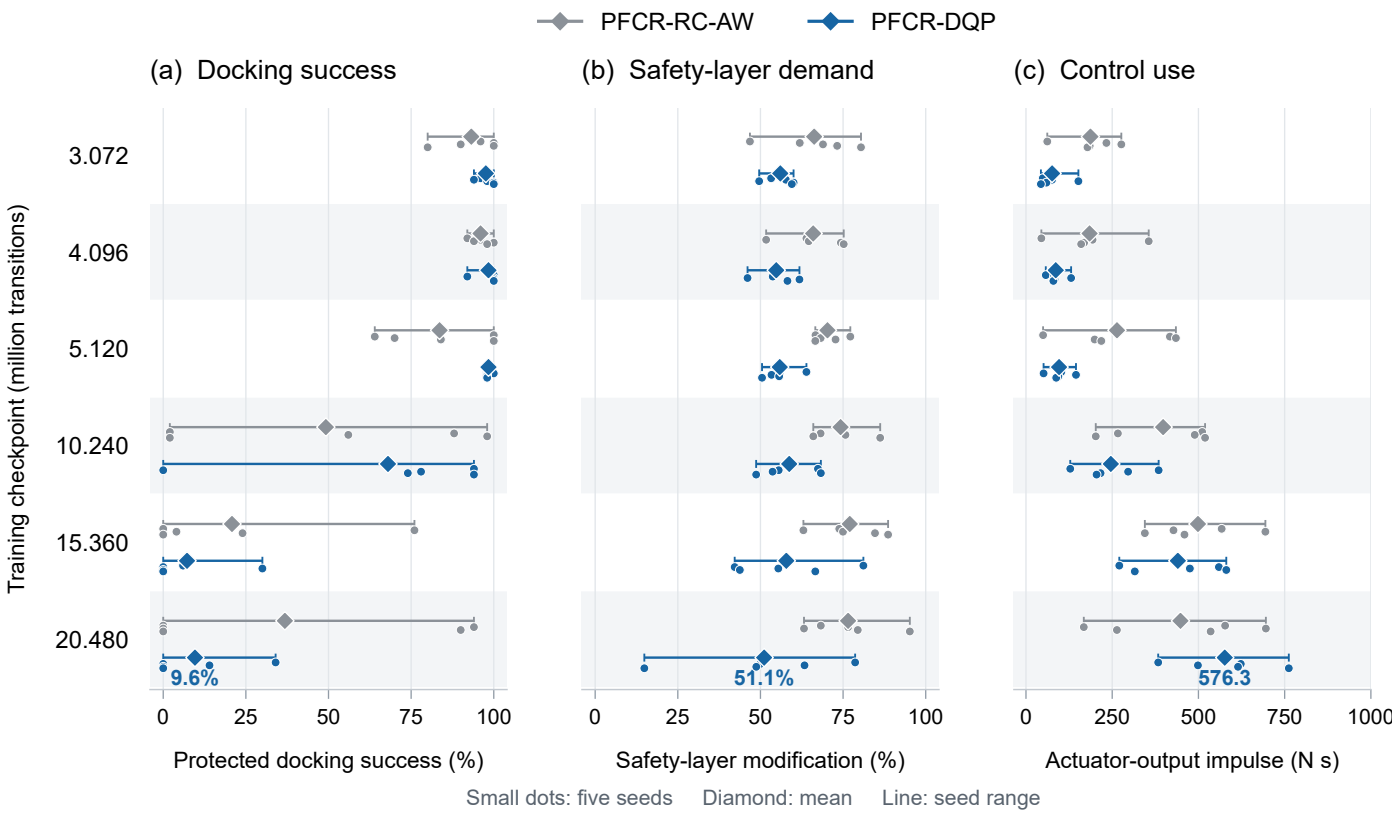


Figure 2: Development-set performance at six training checkpoints for the matched C/E comparison. Each row groups one checkpoint, with PFCR-RC-AW in gray and PFCR-DQP in blue. Small points show the five training seeds, diamonds their arithmetic mean, and horizontal lines their minimum-to-maximum range. Points are offset vertically for visibility; their horizontal coordinates are the measured values. Each seed aggregates 50 deterministic episodes. Checkpoint rows are equally spaced and labeled by their actual training horizons. The ranges describe seed dispersion, not confidence intervals. Final-checkpoint DQP means are annotated. The modification event is $\|\boldsymbol{\delta}_{\mathrm{F}}\|_2 > 10^{-5}$ N.

Table 2: Five-training-seed results on the frozen holdout. Each cell reports the seed mean followed by the seed median [interquartile range].

Method	Protected docking success [%]	Velocity-envelope violation [%]	Safety-layer modification [%]	Total impulse [N s]	c_{F} [N]	c_{A} [N]
PFCR-RC-AW	97.60; 99.50 [97.00, 100.00]	0.587; 0.423 [0.095, 0.712]	67.30; 66.81 [64.69, 67.00]	205.9; 216.5 [170.7, 225.1]	0.3922; 0.4535 [0.4054, 0.4628]	0.003389; 0.003489 [0.002379, 0.004207]
PFCR-RC-AW + LQR	96.30; 98.50 [93.00, 99.00]	1.120; 0.559 [0.316, 0.761]	67.52; 67.54 [67.00, 68.40]	278.9; 277.4 [248.0, 289.2]	0.4360; 0.4367 [0.4216, 0.4446]	0.004059; 0.003778 [0.003276, 0.004103]
PFCR-DQP	99.50; 100.00 [99.00, 100.00]	0.445; 0.559 [0.226, 0.655]	54.84; 56.05 [55.40, 58.16]	77.2; 59.0 [57.3, 89.4]	0.1102; 0.1127 [0.0927, 0.1280]	0.001223; 0.000917 [0.000798, 0.001890]

Table 3: Favorable paired-seed changes for PFCR-DQP. Runtime-safety-layer modification, half-authority override, and velocity-envelope quantities are in percentage points; the half-authority override event is $\|\boldsymbol{\delta}_F\|_2 > 0.5$ N. Intervals are descriptive exhaustive-bootstrap intervals for the five-seed mean. Changes are computed from unrounded seed values, so the last displayed digit can differ from a subtraction of rounded marginal means.

Reference	Metric	Mean favorable change	Median change	Descriptive 95% interval	Improved seeds
C	Protected docking success	+1.90 pp	0.00 pp	[−0.80, +5.40] pp	2/5
C	Velocity-envelope-violation reduction	+0.142 pp	−0.061 pp	[−0.387, +0.858] pp	2/5
C	Safety-layer-modification reduction	+12.46 pp	+11.42 pp	[+9.05, +16.19] pp	5/5
C	Half-authority-override reduction	+30.83 pp	+38.84 pp	[+14.46, +42.44] pp	4/5
C	Median-override-magnitude reduction	+0.2995 N	+0.3737 N	[+0.1540, +0.4213] N	5/5
C	Total-impulse reduction	+128.7 N s	+127.1 N s	[+51.6, +195.2] N s	4/5
C	c_F reduction	+0.2820 N	+0.3318 N	[+0.1484, +0.3656] N	5/5
C	c_A reduction	+0.002167 N	+0.001599 N	[+0.001126, +0.003326] N	5/5
D	Protected docking success	+3.20 pp	+1.00 pp	[+0.40, +6.00] pp	4/5
D	Safety-layer-modification reduction	+12.67 pp	+10.24 pp	[+9.05, +18.42] pp	5/5
D	Half-authority-override reduction	+43.28 pp	+44.72 pp	[+40.50, +45.31] pp	5/5
D	Median-override-magnitude reduction	+0.4411 N	+0.4381 N	[+0.4023, +0.4826] N	5/5
D	Total-impulse reduction	+201.7 N s	+195.8 N s	[+151.5, +262.6] N s	5/5

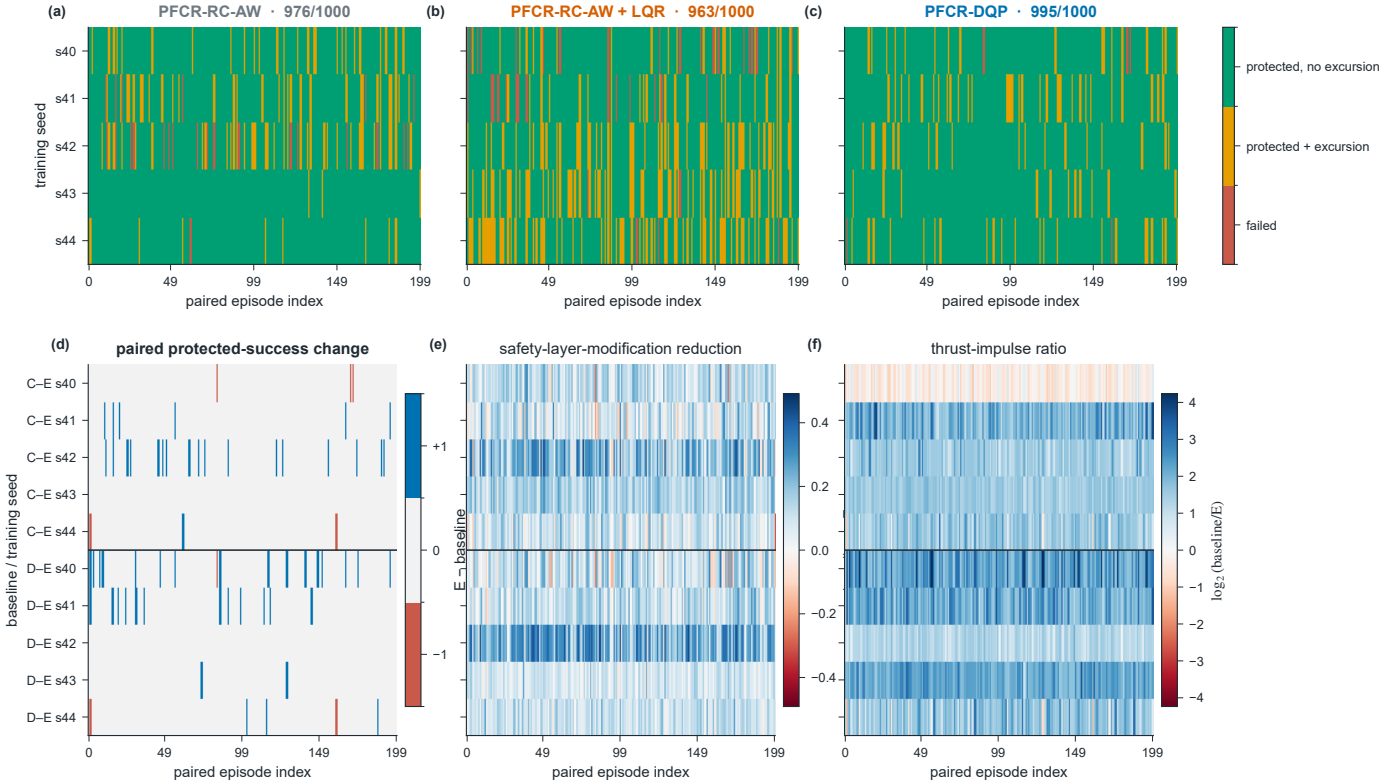


Figure 3: All 3,000 holdout episodes for the three methods and episode-wise paired changes of PFCR-DQP relative to configurations C and D. The upper row retains failures and protected docking successes with or without velocity-envelope violations. The lower row shows changes in protected-docking-success status, reductions in runtime-safety-layer modification ($\|\boldsymbol{\delta}_F\|_2 > 10^{-5}$ N), and actuator-output impulse ratios.

configuration C and 48.40% for configuration D: reductions by factors of 7.0 and 9.5. At full per-axis authority the corresponding rates are 0.06%, 8.41%, and 9.83%. The median correction magnitude was 0.0312 N for PFCR-DQP, 0.3307 N for C, and 0.4723 N for D.

The composition of the corrections differs accordingly. Of the steps that configuration C corrected at all, 52.5% seized more than half the per-axis authority and 6.9% remained below 0.05 N; for PFCR-DQP the same shares are 9.2% and 24.8%. The three controllers therefore differ less in how often the safety layer engages than in what it does when it engages. PFCR-DQP is corrected often and slightly, which is the behavior of a policy operating along the velocity-envelope boundary rather than one repeatedly braked away from it; the accompanying impulse reduction in Table 2 is the control-use consequence of approaching at the admissible speed.

Table 3 gives the paired-seed changes. The half-authority override reduction holds on four of the five seeds against C and on all five against D, and the median-magnitude reduction holds on all five seeds against both references. The single exception is seed 40, whose C policy already overrode on 6.93% of steps against 7.29% for its DQP counterpart. Configuration C's seed-level half-authority rates span 6.93% to 48.85%, whereas PFCR-DQP's span 3.27% to 7.29%; the reference's policy–filter coordination is therefore strongly seed-dependent where the DQP policies are not.

These readings are a post-hoc descriptive re-read of the frozen holdout transitions. They introduce no new safety guarantee, and they do not imply that corrections below any threshold are absent.

6.5. Safety boundaries and backup control

None of the three holdout configurations exhibited a front-plane crossing, a front-plane-buffer violation, an uncertified supervisor sample. Configuration C had one capture-sphere collision-proxy episode, so its mean task-success rate was 97.7%, whereas its mean protected docking success rate was 97.6%. Task success and protected docking success were identical for configurations D and E. None of the three configurations experienced a primary-QP solution failure.

Configurations C and E did not activate backup control on the holdout. The robust LQR in configuration D was active for 14,597 control steps, with 1,853 takeovers and 1,816 releases; its uncertified-supervisor-sample count was zero. This more assertive backup controller preserved the front-plane-buffer constraint, but it was accompanied by greater impulse and a lower mean protected docking success rate. Complete takeover, release, and recoverability telemetry is provided in Supplementary Fig. S1.

6.6. State occupancy and safety-layer modification

Across the 1,375,800 valid holdout transitions, the minimum post-step x over the three configurations was no less than 0.368 m, above the 0.25 m front-plane buffer. Supplementary Fig. S2 provides the full trajectory and time–distance occupancy views without selecting episodes or adding post-termination padding.

Figure 5 locates safety-layer modifications in the distance–speed plane using the same event as Table 2, $\|\boldsymbol{u}_{\mathrm{act}} - \boldsymbol{u}_{\mathrm{des}}\|_2 > 10^{-5}$ N. The blue grid shows episode-balanced state visitation, and the overlaid points show the conditional modification probability in each occupied cell. Every occupied cell is retained. Training seeds are equally weighted, episodes are equally weighted within each seed, and valid transitions are equally weighted within each episode. These conditional probabilities describe where modifications occur; the overall rates and paired-seed differences are reported in Tables 2 and 3.

The lower row traces the continuous slack from $s^{\mathrm{F}}_{\mathrm{min}}(\boldsymbol{u}_{\mathrm{des}})$ to $s^{\mathrm{A}}_{\mathrm{min}}(\boldsymbol{u}_{\mathrm{real}})$. Equation (14) defines $c_{\mathrm{F}} = [-s^{\mathrm{F}}_{\mathrm{min}}(\boldsymbol{u}_{\mathrm{des}})]_+$, so the mean nominal-command deficit in Table 2 measures the magnitude of conflict along constraint normals. Its reduction from 0.3922 to 0.1102 N is the continuous counterpart of the override-magnitude profile in Section 6.4. The slack distribution also retains realized-thrust samples with negative slack; these continuous measurements do not imply that every realized command is strictly feasible.

6.7. Closed-loop action and slack evolution

Figure 6 follows the C/E controllers from the same initial state and actuator draws, using the first holdout episode of seed 40.

The centerline is a 15-step rolling median and the band spans the local 10th–90th percentiles, showing within-episode oscillation. This fixed example illustrates the timing of command deficits, QP corrections, and realized thrust; the paired aggregate results are given in Tables 2 and 3.

6.8. Transfer to nonlinear orbital propagation

Table 4 reports protected docking success rates for the frozen control stack under the two Tudat truth-model fidelity levels. PFCR-DQP achieved 98.4% in T0 and 97.4% in T1; PFCR-RC-AW achieved 94.0% and 94.4%, respectively. The five-seed paired E–C difference was +4.4 percentage points [−2.7, 11.5] in T0 and +3.0 percentage points [−7.6, 13.6] in T1. Both intervals cross zero, leaving the mean success difference uncertain.

The within-seed T1–T0 changes were +0.4 percentage points [−1.9, 2.7] for C, +0.6 percentage points [−1.7, 2.9] for D, and −1.0 percentage points [−2.8, 0.8] for E; all three intervals cross zero. In T1, E had a runtime-safety-layer modification rate of 54.63% at the 10^{-5} N discernibility threshold and an axis-summed actuator-output control impulse of 99.4 N s. The corresponding values were 67.69% and 259.8 N s for C and 67.10% and 359.7 N s for D. The override profile of Section 6.4 is reported for the CWH holdout only. Impulse is measured at the actuator output as defined in Eq. (30). Neither fidelity level exhibited a capture-sphere collision proxy, a front-plane-buffer violation, primary-QP infeasibility, an uncertified supervisor sample, or Tudat integration failure. Velocity-envelope excursions are measured separately. Figure 7 shows the cross-fidelity retention and the same-seed changes.

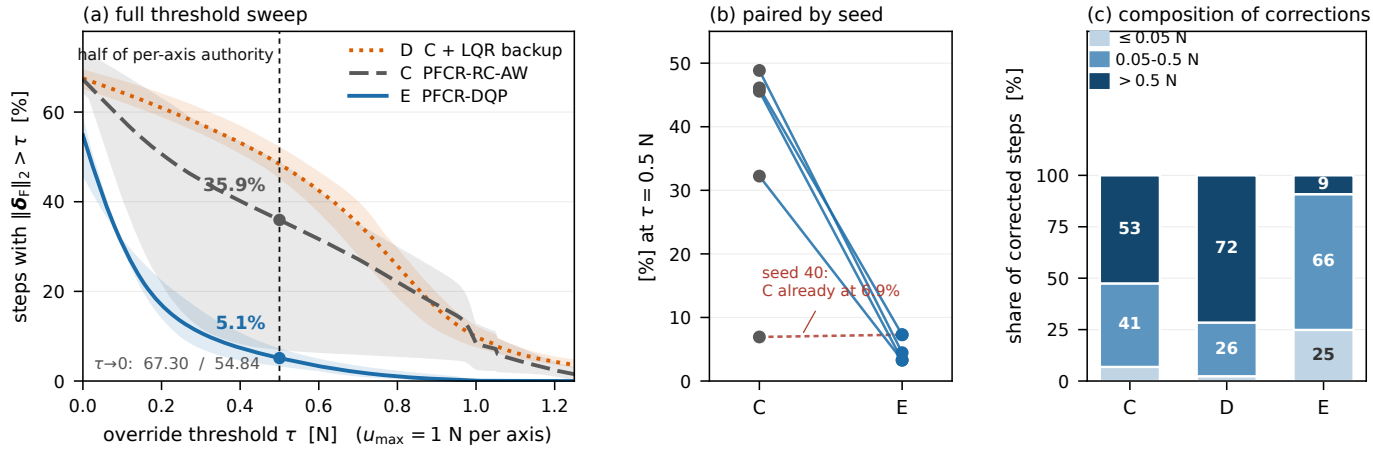


Figure 4: Safety-layer override magnitude on the frozen CWH holdout. (a) Exceedance of the correction norm $\|\boldsymbol{\delta}_F\|_2$ over the full per-axis thrust authority; lines are five-seed means and bands their minimum-to-maximum range. The $\tau \to 0$ limit reproduces the modification rates of Table 2, and the dashed rule marks the half-authority anchor. (b) Per-seed values at that anchor, with configurations C and E paired within training seed. (c) Magnitude composition of the corrected steps alone, so each bar sums to 100%. Training seeds are equally weighted and episodes are equally weighted within each seed. No evaluation was rerun and no episode was selected.

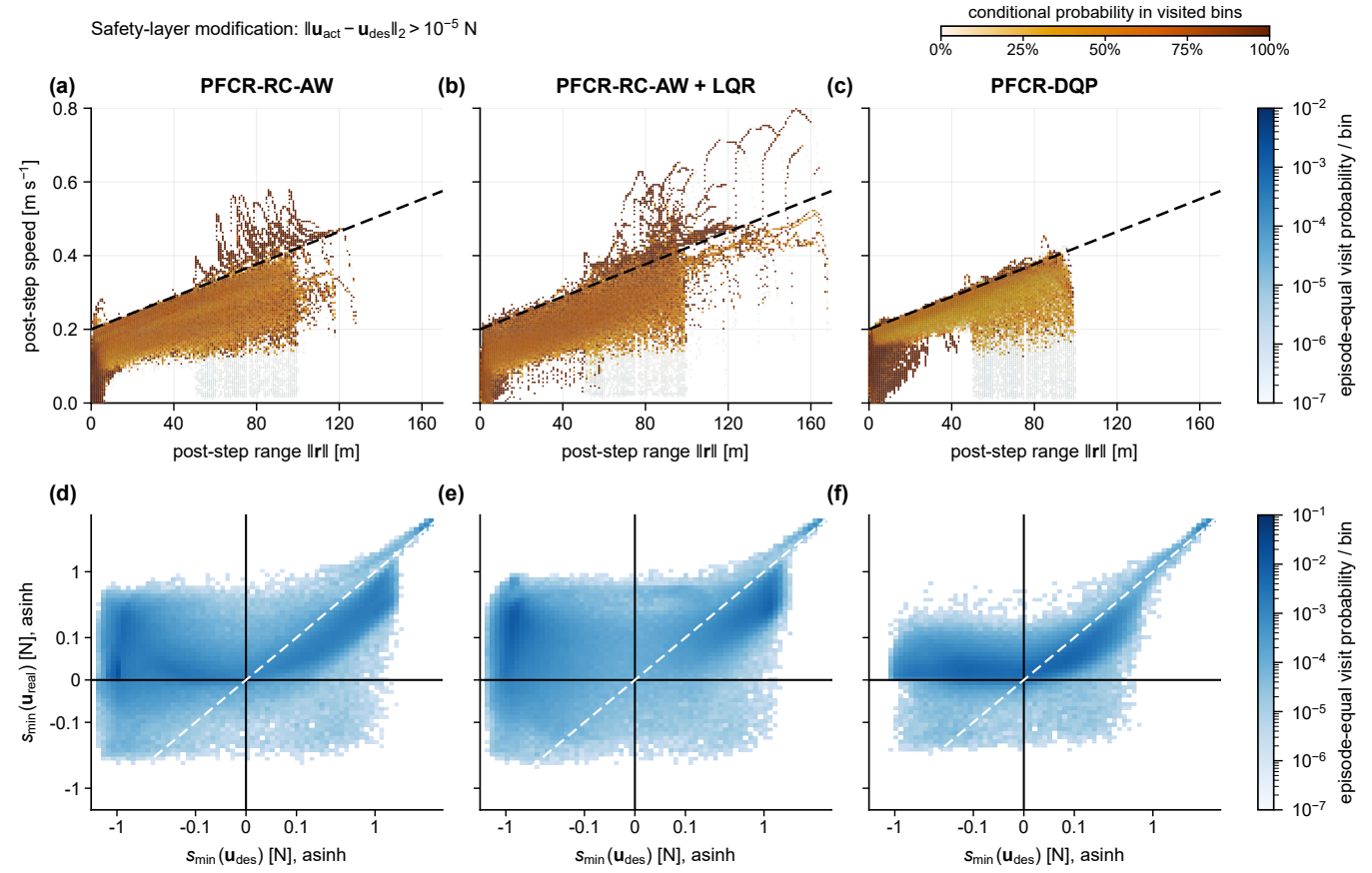


Figure 5: State occupancy, safety-layer modification, and continuous CBF-slack transfer on the frozen CWH holdout. In the upper row, blue cells show episode-balanced distance–speed visitation probability and overlaid points show the conditional probability of $\|\boldsymbol{u}_{act} - \boldsymbol{u}_{des}\|_2 > 10^{-5}$ N in every occupied cell. The lower row shows the joint distribution of minimum normalized CBF slack for nominal and realized thrust under an inverse-hyperbolic-sine transform with scale 0.05 N. Black lines mark zero slack and the dashed diagonal is the identity line. The five training seeds and the 200 episodes within each seed are equally weighted; all valid transitions contribute.

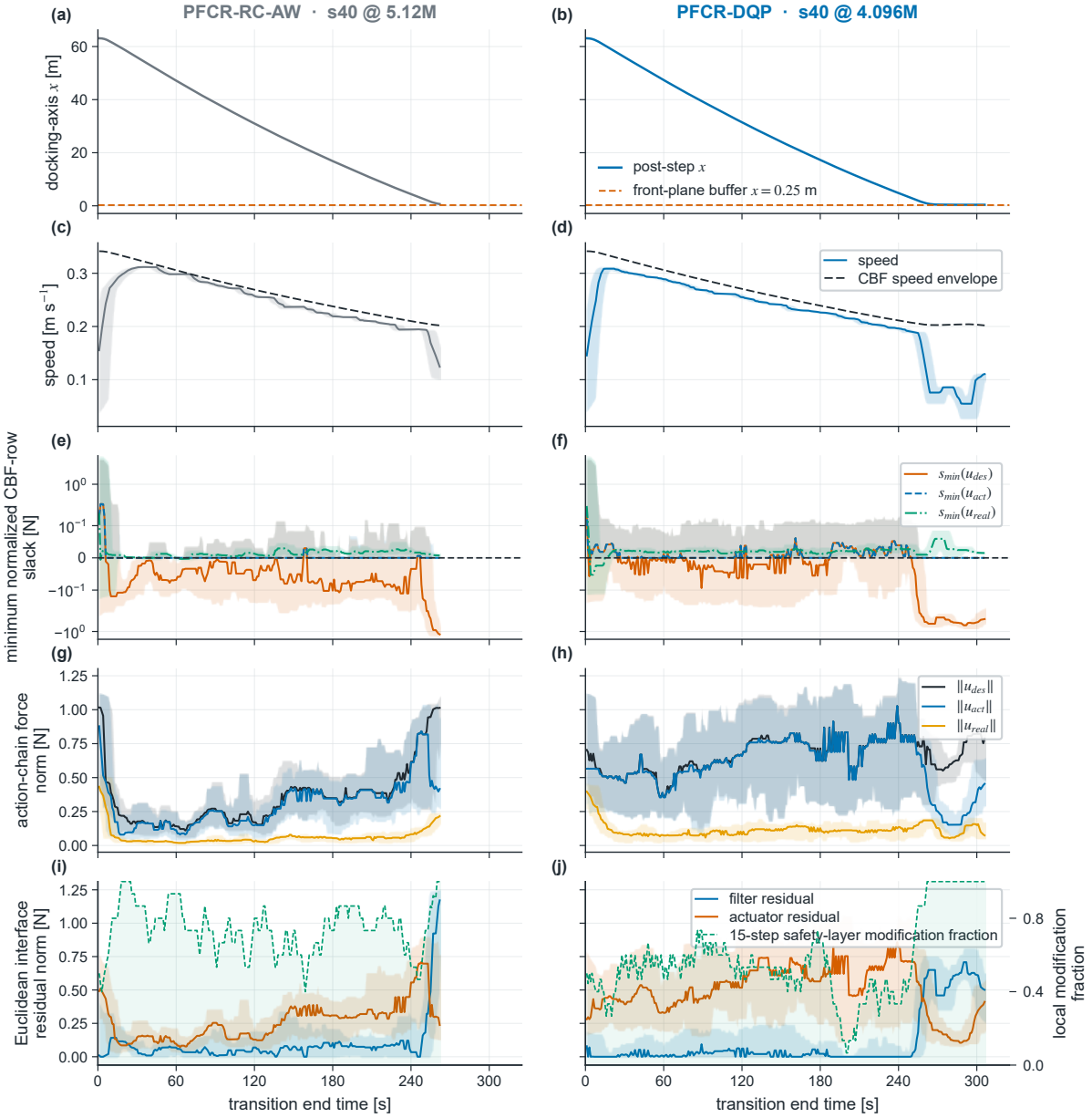


Figure 6: Closed-loop signals for PFCR-RC-AW and PFCR-DQP under the same initial state and actuator parameters. All raw transitions contribute to the ribbon calculation, with no post-termination padding; the exact unsmoothed waveforms are retained in Supplementary Fig. S3. The bottom row is the 15-step local mean of the modification indicator $\mathbf{1}_{\{\|\delta_F\|_2 > 10^{-5}\ \mathrm{N}\}}$.

Table 4: Mean protected docking success rate [%] over five training seeds for the frozen checkpoints in Tudat-T0 and Tudat-T1. Brackets give two-sided 95% Student-t intervals with four degrees of freedom. Each seed, method, and truth-model fidelity contains 100 paired episodes.

Truth-model fidelity	PFCR-RC-AW (C)	PFCR-RC-AW + LQR (D)	PFCR-DQP (E)
Tudat-T0	94.0 [88.5, 99.5]	92.4 [87.2, 97.6]	98.4 [95.3, 100.0]
Tudat-T1	94.4 [86.8, 100.0]	93.0 [88.1, 97.9]	97.4 [93.1, 100.0]

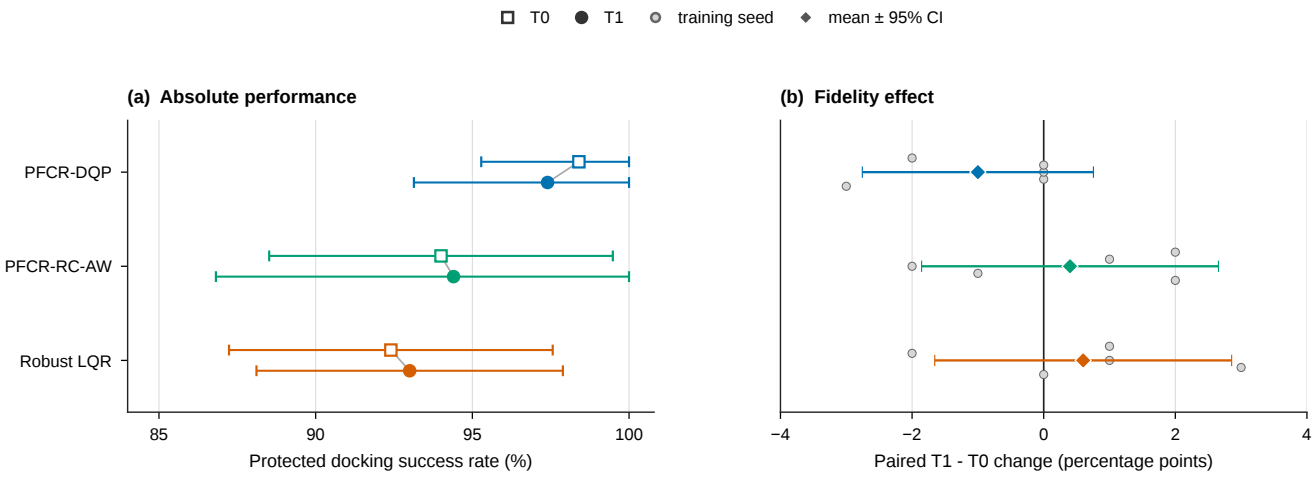


Figure 7: Protected docking success retention from Tudat-T0 to the representative Tudat-T1 model. (a) Open squares and filled circles show the five-training-seed means for T0 and T1, respectively; whiskers are two-sided 95% Student-t intervals with four degrees of freedom, clipped to 0%–100%. (b) Gray points show the five same-seed T1–T0 differences; the colored diamond and whisker show their mean and 95% t interval. Each training seed aggregates 100 paired evaluation episodes at each truth-model fidelity.

7. Discussion

7.1. What projection credit changes

The selected C/E policies show a consistent reduction in override magnitude and in both interface costs across the five seeds. Neither controller activates its backup on the CWH holdout, so these changes reflect coordination between the nominal policy and the primary QP. The override profile, the state-conditioned modification probabilities, the continuous slack distributions, and the paired control traces all locate the effect in the same place: what changes is not principally how often the QP engages, but how far it must move the command when it does. PFCR-DQP's corrections concentrate below 0.05 N, whereas configuration C's remain concentrated above half the per-axis authority. The lower impulse accompanies this change while capture performance remains high.

This is the sense in which the trained policy is post-filter consistent. It does not earn a smaller correction count by retreating from the constraint; it operates along the velocity-envelope boundary, where the admissible approach speed is highest and the filter's corrections are correspondingly small. A single binary modification rate cannot separate those two behaviors, because it counts a boundary-tracking trim and a braking override as the same event.

The effect is strongest in conflict and control use. Protected docking success is already close to its upper bound for the reference: its median is 99.5%, and the paired success interval includes zero. Impulse improves for four of the five seeds; seed 40 uses 8.8 N s more than C. These results support improved policy–filter coordination more clearly than a general success-rate advantage.

Configuration D shows how supervisor design affects the same metrics. Its LQR backup executes 1,853 takeovers with no uncertified samples, but uses more impulse and achieves a lower mean success rate. The holding law, activation region, and release hysteresis therefore merit evaluation together. The matched C/E comparison in Table 1 isolates the learning change.

7.2. Late-training degradation

The development curves show a major limitation of the present training procedure. PFCR-DQP reaches a mean protected docking success rate of 98.4% at both 4.096 and 5.120 million transitions, then falls to 9.6% at 20.480 million. Over the interval from 4.096 to 20.480 million, its mean impulse increases from 86.4 to 576.3 N s. The final safety-layer modification rate remains only 51.12%. A small correction rate can therefore coexist with poor task completion.

All three configurations degrade after their early high-performance checkpoints, and the final DQP success rate is lower than either reference. Local projection credit improves the selected policies but does not stabilize the full training trajectory. The current records do not isolate whether continued policy updates, the dual schedule, or their interaction drives this decline. Checkpoint selection is consequently consequential to the reported performance, and a mechanism study of the late training phase is needed before claiming sustained convergence.

7.3. Locality and control-model assumptions

The projection model is exact only inside the reference active-set region. A PPO update may move the current command outside that region, where the cached affine map is a surrogate for the new QP projection. Numerical acceptance covers 66.78% of training transitions; the remaining samples receive ordinary PPO and dual-cost credit. More complex thruster allocation or obstacle geometry may increase redundancy and active-set switching, reducing the fraction for which this local construction is useful.

The front-plane braking bound in Appendix Appendix A uses forward-Euler CWH dynamics and the declared actuator range, with a drift bound assumed to hold through braking. The speed constraints are evaluated at one-second samples. These assumptions do not establish continuous-time invariance, rigid-body contact safety, or protection against arbitrary actuator failures. The observation model also supplies true-state constraint margins and noise-free realized-force history, although the state block itself is noisy. Replacing these signals with measured or estimated quantities would change both the information available to the policy and the supervisor uncertainty model [27, 28].

Tudat-T0/T1 test transfer to a different propagation model with the CWH controller retained. The high success rates and small, uncertain T1–T0 mean changes are empirical results for the evaluated perturbations. A formal nonlinear safety guarantee would require bounds consistent with that plant. Mission-specific navigation, attitude coupling, thruster calibration, and space-weather conditions remain outside the present model.

7.4. Statistical scope

With five training seeds, intervals for success differences are broad despite the many paired episodes within each seed. The estimates describe the trained policies under a fixed task distribution. Independent seeds and new task distributions are needed to test how well the observed reductions in override magnitude and continuous interface costs generalize. The half-authority override reduction rests on four of the five seeds against configuration C, because that reference's own coordination varies widely by seed.

8. Conclusions

PFCR-DQP uses the local geometry of an operational CBF-QP to improve policy credit assignment. The auxiliary loss acts on the active constraint-normal residual of a sampled affine projection model, alongside task learning and command/actuator interface costs.

On the selected-policy CWH holdout, the method reduces the steps at which the safety layer seizes more than half of the per-axis thrust authority from 35.94% to 5.12%, the median correction magnitude from 0.3307 to 0.0312 N, the nominal-command deficit from 0.3922 to 0.1102 N, and the actuator-output impulse from 205.9 to 77.2 N s, with 99.5% protected docking success. The same policies reach 98.4% and 97.4% success under Tudat-T0 and T1. The development curves also show severe late-training degradation, so the control gains of the selected policies should be distinguished from training stability. Improving that stability while retaining the directional credit mechanism is the next problem.

Appendix A. Controller and observation interface

Appendix A.1. Actuator update and policy input

Let $z_t = \boldsymbol{u}_{\mathrm{real},t-1}$, with $z_0 = \mathbf{0}$. For a scalar command w, define the dead-zone map

$$D_d(w) = \begin{cases} 0, & |w| < d, \\ w, & |w| \geq d. \end{cases} \tag{A.1}$$

The map acts componentwise. With $S_\sigma = \mathrm{diag}(\sigma_x, \sigma_y, \sigma_z)$, the actuator update is

$$\begin{aligned} a_\tau &= \begin{cases} 1, & \tau \leq 10^{-9}\ \mathrm{s}, \\ \Delta t/(\tau + \Delta t), & \tau > 10^{-9}\ \mathrm{s}, \end{cases} \\ \boldsymbol{u}_{\mathrm{real},t} &= \mathrm{sat}_{\mathcal{U}}\left[(1 - a_\tau)z_t + a_\tau S_\sigma D_d(\boldsymbol{u}_{\mathrm{act},t})\right], \\ z_{t+1} &= \boldsymbol{u}_{\mathrm{real},t}. \end{aligned} \tag{A.2}$$

Thus dead zone precedes scaling, lag, and output saturation. The saturation output is also the lag state for the next step. At reset, τ, the three σ_i, and d are drawn independently and uniformly from the ranges in Section 3.3 and held for the episode.

All three configurations use the same 36-component input. Before statistical normalization, its ordered blocks are

$$\boldsymbol{o}_t^{\mathrm{raw}} = \begin{bmatrix} \tilde{\boldsymbol{r}}_t/(100\ \mathrm{m}) \\ \tilde{\boldsymbol{v}}_t/(1\ \mathrm{m\,s^{-1}}) \\ \mathrm{col}_{j=t-4}^{t-1}(\boldsymbol{u}_{\mathrm{des},j}/u_{\max}) \\ \mathrm{col}_{j=t-4}^{t-1}(\boldsymbol{u}_{\mathrm{real},j}/u_{\max}) \\ \boldsymbol{h}_t^{\mathrm{obs}} \\ I_{t-1} \\ \|\boldsymbol{\delta}_{\mathrm{F},t-1}\|_2/(1\ \mathrm{N}) \end{bmatrix}. \tag{A.3}$$

The block sizes are $6 + 12 + 12 + 4 + 2$. Histories run from oldest to newest; entries with negative indices and the two intervention signals are zero at reset. The current state measurements are $\tilde{\boldsymbol{r}}_t = \boldsymbol{r}_t + \boldsymbol{\eta}_{r,t}$ and $\tilde{\boldsymbol{v}}_t = \boldsymbol{v}_t + \boldsymbol{\eta}_{v,t}$, with independent, zero-mean Gaussian components of standard deviations 0.5 m and 0.03 m s^{-1}. There is no state-observation delay. The remaining features are

$$\begin{aligned} \boldsymbol{h}_t^{\mathrm{obs}} &= \begin{bmatrix} h_s(\boldsymbol{x}_t)/(1\ \mathrm{m\,s^{-1}}) \\ h_{v,x}(\boldsymbol{x}_t)/v_{\max}^2 \\ h_{v,y}(\boldsymbol{x}_t)/v_{\max}^2 \\ h_{v,z}(\boldsymbol{x}_t)/v_{\max}^2 \end{bmatrix}, \\ I_{t-1} &= \mathbf{1}_{\{\|\boldsymbol{\delta}_{\mathrm{F},t-1}\|_2 > 10^{-5}\ \mathrm{N}\}}. \end{aligned} \tag{A.4}$$

These four margins use the true current state, and the realized-force history has no added measurement noise. The recovery margin and hidden actuator parameters are absent from the input. The policy receives

$$o_{t,i} = \mathrm{clip}_{[-10,10]}\, \frac{o_{t,i}^{\mathrm{raw}} - \mu_i}{\sqrt{v_i^{\mathrm{obs}} + 10^{-8}}}, \tag{A.5}$$

where $\mu_i, v_i^{\mathrm{obs}}$ are the running observation mean and variance during training and the selected checkpoint's saved values during evaluation. The policy samples a three-component Gaussian action during PPO collection; evaluation uses its mean. Componentwise clipping to $[-1, 1]$, followed by multiplication by $u_{\max}$, gives $\boldsymbol{u}_{\mathrm{des},t}$. After action t, the force histories and intervention signals are updated, the plant advances, and the margins at $\boldsymbol{x}_{t+1}$ enter the next observation.

Appendix A.2. Axial braking tube

The controller uses the forward-Euler CWH update

$$\begin{aligned} \boldsymbol{r}_{t+1} &= \boldsymbol{r}_t + \Delta t\, \boldsymbol{v}_t, \\ \boldsymbol{v}_{t+1} &= \boldsymbol{v}_t + \Delta t\left[(A_{\mathrm{CWH}}\boldsymbol{x}_t)_{4:6} + \boldsymbol{u}_{\mathrm{real},t}/m\right]. \end{aligned} \tag{A.6}$$

Write $p = r_{x,t}$, $v = v_{x,t}$, and $f_- = u_{\mathrm{real},x,t-1}$. The declared domain is $|p| \le p_{\max} = 210$ m, $\|\boldsymbol{v}_t\|_\infty \le 10$ m s^{-1}, and $\|\boldsymbol{u}_{\mathrm{real},t-1}\|_\infty \le u_{\max}$. Within this domain, the magnitude of the axial CWH drift is bounded by

$$d_b = 3n^2 p_{\max} + 2n v_{\max}. \tag{A.7}$$

For braking with $u_{\mathrm{act},x} = +u_{\max}$, set

$$\begin{aligned} a_{\min} &= \frac{\Delta t}{\Delta t + \tau_{\max}} = 0.2, \qquad \rho = 1 - a_{\min} = 0.8, \\ f_\infty &= \sigma_{\min} u_{\max} = 0.7 \text{ N}, \qquad a_b = f_\infty/m - d_b > 0. \end{aligned} \tag{A.8}$$

Starting from $\underline{p}_0 = p$, $\underline{v}_0 = v$, and $\underline{f}_0 = \min(f_-, f_\infty)$, compute $H = 512$ steps:

$$\begin{aligned} \underline{f}_{k+1} &= \rho \underline{f}_k + (1-\rho) f_\infty, \\ \underline{p}_{k+1} &= \underline{p}_k + \Delta t\, \underline{v}_k, \\ \underline{v}_{k+1} &= \underline{v}_k + \Delta t(\underline{f}_{k+1}/m - d_b). \end{aligned} \tag{A.9}$$

The new force enters the velocity update; the position update uses the old velocity. The tube has a fixed horizon, with no early stopping. Its minimum buffer margin is $M = \min_{0\le k\le H} \underline{p}_k - x_{\mathrm{buf}}$. The tube passes when $M \ge 0$, $\underline{v}_H \ge 0$, and $\underline{f}_H/m \ge d_b$. These terminal conditions make the continued lower-bound motion nondecreasing under the same drift bound.

For the command calculation below, the closed form uses

$$C_k = \sum_{\ell=1}^{k} \rho^\ell, \qquad B_k = \sum_{\ell=0}^{k-1} C_\ell, \qquad C_0 = B_0 = 0. \tag{A.10}$$

Then $\underline{f}_k = f_\infty + (\underline{f}_0 - f_\infty)\rho^k$ and

$$\begin{aligned} \underline{v}_k &= v + \Delta t \left[k a_b + \frac{\underline{f}_0 - f_\infty}{m} C_k \right], \\ \underline{p}_k &= p + k\Delta t\, v + \Delta t^2 \left[\frac{a_b k(k-1)}{2} + \frac{\underline{f}_0 - f_\infty}{m} B_k \right]. \end{aligned} \tag{A.11}$$

Appendix A.3. From a braking requirement to the eleventh QP row

Let F be the realized axial force during the candidate step. Its next position is $p_1 = p + \Delta t\, v$, independent of F under Euler integration, and a lower bound on its next velocity is $v + \Delta t(F/m - d_b)$. Starting the subsequent braking tube with force $F \le f_\infty$ gives

$$\begin{aligned} L_k &= p_1 + k\Delta t\, v + \Delta t^2 \left[-k d_b + \frac{a_b k(k-1)}{2} - \frac{f_\infty}{m} B_k \right], \\ L_v &= v - \Delta t\, d_b + \Delta t \left[H a_b - \frac{f_\infty}{m} C_H \right]. \end{aligned} \tag{A.12}$$

The ensuing positions are bounded below by $L_k + \Delta t^2 (k + B_k) F/m$, and the terminal velocity by $L_v + \Delta t(1 + C_H)F/m$. Solving these inequalities yields the physical force floor

$$F_{\mathrm{req}} = \max\left\{ \max_{1\le k\le H} \frac{m(x_{\mathrm{buf}} - L_k)}{\Delta t^2 (k + B_k)}, \frac{-m L_v}{\Delta t(1 + C_H)} \right\}. \tag{A.13}$$

Certification also requires the current tube and domain checks above, $p_1 \ge x_{\mathrm{buf}}$, and $F_{\mathrm{req}} \le f_\infty$. If an actual force exceeds f_∞, the tube conservatively starts at f_∞.

To obtain a command bound without using the hidden parameter draws, define the actuator-domain vertices

$$\mathcal{V} = \left\{ \frac{\Delta t}{\Delta t + \tau_{\min}}, \frac{\Delta t}{\Delta t + \tau_{\max}} \right\} \times \{\sigma_{\min}, \sigma_{\max}\}. \tag{A.14}$$

The worst vertex determines the required dead-zone output:

$$\begin{aligned} w_{\mathrm{req}} &= \max_{(a,\sigma)\in\mathcal{V}} \frac{F_{\mathrm{req}} - (1-a) f_-}{a\sigma}, \\ b_{\mathrm{cmd}} &= \begin{cases} \max(w_{\mathrm{req}}, d_{\max}), & w_{\mathrm{req}} > 0, \\ w_{\mathrm{req}}, & w_{\mathrm{req}} \le 0, \end{cases} + 10^{-9} \text{ N}. \end{aligned} \tag{A.15}$$

For negative commands, ignoring dead-zone suppression gives a conservative lower force. For positive commands, the threshold $d_{\max}$ ensures that the command passes every dead zone in the declared range. The appended row is

$$\boldsymbol{e}_x^\top \boldsymbol{u} \ge b_{\mathrm{cmd}}, \qquad \boldsymbol{e}_x = [1, 0, 0]^\top. \tag{A.16}$$

The supervisor accepts the primary QP candidate only if its current geometry is certified, its axial command satisfies this row, and it lies in $\mathcal{U}$. Domain, position, force-floor, and command comparisons allow 10^{-8} in their respective SI units; the terminal velocity and terminal force-drift checks in the braking tube use zero tolerance. Thus the numerical checks as well as the force floor are part of the controller definition.

Appendix A.4. Constraint sets for the two interface costs

Let $\boldsymbol{\ell}(\boldsymbol{x}) = A_{\mathrm{CWH}}\boldsymbol{x}$. The four shared safety rows are

$$\begin{aligned} g_1^\top &= \nabla\phi_s^\top B_{\mathrm{CWH}}, \qquad b_1 = -\nabla\phi_s^\top \boldsymbol{\ell} - 0.05 h_s - 0.1 h_s^3, \\ g_{i+1}^\top &= \nabla h_{v,i}^\top B_{\mathrm{CWH}}, \quad b_{i+1} = -\nabla h_{v,i}^\top \boldsymbol{\ell} - 0.0005 h_{v,i} - 0.001 h_{v,i}^3. \end{aligned} \tag{A.17}$$

Here $i = 1, 2, 3$ denotes the x, y, z axes, respectively. The implemented range regularization evaluates the position block of $\nabla\phi_s$ as $2\nu_1^2 \boldsymbol{r} + 2\nu_1\nu_0 \boldsymbol{r}/(R + 10^{-12}\text{ m})$; its velocity block is $-2\boldsymbol{v}$. The command and realized-force cost sets each contain five rows:

$$\begin{aligned} \mathcal{R}_t^{\mathrm{F}} &= \{(g_j, b_j)\}_{j=1}^4 \cup \{(\boldsymbol{e}_x, b_{\mathrm{cmd}})\}, \\ \mathcal{R}_t^{\mathrm{A}} &= \{(g_j, b_j)\}_{j=1}^4 \cup \{(\boldsymbol{e}_x, F_{\mathrm{req}})\}. \end{aligned} \tag{A.18}$$

Their slacks are those in Eq. (13). Only finite rows with $\|g_j\|_2 > 10^{-8}$ under the stated SI scaling contribute to either maximum; a set with no valid rows has cost zero. The six thrust-box rows belong to the QP and its DQP active-set geometry, but are excluded from both interface costs. In solver order, the QP consists of $(+\boldsymbol{e}_i, -u_{\max})$, $(-\boldsymbol{e}_i, -u_{\max})$ for $i = x, y, z$, followed by the four shared rows and the command front-plane row.

Appendix A.5. Braking/hold backup and release

For C and E, rejection of the primary candidate activates the backup and stores $p_{\mathrm{hold}} = \max(p, 3\text{ m})$. With $\boldsymbol{r}_{\mathrm{hold}} = [p_{\mathrm{hold}}, 0, 0]^\top$, it computes

$$\boldsymbol{u}_{\mathrm{hold}} = m\left[0.01(\boldsymbol{r}_{\mathrm{hold}} - \boldsymbol{r}) - 0.20\boldsymbol{v} - (A_{\mathrm{CWH}}\boldsymbol{x})_{4:6}\right], \tag{A.19}$$

where the gains have units s^{-2} and s^{-1}. If $v_x < 0$, the axial component is replaced by $+u_{\max}$. After box saturation, the axial command is raised to at least b_{cmd}, subject to $u_{\max}$. Uncertifiable geometry receives $+u_{\max}$ on the axial channel and is recorded as uncertified. The primary QP is still attempted during backup. Release requires an admissible primary candidate, $M \geq 0.10$ m, at least eight completed backup steps, and $v_x \geq -0.02\ \text{m s}^{-1}$. The hold point remains fixed until the next takeover. Configuration D uses the LQR hold law of Section 4.4 and the same recovery floor, with its activation and release parameters specified there.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The core implementation, configuration files, and tests are maintained in a private GitHub repository (`https://github.com/Strtus/pfcr-rpod`). Access to the repository and the simulation data supporting this study can be requested from the corresponding author. Model checkpoints, episode-level data, and training logs are not included in the repository.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used OpenAI Codex to assist with English translation, manuscript organization, language editing, LaTeX preparation, and reference-format verification. The authors reviewed, verified, and edited all resulting content and take full responsibility for the content of the manuscript.

References

[1] M. Tipaldi, R. Iervolino, P. R. Massenio, Reinforcement learning in spacecraft control applications: Advances, prospects, and challenges, Annual Reviews in Control 54 (2022) 1–23. doi:10.1016/j.arcontrol.2022.07.004.

[2] K. Hovell, S. Ulrich, Deep reinforcement learning for spacecraft proximity operations guidance, Journal of Spacecraft and Rockets 58 (2) (2021) 254–264. doi:10.2514/1.A34838.

[3] C. E. Oestreich, R. Linares, R. Gondhalekar, Autonomous six-degree-of-freedom spacecraft docking with rotating targets via reinforcement learning, Journal of Aerospace Information Systems 18 (7) (2021) 417–428. doi:10.2514/1.I010914.

[4] Q. Qu, K. Liu, W. Wang, J. Lü, Spacecraft proximity maneuvering and rendezvous with collision avoidance based on reinforcement learning, IEEE Transactions on Aerospace and Electronic Systems 58 (6) (2022) 5823–5834. doi:10.1109/TAES.2022.3180271.

[5] A. D. Ames, X. Xu, J. W. Grizzle, P. Tabuada, Control barrier function based quadratic programs for safety critical systems, IEEE Transactions on Automatic Control 62 (8) (2017) 3861–3876. doi:10.1109/TAC.2016.2638961.

[6] J. Breeden, D. Panagou, Guaranteed safe spacecraft docking with control barrier functions, IEEE Control Systems Letters 6 (2022) 2000–2005. doi:10.1109/LCSYS.2021.3136813.

[7] K. Dunlap, M. Hibbard, M. Mote, K. Hobbs, Comparing run time assurance approaches for safe spacecraft docking, IEEE Control Systems Letters 6 (2022) 1849–1854. doi:10.1109/LCSYS.2021.3135260.

[8] K. Dunlap, M. Mote, K. Delsing, K. L. Hobbs, Run time assured reinforcement learning for safe satellite docking, Journal of Aerospace Information Systems 20 (1) (2023) 25–36. doi:10.2514/1.I011126.

[9] A. T. Harris, H. Schaub, Spacecraft command and control with safety guarantees using shielded deep reinforcement learning, in: AIAA Scitech 2020 Forum, American Institute of Aeronautics and Astronautics, 2020, pp. AIAA 2020–0386. doi:10.2514/6.2020-0386.

[10] D. van Wijk, K. Dunlap, M. Majji, K. Hobbs, Safe spacecraft inspection via deep reinforcement learning and discrete control barrier functions, Journal of Aerospace Information Systems 21 (12) (2024) 996–1013. doi:10.2514/1.I011391.

[11] H. Krasowski, J. Thumm, M. Müller, L. Schäfer, X. Wang, M. Althoff, Provably safe reinforcement learning: Conceptual analysis, survey, and benchmarking, Transactions on Machine Learning Research (2023).
URL https://openreview.net/forum?id=mcN0ezbnz0

[12] N. Hamilton, K. Dunlap, T. T. Johnson, K. L. Hobbs, Ablation study of how run time assurance impacts the training and performance of reinforcement learning agents, in: 2023 IEEE 9th International Conference on Space Mission Challenges for Information Technology (SMC-IT), IEEE, 2023, pp. 45–55. doi:10.1109/SMC-IT56444.2023.00014.

[13] H. Markgraf, S. Sawant, H. Krasowski, L. Schäfer, S. Gros, M. Althoff, Safe reinforcement learning using action projection: Safeguard the policy or the environment?, Transactions on Machine Learning Research (Mar. 2026).
URL https://openreview.net/forum?id=DDrGSEYxGU

[14] X. Wang, Ensuring safety of learning-based motion planners using control barrier functions, IEEE Robotics and Automation Letters 7 (2) (2022) 4773–4780. doi:10.1109/LRA.2022.3152313.

[15] J. Thumm, G. Pelat, M. Althoff, Reducing safety interventions in provably safe reinforcement learning, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2023, pp. 7515–7522. doi:10.1109/IROS55552.2023.10342464.

[16] F. Pizarro Bejarano, L. Brunke, A. P. Schoellig, Safety filtering while training: Improving the performance and sample efficiency of reinforcement learning agents, IEEE Robotics and Automation Letters 10 (1) (2025) 788–795. doi:10.1109/LRA.2024.3512374.

[17] A. Stooke, J. Achiam, P. Abbeel, Responsive safety in reinforcement learning by PID lagrangian methods, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 9133–9143.
URL https://proceedings.mlr.press/v119/stooke20a.html

[18] J. Achiam, D. Held, A. Tamar, P. Abbeel, Constrained policy optimization, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 22–31.
URL https://proceedings.mlr.press/v70/achiam17a.html

[19] D. Goldfarb, A. Idnani, A numerically stable dual method for solving strictly convex quadratic programs, Mathematical Programming 27 (1) (1983) 1–33. doi:10.1007/BF02591962.

[20] B. Amos, J. Z. Kolter, OptNet: Differentiable optimization as a layer in neural networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 136–145.
URL https://proceedings.mlr.press/v70/amos17a.html

[21] S. Barratt, On the differentiability of the solution to convex optimization problems, arXiv (2018). arXiv:1804.05098.

[22] W. H. Clohessy, R. S. Wiltshire, Terminal guidance system for satellite rendezvous, Journal of the Aerospace Sciences 27 (9) (1960) 653–658. doi:10.2514/8.8704.

[23] D. Dirkx, M. Fayolle, G. Garrett, M. Avillez, K. Cowan, S. Cowan, J. Encarnacao, C. Fortuny Lombrana, J. Gaffarel, J. Hener, X. Hu, M. van Nistelrooij, F. Oggionni, M. Plumaris, The open-source astrodynamics Tudatpy software—overview for planetary mission design and science analysis, Europlanet Science Congress 2022, EPSC Abstracts, Vol. 16, EPSC2022-253 (2022). doi:10.5194/epsc2022-253.

[24] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, U. Topcu, Safe reinforcement learning via shielding, Proceedings of the AAAI Conference on Artificial Intelligence 32 (1) (2018). doi:10.1609/aaai.v32i1.11797.

[25] R. Cheng, G. Orosz, R. M. Murray, J. W. Burdick, End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks, Proceedings of the AAAI Conference on Artificial Intelligence 33 (01) (2019) 3387–3395. doi:10.1609/aaai.v33i01.33013387.

[26] K. P. Wabersich, M. N. Zeilinger, A predictive safety filter for learning-based control of constrained nonlinear dynamical systems, Automatica 129 (2021) 109597. doi:10.1016/j.automatica.2021.109597.

[27] Y. Zhang, S. Walters, X. Xu, Control barrier function meets interval analysis: Safety-critical control with measurement and actuation uncertainties, in: 2022 American Control Conference (ACC), IEEE, 2022, pp. 3814–3819. doi:10.23919/ACC53348.2022.9867681.

[28] S. Kolathaya, A. D. Ames, Input-to-state safety with control barrier functions, IEEE Control Systems Letters 3 (1) (2019) 108–113. doi:10.1109/LCSYS.2018.2853698.

[29] J. M. Picone, A. E. Hedin, D. P. Drob, A. C. Aikin, NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues, Journal of Geophysical Research: Space Physics 107 (A12) (2002) SIA 15–1–SIA 15–16. doi:10.1029/2002JA009430.

[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv (2017). arXiv:1707.06347.

Supplementary material

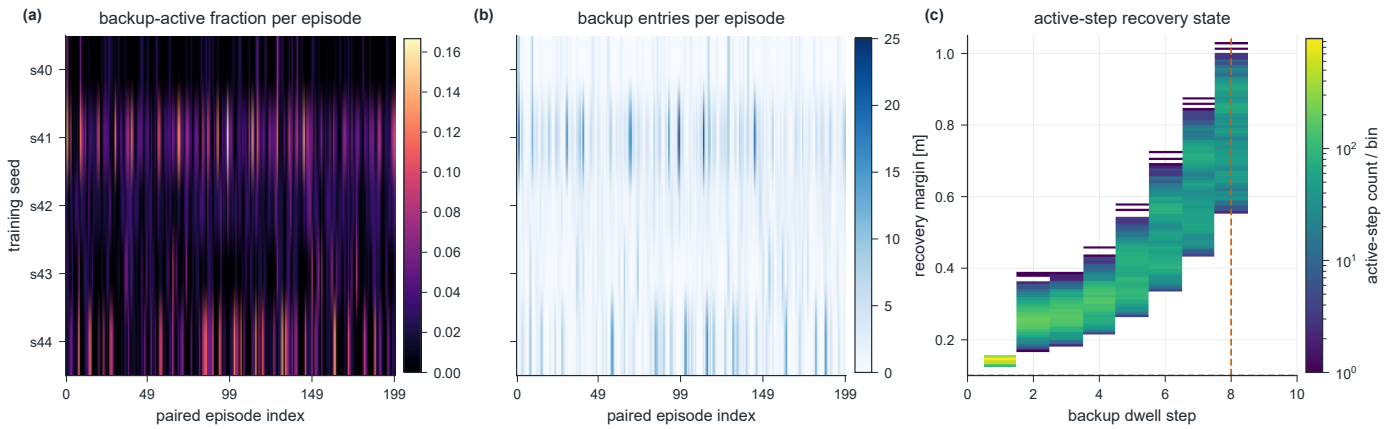


Figure S1: Complete telemetry for the robust saturated-LQR backup controller. The panels retain every takeover and release event and report the associated recoverability margins. These data describe the deterministic supervisory fallback and are not used to attribute an effect to the DQP credit channel.

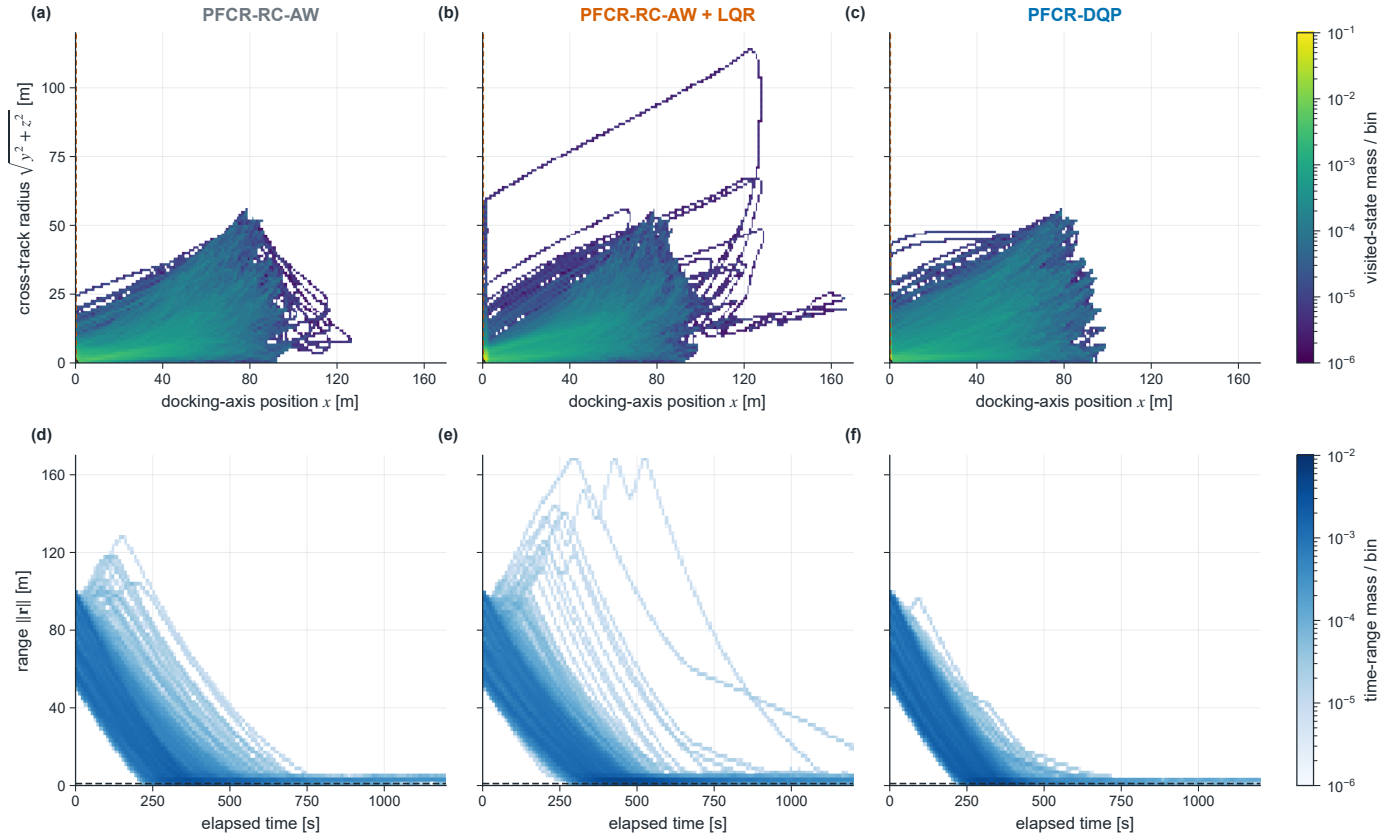


Figure S2: Trajectory geometry and time–distance occupancy for the three frozen configurations. All valid held-out transitions are retained; the panels provide a state-occupancy complement to the scalar aggregate metrics in the main manuscript.

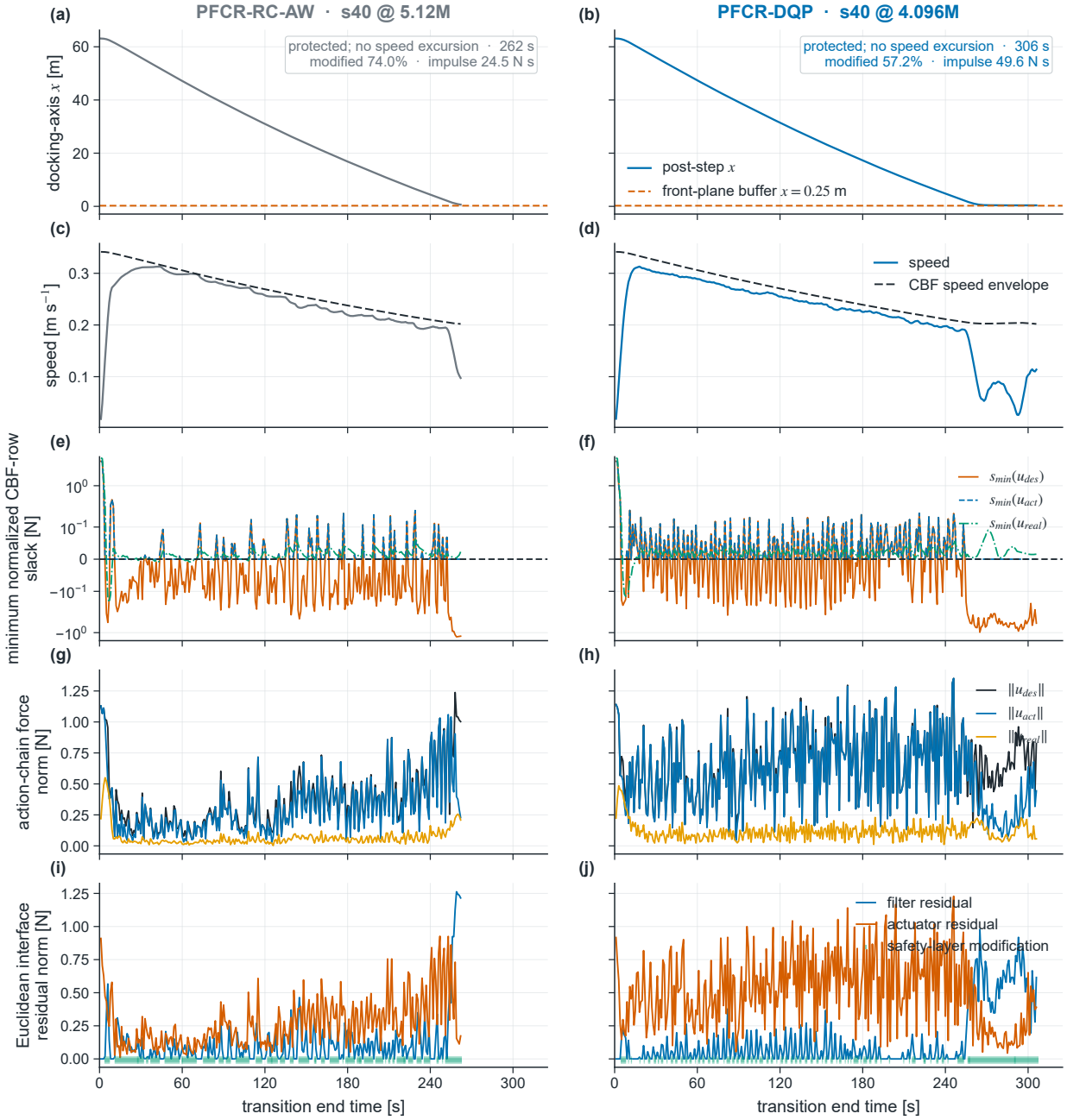


Figure S3: Unsmoothed closed-loop traces for the same paired C/E episode shown in the main manuscript (training seed 40, episode index 0, episode seed 350000). Every original transition is shown and no post-termination padding is added. The figure shows the sample-aligned temporal relationship among nominal, filtered, and realized thrust magnitudes, CBF slack, and runtime-safety-layer modification, defined by $\|\delta_F\|_2 > 10^{-5}$ N.

S1. Evaluation and checkpoint selection

The candidate training horizons are fixed in advance at 1.024, 2.048, 3.072, 4.096, 5.12, 10.24, 15.36, and 20.48 million transitions. A checkpoint is eligible only if it occurs at least 1.5 million transitions after the most recent activation of the dual variables. Across the five training seeds, the first eligible steps lie between 2.309 and 2.522 million transitions. Consequently, the actual comparison contains six candidate horizons: 3.072, 4.096, 5.12, 10.24, 15.36, and 20.48 million transitions.

The development set uses episode seeds 340000–340049 and evaluates each checkpoint for 50 deterministic episodes. Three methods, five training seeds, and six eligible candidate horizons yield 90 development cells and 4,500 episodes. One checkpoint is selected independently for each training seed using the following fixed lexicographic ranking:

1. maximize the protected docking success rate;
2. minimize the velocity-envelope violation rate;

3. minimize the mean total impulse; and
4. prefer the earlier training horizon.

An eligible candidate must additionally record zero front-plane crossings, zero front-plane-buffer violations, zero uncertified supervisor samples, zero capture-sphere collision proxies, zero triangle-relation anomalies, and zero one-step input-bound anomalies on the development set. The selection manifest is frozen before the independent holdout is read.

The independent holdout uses episode seeds 350000–350199. The model selected for each training seed is evaluated for 200 deterministic episodes, producing 15 evaluation cells and 3,000 episodes. For matching training seeds and episode indices, the initial state, observation-noise sequence, and actuator parameters are paired across methods; the identity hashes match across all three holdouts. The holdout is not used for reselection, hyperparameter tuning, or additional training.

Table S1: Frozen checkpoint for each training seed, in millions of transitions.

Method	Seed 40	Seed 41	Seed 42	Seed 43	Seed 44
PFCR-RC-AW	5.120	3.072	4.096	4.096	5.120
PFCR-RC-AW + LQR	5.120	4.096	4.096	3.072	5.120
PFCR-DQP	4.096	5.120	4.096	3.072	4.096

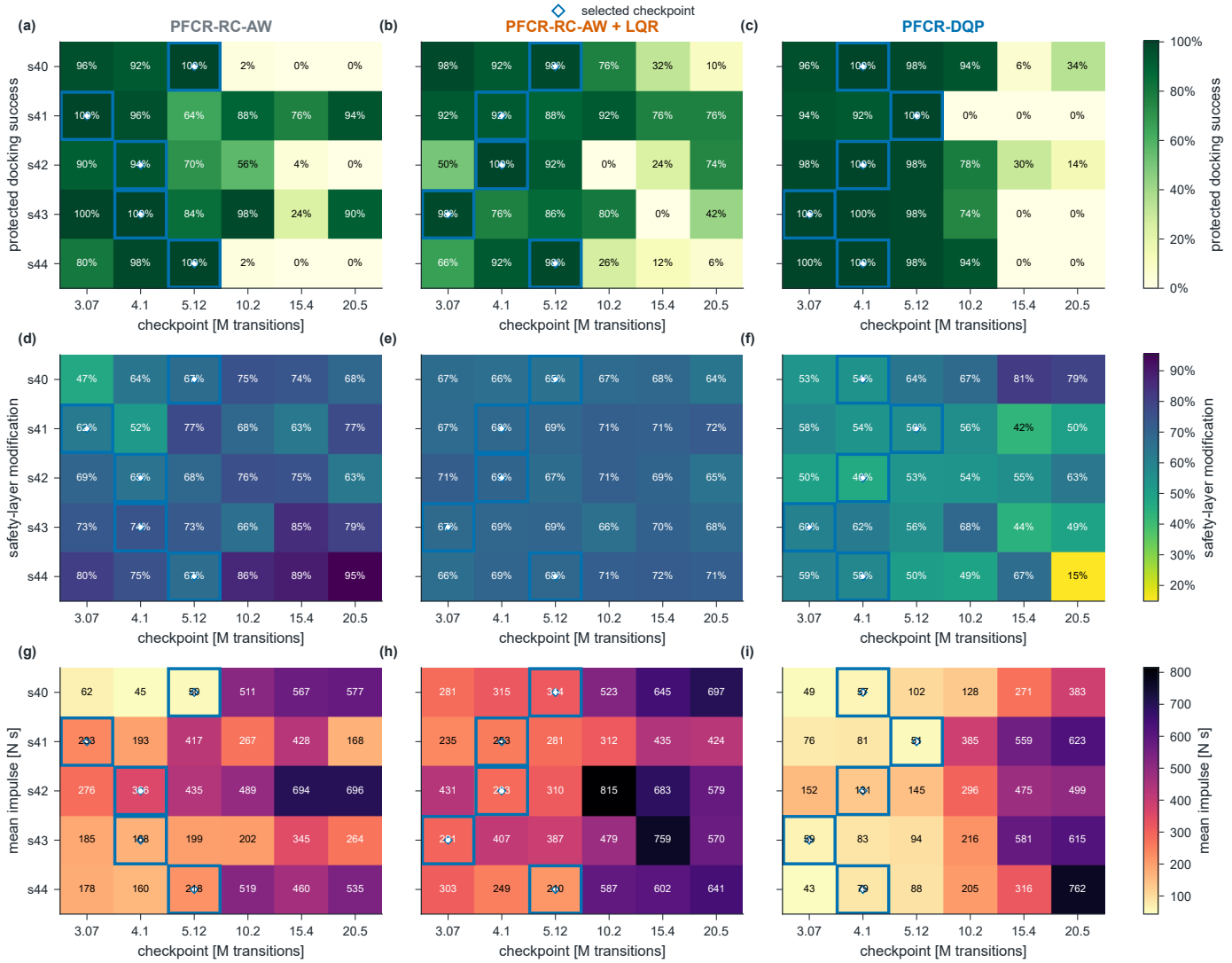


Figure S4: Protected docking success rate, runtime-safety-layer modification rate, and total impulse in the 90 frozen development-set cells. Outlined cells indicate the checkpoints written to the selection manifest before any holdout run. Modification uses $\|\boldsymbol{\delta}_{\mathrm{F}}\|_2 > 10^{-5}$ N.

The Tudat evaluation uses episode seeds 380000–380099 at each training seed and truth level. Initial relative states, actuator parameters, observation-noise sequences, inertial initial states, orbital angles, epochs, and T1 thrust-error sequences are paired across methods. Selected policy and normalizer files are verified by SHA-256 before copying to a separate, write-protected evaluation root. Coordinate round trips, thrust-axis signs, short-arc T0/CWH agreement, integration-step convergence, and the first control step are checked before evaluation. All 30 cells complete without artifact mismatches, control-interface failures, pairing violations, or integration failures. The T0/T1 results use the CWH-trained checkpoints; navigation-loop and later truth-fine-tuning studies are separate experiments. The CWH and Tudat evaluation episodes are not paired with one another.

S2. Projection and gradient diagnostics

All 102.4 million PFCR-DQP training transitions had both a valid operational QP solution and valid projection diagnostics; the invalid-sample count was zero. Under the stated numerical tolerances, 66.78% of the transitions were accepted as fixed-active-set locally regular. The remaining 33.22% were conservatively classified as boundary or near-degenerate samples and were masked. Among the locally regular samples, 69.45% contained at least one active constraint. The weighted mean active rank was 0.814, and the weighted mean Jacobian trace was 2.186, consistent with approximately one normal degree of freedom being constrained in the three-dimensional control space. The maximum KKT stationarity residual over all training records was 6.66×10^{-16}.

The first five policy updates for each of the five training seeds produced 25 gradient diagnostics. The recorded ratio of DQP to PPO policy-gradient norms ranged from 0.00406 to 0.04095. Every value was finite, and the clipping-saturation rate of the policy mean action was zero. These observations show that the DQP auxiliary loss supplied nonzero credit to the policy network rather than merely producing numerical logs; by themselves, they are not evidence of task performance. Figure S5 summarizes the projection and initial-gradient diagnostics.

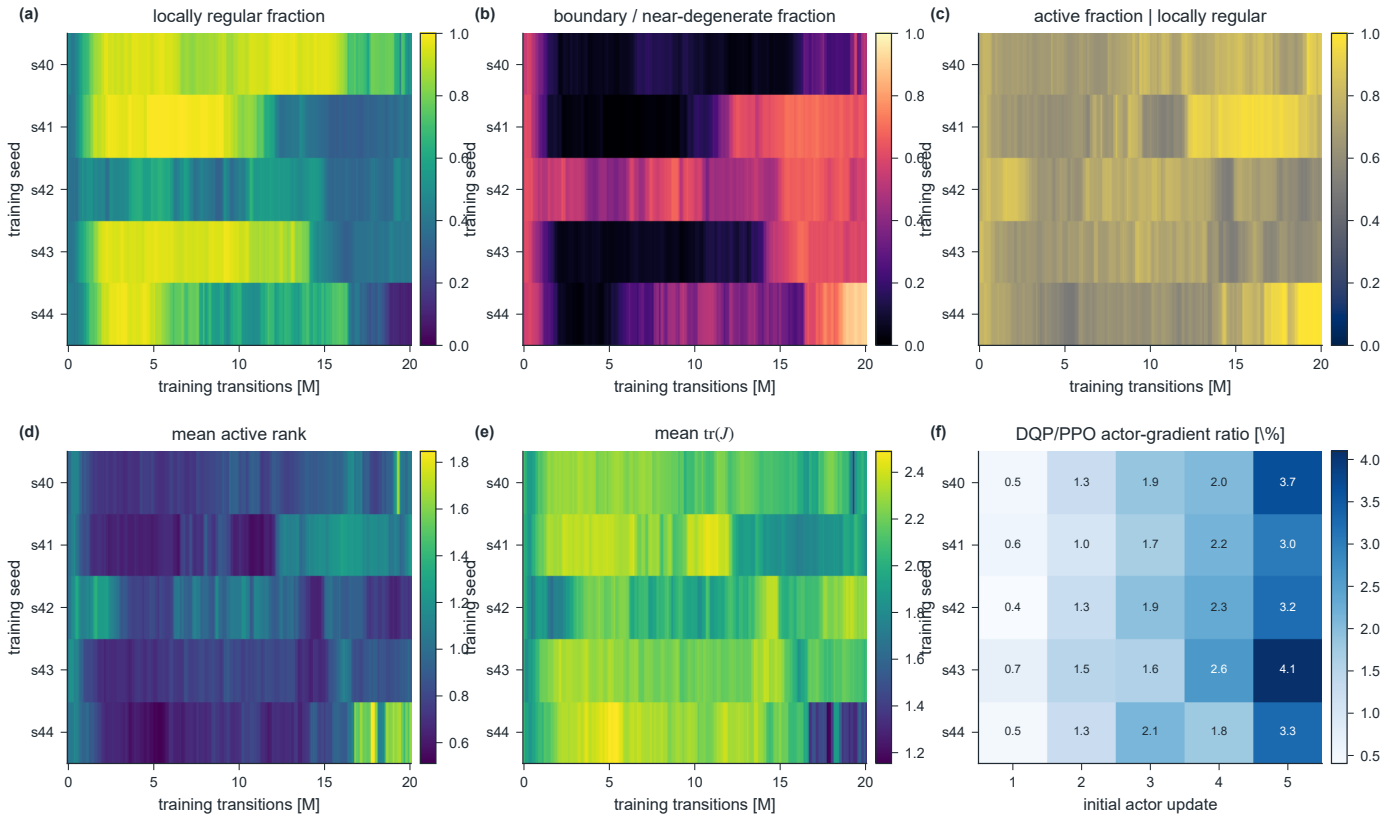


Figure S5: Projection validity, numerical local-regularity acceptance, active rank, Jacobian trace, and initial-gradient diagnostics for PFCR-DQP. The statistics cover all 102.4 million PFCR-DQP training transitions; boundary or near-degenerate samples are retained in the counts but excluded from the DQP loss.