Recoverability-shielded reinforcement learning for far-range spacecraft rendezvous

Abstract

Repeated replanning from uncertain navigation estimates can produce unnecessary corrective manoeuvres. This paper develops recurrent impulsive guidance for the approach from approximately 10 km to a 100 m handover in geostationary orbit. Every five minutes, a policy uses the navigation solution, reported uncertainty and manoeuvre history to select a burn gate and a velocity increment. A recoverability shield checks the thrust commands against a predicted braking manoeuvre with a four-standard-deviation navigation margin. Dataset-aggregation imitation of an adaptive certainty-equivalent planner initializes the policy, and imitation-regularized proximal policy optimization improves its fuel–time objective under randomized navigation errors and thruster faults. Native Tudat evaluation includes Earth J_2, lunisolar gravity and solar radiation pressure, with eighteen classical designs flying matched missions. In nominal and combined-fault missions at navigation errors of 0.5–5 % of range, the policy achieves the lowest mean fuel–time cost among the evaluated methods: 4–26 % below its adaptive teacher and 17–68 % below feedback guidance on paired successful missions. Its mean propellant use over successful missions is lower than that of every classical design at each error level. At 5 % error, nominal completion reaches 91.7 %, compared with 76.5 % for the teacher. The constraint audit records zero violations in 22 800 test missions and five in 582 010 training episodes.

Keywords: Spacecraft rendezvous, far-range approach, safe reinforcement learning, run-time assurance, impulsive guidance, imitation learning

1. Introduction

Autonomous on-orbit servicing requires a spacecraft to approach a client from kilometres away while respecting restrictions on approach direction and speed [1]. In the geostationary-orbit (GEO) task considered here, the far-range phase connects an initial separation of approximately 10 km to the 100 m handover to close-range operations. The servicer must remain behind the target, enter its vicinity through an approach corridor and reduce its closing speed as the range decreases. Guidance must accomplish this transfer with limited propellant while responding to navigation errors and changes in delivered thrust.

Navigation uncertainty makes the timing of a correction as consequential as its magnitude. A noncooperative client can be tracked through line-of-sight measurements, with range information supplied by the relative motion. The ARGON and AVANTI flight experiments demonstrated this form of navigation and autonomous approach. ARGON reported a mean along-track navigation error of about 1 % of the initial separation, while AVANTI reported boresight accuracy of about 10 % of the current separation [2–4]. These flight results motivate evaluating guidance with range-dependent navigation errors. A certainty-equivalent planner treats the resulting state estimate as the true state when computing a manoeuvre. Repeated replanning therefore responds both to spacecraft motion and to changes in the estimation error. This paper asks whether a policy trained with navigation uncertainty can learn when to burn and how much velocity increment to apply, improving the fuel–time cost of the entire approach while its commands are checked for recoverability.

Classical designs provide efficient solutions to the underlying targeting problem. The Clohessy–Wiltshire (CW) model [5] supports impulsive targeting and periodic replanning [1, 6]; glideslope guidance [6], artificial potential functions [7] and model predictive control [8] provide feedback and constraint handling. Disturbance observers estimate and compensate for acceleration disturbances [9]. A dead band or longer replanning interval reduces sensitivity to navigation fluctuations but delays trajectory corrections. The design problem is therefore to select burn timing and increments under uncertainty over a multi-hour transfer.

Learning-based guidance has been applied to spacecraft proximity operations [10, 11], docking [12] and adaptive guidance [13]. Recent work brings learning and safety assurance closer to operational rendezvous tasks. Van Wijk et al. [14] use discrete control barrier functions within optimization-based run-time assurance for learned spacecraft inspection. For far-range rendezvous, Wijayatunga et al. [15] develop nominal trajectory planning and PPO guidance with safety and angles-only observability considerations. Their policy selects a contraction parameter for a convex optimization problem that computes corrective impulses. Stephenson and Schaub [16] formulate shielded inspection using impulsive velocity changes and selectable drift durations; their action-projection shield preserves passive collision avoidance under eccentric relative-motion dynamics.

Constrained policy optimization [17] and Lagrangian methods [18] incorporate constraint costs in learning, while shielding [19] checks the actions issued by a policy. Backup-based run-time assurance [20–22] and predictive safety filters [23] motivate checking the continuation of a candidate command in addition to its immediate effect. Dynamic model predictive shielding [24] uses a local planner and learned value estimates to select

recovery actions that account for subsequent task reward. Gameplay filters [25] evaluate candidate actions through simulated interactions between a learned safety policy and an adversarial disturbance policy. For spacecraft proximity operations, Wijayatunga et al. [26] learn parameters and a residual barrier within an input-constrained control-barrier-function formulation to improve fuel use and recoverability.

Here, a navigation-conditioned policy allocates corrections over a multi-hour GEO approach. The policy directly selects a burn/coast gate and velocity increment at five-minute manoeuvre opportunities. A fixed braking rule and five thrust blends supply candidate commands, checked against the rear plane, approach corridor and speed zone with navigation-dependent margins. The learned quantities are thus mission-level corrections, coupled to execution-level braking checks.

The contributions are:

1. A recurrent impulsive guidance formulation that directly selects correction timing and magnitude from uncertain navigation and manoeuvre history. Dataset aggregation (DAgger) [27] supplies targeting actions from an adaptive certainty-equivalent teacher; imitation-regularized PPO [28, 29] optimizes fuel–time cost over the complete approach.
2. A command-level recoverability shield that checks both the next thrust interval and a subsequent braking segment against rendezvous geometry and speed constraints. Matched ablations quantify the effect of this continuation check relative to a one-step filter, a Lagrangian penalty and unfiltered execution.
3. A matched native-Tudat evaluation against eighteen classical designs and the imitated policy, spanning exact and uncertain navigation, actuator faults and starts near constraints. It quantifies fuel–time cost and mission completion across the tested regimes, with constraint violations and shield exhaustion reported separately.

2. Problem formulation

The task is to guide a chaser from a kilometre-scale separation to a close-range handover while balancing propellant and flight time. Guidance uses an uncertain relative-state estimate and commands thrusters whose efficiency, alignment and bias can change during the approach. The formulation below defines the dynamics available to the controller, the uncertainty encountered by each mission, and the geometric and terminal conditions used to evaluate the approach.

2.1. Relative dynamics

The target flies a near-circular geostationary orbit ($a = 42\,164\,\mathrm{km}$, $e = 10^{-4}$). The chaser state in the target local-vertical local-horizontal (LVLH) frame is $\boldsymbol{x} = [\boldsymbol{r}^\top, \boldsymbol{v}^\top]^\top$, with x radial outward, y along the orbital velocity and z along the orbit normal. For guidance design and training, the relative motion is described by the CW equations with mean motion n,

$$\begin{aligned} \ddot{x} - 2n\dot{y} - 3n^2 x &= a_x + d_x, \\ \ddot{y} + 2n\dot{x} &= a_y + d_y, \\ \ddot{z} + n^2 z &= a_z + d_z, \end{aligned} \tag{1}$$

or $\dot{\boldsymbol{x}} = A\boldsymbol{x} + B(\boldsymbol{a} + \boldsymbol{d})$, where $\boldsymbol{a}$ is the realized thrust acceleration and $\boldsymbol{d}$ collects unmodelled accelerations. Commands are held constant over the control interval $\Delta t = 1\,\mathrm{s}$, and wherever a prediction model is needed the exact zero-order-hold transition of (1) is used,

$$\begin{aligned} \boldsymbol{x}_{j+1} &= \Phi \boldsymbol{x}_j + \Gamma(\boldsymbol{a}_j + \boldsymbol{d}_j), \\ \Phi &= e^{A\Delta t}, \qquad \Gamma = \int_0^{\Delta t} e^{As} B \,\mathrm{d}s, \end{aligned} \tag{2}$$

where j counts control steps. The test evaluation propagates target and chaser in Earth-centred J2000 with native TudatPy [30] integration (fixed-step fourth-order Runge–Kutta, 0.25 s): Earth gravity with J_2, point-mass gravity of Sun and Moon from DE440 ephemerides, and cannonball solar radiation pressure (SRP) with Earth shadow. The target has an area-to-mass ratio of $0.015\,\mathrm{m^2/kg}$ and $C_r = 1.3$; each chaser draws its own ratio in $0.005\,\mathrm{m^2/kg}$ to $0.03\,\mathrm{m^2/kg}$ and $C_r \in [1.2, 1.5]$, so the differential SRP differs between missions. The epoch and the target's orbital phase are drawn per scenario. Guidance and the safety layer use the CW transition throughout this evaluation, exposing the closed loop to the dynamics mismatch between the training and test models.

2.2. Thruster fault model

The commanded LVLH acceleration $\boldsymbol{u}$, with $\|\boldsymbol{u}\| \le u_{\max} = 0.02\,\mathrm{m/s^2}$, is realized as

$$\boldsymbol{a} = \mathrm{diag}(\boldsymbol{\eta})\, R\, \boldsymbol{u} + \boldsymbol{b}, \qquad t \ge t_f, \tag{3}$$

and $\boldsymbol{a} = \boldsymbol{u}$ before the unknown onset time $t_f \in [0, 3\,\mathrm{h}]$. The per-axis efficiencies $\eta_i \in [0.4, 1]$, the misalignment rotation R (random axis, angle up to 5°) and the constant bias $\|\boldsymbol{b}\| \le 2 \times 10^{-4}\,\mathrm{m/s^2}$ are hidden from every controller; thrusters are assumed held in the LVLH axes, so a per-axis efficiency loss both shortens and deflects a burn. The out-of-distribution (OOD) test scenario applies all three faults with extended ranges $\eta_i \ge 0.25$, angles up to 8° and $\|\boldsymbol{b}\| \le 4 \times 10^{-4}\,\mathrm{m/s^2}$.

2.3. Relative navigation

Every controller and the safety layer act on a navigation solution $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{e}$, while the constraints and the handover are evaluated on the true state. The position error per LVLH axis has the standard deviation

$$\sigma_p(\rho) = 0.5\,\mathrm{m} + f\rho, \tag{4}$$

where ρ is the range and f the navigation-error level, the error as a fraction of the range. On each axis i a unit second-order Gauss–Markov process ξ_i with the natural frequency $\omega_n = 2\pi/2\,\mathrm{h}$ and the damping $\zeta = 0.7$ sets the position and velocity errors,

$$\begin{aligned} \ddot{\xi}_i + 2\zeta\omega_n\dot{\xi}_i + \omega_n^2\xi_i &= \sqrt{4\zeta\omega_n^3}\,\nu_i(t), \\ e_{r,i} = \sigma_p(\rho)\,\xi_i, \qquad e_{v,i} &= \sigma_p(\rho)\,\dot{\xi}_i, \end{aligned} \tag{5}$$

driven by unit white noise ν_i. Position and velocity errors therefore share a temporally correlated process. The process is initialized in its stationary distribution, with unit position variance and velocity-error standard deviation $\sigma_v = \omega_n\sigma_p$ per axis.

The measured acceleration innovation carries white noise of $2\times10^{-5}\,\mathrm{m/s^2}$, and every controller receives σ_p and σ_v evaluated at the estimated range $\hat{\rho}$. The evaluated levels $f = 0.5$–$5\,\%$ cover errors of the order encountered by angles-only navigation in flight: about 1 % of the separation along track in ARGON [2] and about 10 % in the boresight direction in AVANTI [3]. Every method receives the same unit error process for a given mission and level. Exact-navigation runs set the state error and reported uncertainty to zero.

2.4. Mission, constraints and performance

Missions start 9.5 km to 10.5 km from the target, within 60° of the $-y$ axis, with relative speed below 0.5 m/s. The approach hands over to a close-range controller at 100 m. The trajectory is required to satisfy

$$y \le 0 \qquad \text{(rear half-space)}, \tag{6}$$

$$\|\boldsymbol{r}\| \le R_c \;\Rightarrow\; \sqrt{x^2+z^2} \le -y\tan\theta \tag{7}$$

$$\|\boldsymbol{r}\| \le R_s \;\Rightarrow\; \|\boldsymbol{v}\| \le v_{\lim}, \tag{8}$$

$$\|\boldsymbol{r}\| \ge R_{ko} \qquad \text{(keep-out)}. \tag{9}$$

with $R_c = 200\,\mathrm{m}$, $\theta = 30°$, $R_s = 100\,\mathrm{m}$, $v_{\lim} = 1\,\mathrm{m/s}$ and $R_{ko} = 10\,\mathrm{m}$. The corridor constraint (7) makes the safe set non-convex: outside the 200 m sphere the chaser may be anywhere behind the target, inside it must lie in the line-of-sight (LOS) cone. A mission succeeds when the chaser enters the handover sphere $\|\boldsymbol{r}\| \le 100\,\mathrm{m}$ at a speed within the speed-zone limit $v_{\lim}$ within 12 h and without an audited violation of (6)–(9). The audit evaluates the true state at 0.25 s resolution and checks every connecting straight segment against the corridor and keep-out sphere (Section 3.3). Fuel is the commanded velocity expenditure $\Delta v = \int \|\boldsymbol{u}\|\,dt$, and the flight time T is the time to the handover. Both enter the performance index

$$J = \Delta v + w\,T, \qquad w = \tfrac{1}{3}\,\mathrm{m/s\ per\ hour}, \tag{10}$$

which values an hour of flight at 0.33 m/s of propellant. Success, violations, propellant and flight time are evaluated on the true trajectory. With $\mathcal{S}$ the set of states that satisfy (6)–(9), the guidance problem is

$$\begin{aligned} &\min_{\pi}\ \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\vartheta},\xi}[\Delta v + w\,T] \\ &\text{s.t. } \boldsymbol{x}(t)\in\mathcal{S}, \quad \|\boldsymbol{u}(t)\| \le u_{\max}, \quad t\in[0,T], \\ &\qquad \|\boldsymbol{r}(T)\| \le R_s, \quad \|\boldsymbol{v}(T)\| \le v_{\lim}, \quad T \le 12\,\mathrm{h}, \end{aligned} \tag{11}$$

where the expectation runs over the initial state, the hidden fault parameters $\boldsymbol{\vartheta} = (\boldsymbol{\eta}, R, \boldsymbol{b}, t_f)$ and the navigation-error process (5). The policy π selects burns from its observation history; the hidden faults and state-estimation error make this a partially observable problem. Section 3 combines a recurrent policy for burn timing and velocity increments with a shield that checks the resulting commands against the geometric constraints and an assumed braking continuation.

3. Method

Figure 1 follows the online control loop from recurrent guidance through burn execution and recoverability shielding to the rendezvous environment. Every $T_d = 300\,\mathrm{s}$, the policy uses the observation $\boldsymbol{o}_k$ to select a burn gate g_k and a three-component latent increment $\boldsymbol{\ell}_k$. The bounded mapping converts $\boldsymbol{\ell}_k$ into the requested velocity increment $\Delta\boldsymbol{v}_k$, and the burn/coast executor produces the proposed acceleration $\boldsymbol{u}_{\mathrm{prop},j}$ at 1 s intervals. The shield blends this proposal with a braking backup and checks the candidates before selecting the command $\boldsymbol{u}_j$ sent to the actuator.

The actuator model maps the selected command to the realized acceleration $\boldsymbol{a}_j$, which drives the relative dynamics. Navigation supplies the estimated state $\hat{\boldsymbol{x}}_j$, reported position uncertainty σ_p and acceleration innovation $\boldsymbol{w}_j$ to the shield and the policy observation construction. The lower feedback paths carry the executed command, selected blend and previous gated increment into the manoeuvre history used at the next guidance decision. The policy therefore selects burn timing and magnitude while the shield checks individual acceleration commands.

3.1. Impulsive action

Burn selection and burn magnitude form a parameterized action space [31]. At decision k the policy outputs the latent $(g_k, \boldsymbol{\ell}_k) \in \mathbb{R}\times\mathbb{R}^3$, with velocity increment

$$\Delta\boldsymbol{v}_k = \Delta v_{\max}\,\frac{\boldsymbol{\ell}_k}{\sqrt{1+\|\boldsymbol{\ell}_k\|^2}}, \qquad \Delta v_{\max} = 2\,\mathrm{m/s}, \tag{12}$$

and the burn/coast executor fires if the gate is open, $g_k > 0$, and $\|\Delta\boldsymbol{v}_k\| \ge \delta = 0.02\,\mathrm{m/s}$, as indicated by the two conditions in Fig. 1; otherwise it proposes zero thrust. At the control steps $j = 0, 1, \ldots$ after the decision the executor proposes

$$\boldsymbol{u}_{\mathrm{prop},j} = u_{\max}\,\min\!\Big(1, \frac{t_b - j\Delta t}{\Delta t}\Big)^{+}\frac{\Delta\boldsymbol{v}_k}{\|\Delta\boldsymbol{v}_k\|}, \tag{13}$$

where $(\cdot)^+ = \max(\cdot, 0)$ and $t_b = \|\Delta\boldsymbol{v}_k\|/u_{\max}$. With a nominal actuator, this command delivers $\Delta\boldsymbol{v}_k$ and then coasts; a new decision replaces any unfinished burn. The increment is independent of spacecraft mass. Each burn is open loop, so an efficiency fault changes the delivered increment and the policy receives the measured outcome at its next decision. A closed gate explicitly requests a coast. The policy samples a Gaussian latent and uses the following masked log-likelihood in its policy-gradient update:

$$\log\pi(g,\boldsymbol{\ell}) = \log\mathcal{N}(g) + \mathbb{I}_{g>0}\,\log\mathcal{N}(\boldsymbol{\ell}). \tag{14}$$

The increment contribution is active only when the sampled gate is open.

3.2. Observation and network

The observation $\boldsymbol{o}_k \in \mathbb{R}^{34}$ describes the approach geometry, recent control response and reported navigation uncertainty. Geometry is represented by the estimated relative state with positions scaled by 10 km, the range and its logarithm, the speed, and the LOS-cone and corridor-sphere margins scaled

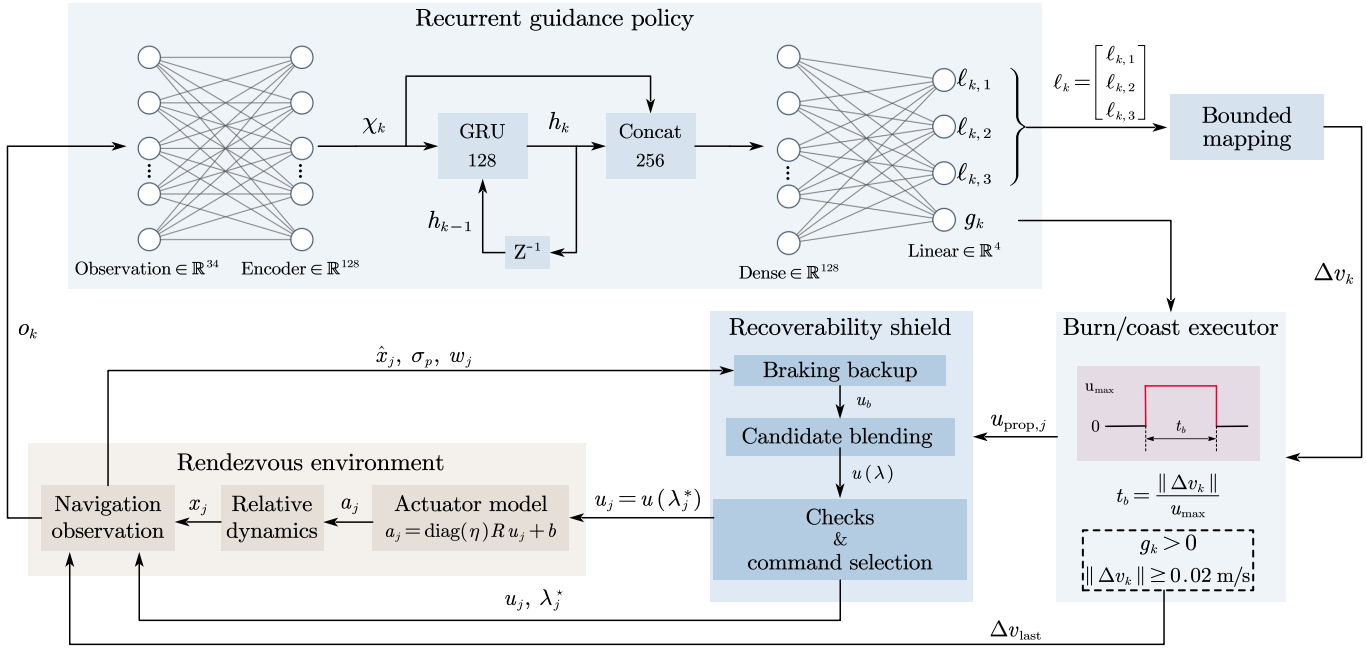


Figure 1: Closed-loop architecture of recurrent impulsive guidance with burn/coast execution and recoverability shielding.

by 200 m. The LOS direction $\hat{\boldsymbol{r}}/\|\hat{\boldsymbol{r}}\|$ and clipped near-field position $\mathrm{clip}(\hat{\boldsymbol{r}}/300\,\mathrm{m}, -1, 1)$ resolve the corridor entry at short range. The manoeuvre history supplies the interval means of the executed command, acceleration innovation and shield blend, together with the previous decision's gated increment $\Delta\boldsymbol{v}_{\mathrm{last}}$ and mission clock. This stored increment is zero after a coast decision. These features correspond to the lower feedback paths in Fig. 1. Two uncertainty inputs, $\mathrm{clip}(20\,\sigma_p/\hat{\rho}, 0, 2)$ and $\log_{10}(1 + \sigma_p/\mathrm{m})/3$, describe its relative and absolute scales.

The control-response features use the acceleration innovation

$$\boldsymbol{w}_j = \Gamma^{+}(\boldsymbol{x}_{j+1} - \Phi\boldsymbol{x}_j) - \boldsymbol{u}_j, \tag{15}$$

which is the realized minus commanded acceleration of control step j, measured with the accelerometer noise of Section 2.3. Over the coasting steps C_k and the full-thrust steps $\mathcal{B}_k$ ($\|\boldsymbol{u}_j\| > u_{\max}/2$) of the last decision interval, the two fault measurements are

$$\bar{\boldsymbol{w}}_k = \frac{1}{|C_k|}\sum_{j\in C_k}\boldsymbol{w}_j, \qquad \epsilon_k = \frac{1}{|\mathcal{B}_k|}\sum_{j\in\mathcal{B}_k}\frac{\langle \boldsymbol{w}_j, \boldsymbol{u}_j\rangle}{\|\boldsymbol{u}_j\|^2}. \tag{16}$$

The coast innovation $\bar{\boldsymbol{w}}_k$, scaled by $2\times10^{-4}\,\mathrm{m/s^2}$, measures the additive disturbance, while ϵ_k measures the thrust error along a burn and is retained through intervals without a burn. A 128-unit layer encodes $\boldsymbol{o}_k$ into $\boldsymbol{\chi}_k$, and a 128-unit gated recurrent unit (GRU) [32] updates the memory $\boldsymbol{h}_k$. In Fig. 1, Z^{-1} returns $\boldsymbol{h}_{k-1}$ to the GRU, while the bypass forwards $\boldsymbol{\chi}_k$ directly to the concatenation block. The resulting 256-dimensional vector $[\boldsymbol{\chi}_k; \boldsymbol{h}_k]$ passes through a 128-unit dense layer and a four-output linear layer to obtain the mean of the gate and three increment latents:

$$\begin{aligned}
&\boldsymbol{\chi}_k = \tanh(W_o\boldsymbol{o}_k + \boldsymbol{c}_o), \qquad \boldsymbol{h}_k = \mathrm{GRU}(\boldsymbol{\chi}_k, \boldsymbol{h}_{k-1}),\\
&\boldsymbol{\mu}_k = W_2\tanh(W_1[\boldsymbol{\chi}_k; \boldsymbol{h}_k] + \boldsymbol{c}_1) + \boldsymbol{c}_2,\\
&(g_k, \boldsymbol{\ell}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathrm{diag}(\boldsymbol{\sigma}^2)),
\end{aligned} \tag{17}$$

with $\boldsymbol{h}_0 = \boldsymbol{0}$ at the start of a mission and a learned state-independent standard deviation $\boldsymbol{\sigma}$. Training samples the latent, and the online path in Fig. 1 uses its mean $\boldsymbol{\mu}_k$ for evaluation. The critic receives the additional training input $\boldsymbol{p}_k$: hidden fault parameters, navigation error $\hat{\boldsymbol{x}} - \boldsymbol{x}$ and elapsed mission time, following asymmetric actor–critic training [33]. Deployment uses the actor inputs above.

3.3. Recoverability shield

At every control step, the braking-backup block in Fig. 1 computes $\boldsymbol{u}_{\mathrm{b}}$ from the navigation state and the filtered acceleration innovation. Candidate blending combines it with the executor proposal (13); the checks and command-selection block then selects the smallest admissible blend,

$$\begin{aligned}
&\boldsymbol{u}(\lambda) = (1-\lambda)\,\boldsymbol{u}_{\mathrm{prop},j} + \lambda\,\boldsymbol{u}_{\mathrm{b}},\\
&\lambda^{\star} = \min\{\lambda\in\Lambda:\ \boldsymbol{u}(\lambda)\ \text{passes tests (i) and (ii)}\},
\end{aligned} \tag{18}$$

with $\Lambda = \{0, 0.25, 0.5, 0.75, 1\}$. The selected weight at control step j is denoted $\lambda_j^{\star}$ in Fig. 1, and the command is $\boldsymbol{u}_j = \boldsymbol{u}(\lambda_j^{\star})$. The reported uncertainty σ_p enters the candidate checks through the prediction and braking margins below. If no candidate passes, the shield reports exhaustion and executes the backup for the current interval; the mission then terminates under the protocol of Section 4.2. The backup brakes against the relative velocity, cancels the CW drift and the shield's own disturbance estimate $\tilde{\boldsymbol{d}}$, a low-pass filter of the innovation (15) with $\tau = 30\,\mathrm{s}$, and respects the thrust limit,

$$\boldsymbol{u}_{\mathrm{b}} = \mathrm{sat}_{u_{\max}}\Big[\,\mathrm{sat}_{a_b}\Big(\frac{\boldsymbol{v}_c - \hat{\boldsymbol{v}}}{\Delta t}\Big) - \boldsymbol{f}(\hat{\boldsymbol{x}}) - \tilde{\boldsymbol{d}}\Big], \tag{19}$$

where sat_a limits the norm of a vector to a, $a_b = 0.015\,\mathrm{m/s^2}$, $\boldsymbol{f}(\boldsymbol{x})$ is the acceleration part of $A\boldsymbol{x}$ in (1), and the commanded

velocity $\boldsymbol{v}_c$ is zero, or $[0, -0.01, 0]^\top$ m/s, a slow retreat along $-y$, once the chaser has stopped within 5 m of a boundary.

Test (i) predicts the next control interval from the navigation solution with (2), the nominal actuator $\boldsymbol{a}_j = \boldsymbol{u}(\lambda)$ and $\boldsymbol{d}_j = \boldsymbol{0}$, at $\tau = 0.25, 0.5, 0.75$ and 1 s. The predicted states and the straight segments between them must keep the margins of (6), (7) and (9) at least

$$\rho_{\mathrm{i}}(\tau) = \tfrac{1}{2}\varepsilon\tau^2 + 4\sigma_p(\hat{\rho}) + \rho_{\mathrm{res}}\,\mathbb{I}_{\tau=\Delta t}, \tag{20}$$

a prediction margin that grows with $\varepsilon = 2 \times 10^{-4}\,\mathrm{m/s^2}$, and within $R_s + \rho_{\mathrm{i}}$ of the target the predicted speed must stay below $v_{\lim} - 0.02\,\mathrm{m/s} - \varepsilon\tau$. The term $4\sigma_p(\hat{\rho})$ expands the geometric margin with the reported position uncertainty at the estimated range. The endpoint reserve is $\rho_{\mathrm{res}} = 0.5$ m for $\lambda < 1$ and zero for the full-backup candidate.

Test (ii) brakes straight from the predicted endpoint $(\boldsymbol{r}_1, \boldsymbol{v}_1)$ at the assumed deceleration $b = 0.006\,\mathrm{m/s^2}$. The chaser stops after t_s at $\boldsymbol{r}_s$, and it slows to $\bar{v} = v_{\lim} - 0.02$ m/s after t_v at $\boldsymbol{r}_v$,

$$\begin{aligned} t_s &= \frac{\|\boldsymbol{v}_1\|}{b}, \qquad \boldsymbol{r}_s = \boldsymbol{r}_1 + \tfrac{1}{2}t_s\boldsymbol{v}_1, \\ t_v &= \frac{(\|\boldsymbol{v}_1\| - \bar{v})^+}{b}, \qquad \boldsymbol{r}_v = \boldsymbol{r}_1 + \frac{\|\boldsymbol{v}_1\| + \bar{v}}{2\|\boldsymbol{v}_1\|}\,t_v\boldsymbol{v}_1. \end{aligned} \tag{21}$$

The segment $[\boldsymbol{r}_1, \boldsymbol{r}_s]$ must keep the margins of (6), (7) and (9) at least $\rho_{\mathrm{ii}}(t_s)$, and unless $\|\boldsymbol{v}_1\| \le \bar{v}$ the segment $[\boldsymbol{r}_1, \boldsymbol{r}_v]$ must stay at least $\rho_{\mathrm{ii}}(t_v)$ outside the speed zone, with the margin

$$\rho_{\mathrm{ii}}(t) = \rho_{\mathrm{res}} + \rho_{\mathrm{pos}} + 4\sigma_p(\|\boldsymbol{r}_1\|) + \tfrac{1}{2}\varepsilon t^2, \tag{22}$$

where $\rho_{\mathrm{pos}} = 0.5$ m. Thus a candidate is admitted when both its predicted hold interval and its straight braking continuation pass the expanded geometric and speed checks. The constants b, ε, ρ_{pos} and ρ_{res} define the check, and the true-state trajectory audit measures the closed-loop constraint outcomes. The one-step-filter baseline applies test (i) only.

The nonconvex corridor set is handled exactly on straight segments. Its point margin is $\max(\|\boldsymbol{r}\| - R_c,\, m_\theta(\boldsymbol{r}))$ with the LOS margin

$$m_\theta(\boldsymbol{r}) = -y\sin\theta - \sqrt{x^2 + z^2}\cos\theta, \tag{23}$$

the signed distance to the cone surface, positive inside the cone. On a segment $\boldsymbol{r}(s) = \boldsymbol{r}_0 + s\,\Delta\boldsymbol{r}$, $s \in [0, 1]$, the part inside a sphere of radius R is $[s_-, s_+]$, with

$$s_\pm = \frac{-\boldsymbol{r}_0^\top\Delta\boldsymbol{r} \pm \sqrt{(\boldsymbol{r}_0^\top\Delta\boldsymbol{r})^2 - \|\Delta\boldsymbol{r}\|^2(\|\boldsymbol{r}_0\|^2 - R^2)}}{\|\Delta\boldsymbol{r}\|^2} \tag{24}$$

clipped to $[0, 1]$. Because m_θ is concave along any line, its minimum over the chord is $\min\{m_\theta(\boldsymbol{r}(s_-)), m_\theta(\boldsymbol{r}(s_+))\}$, an exact test of the segment. The shield applies it with $R = R_c + \rho$ for its margin ρ, and with $R = R_c$ the same test audits the realized trajectory. The resulting algorithm uses the same geometric tests for command admission and trajectory auditing, with uncertainty and braking margins applied during admission. Section 5.3 reports the audited outcomes in Tudat.

3.4. Imitation of a certainty-equivalent planner

The teacher is a certainty-equivalent (CE) impulsive planner using the same navigation solution, action and executor as the student. At each transfer-leg decision it searches arrival times $\tau \in [0.75\,\mathrm{h}, 8\,\mathrm{h}]$ on a grid, targeting the corridor entry point $\boldsymbol{r}_E = [0, -300, 0]^\top$ m. For each candidate time it solves the forced CW targeting problem with constant estimated disturbance $\hat{\boldsymbol{d}}$. Writing the exact transition as $\boldsymbol{x}(\tau) = \Phi(\tau)\boldsymbol{x} + G(\tau)\hat{\boldsymbol{d}}$ and partitioning it into position and velocity blocks gives the required departure and arrival velocities, evaluated at the navigation solution $(\boldsymbol{r}, \boldsymbol{v})$:

$$\begin{aligned} \boldsymbol{v}^\star(\tau) &= \Phi_{rv}^{-1}(\boldsymbol{r}_E - \Phi_{rr}\boldsymbol{r} - G_r\hat{\boldsymbol{d}}), \\ \boldsymbol{v}_E(\tau) &= \Phi_{vr}\boldsymbol{r} + \Phi_{vv}\boldsymbol{v}^\star(\tau) + G_v\hat{\boldsymbol{d}}. \end{aligned} \tag{25}$$

It then minimizes

$$\begin{aligned} &\|\boldsymbol{v}^\star(\tau) - \boldsymbol{v}\| + \|\boldsymbol{v}_E(\tau) - \boldsymbol{v}_D(\tau)\| \\ &\qquad + w\left(\tau + \frac{200\,\mathrm{m}}{s_D(\tau)}\right) \end{aligned} \tag{26}$$

over arcs that remain at least 30 m behind the rear plane. The entry drift velocity is $\boldsymbol{v}_D = [0, s_D, 0]^\top$, where s_D is the along-track arrival velocity $v_{E,y}$ clipped to 0.1 m/s to 0.8 m/s; w is the time weight in (10). Replanning replaces the current plan when its cost decreases by at least 0.01 m/s. At the last decision before arrival, or whenever the measured state enters $\|\boldsymbol{r}\| \le \|\boldsymbol{r}_E\|$, the planner switches to a drift toward $[0, -50, 0]^\top$ m and re-aims that velocity at each decision. The state-based entry rule also assigns a drift command to student trajectories that reach the entry sphere before the teacher's planned arrival. During coasts, the planner estimates the constant disturbance by low-pass filtering the innovation (15):

$$\hat{\boldsymbol{d}} \leftarrow \hat{\boldsymbol{d}} + \frac{\Delta t}{\tau_d}(\boldsymbol{w}_j - \hat{\boldsymbol{d}}), \qquad j \in C, \quad \tau_d = 60\,\mathrm{s}. \tag{27}$$

The teacher combines three extensions of this planner, Imp-CE. A corridor re-entry rule returns the chaser to the approach axis when it nears the corridor sphere off axis. A per-axis efficiency estimate then compensates the commanded increments (Imp-CE-adapt). Using the full-thrust steps $\mathcal{B}$ accumulated during the mission, the estimate for axis i is

$$\hat{\eta}_i = 1 + \frac{\sum_{j\in\mathcal{B}}(w_{j,i} - \hat{d}_i)u_{j,i}}{\sum_{j\in\mathcal{B}} u_{j,i}^2}, \tag{28}$$

clipped to $[0.15, 1.2]$ and used once $\sum_{j\in\mathcal{B}} u_{j,i}^2 > (0.02\,u_{\max})^2$, and the planned increment is divided by $\hat{\eta}_i$ on each axis. On the transfer leg, finally, a re-targeting increment is fired only if

$$\|\Delta\boldsymbol{v}\| \ge \max\left(\delta,\ k\sqrt{3}\,\sigma_v\right), \tag{29}$$

k times the root-mean-square magnitude of the velocity error (Imp-CE-adapt-$k\sigma$). This dead band suppresses corrections at the scale of the navigation uncertainty. The teacher uses $k = 3$ and is denoted Imp-CE-adapt-3σ; with exact navigation it equals Imp-CE-adapt. Section 4.1 evaluates each extension as a separate baseline.

Dataset aggregation (DAgger) proceeds in nine rounds $k = 0, \dots, 8$ of 512 missions [27] from the training fault mixture, with navigation errors whose level f is drawn per mission uniformly from $[0, 4\,\%]$. Each mission of round k is flown by the policy π_k, and the teacher $\pi^\star$ labels every decision using the navigation solution visited by that policy. The aggregated dataset grows as

$$\begin{aligned} \mathcal{D}_k &= \mathcal{D}_{k-1} \cup \{(\boldsymbol{o}_{0:K}, \Delta\boldsymbol{v}^\star_{0:K})\}_{\pi_k}, \\ \pi_k &= \begin{cases} \pi^\star & \text{with probability } \beta_k = 2^{-k}, \\ \pi_{\theta_{k-1}} & \text{otherwise,} \end{cases} \end{aligned} \tag{30}$$

where $\boldsymbol{o}_{0:K}$ and $\Delta\boldsymbol{v}^\star_{0:K}$ are the observations and teacher increments of a mission and $\pi_{\theta_{k-1}}$ is the student fitted after round $k-1$. The student is fitted on the aggregated sequences by unrolling the GRU over whole missions with the loss

$$\begin{aligned} L_{\mathrm{BC}} = \frac{1}{\sum_k \omega_k} \sum_k \omega_k \Bigg[& \mathrm{BCE}(g_k, \mathbb{I}_{\|\Delta\boldsymbol{v}^\star_k\| \ge \delta}) \\ & + \mathbb{I}_{\|\Delta\boldsymbol{v}^\star_k\| \ge \delta} \frac{\|\Delta\boldsymbol{v}_k - \Delta\boldsymbol{v}^\star_k\|^2}{(0.05\ \mathrm{m/s})^2} \Bigg], \end{aligned} \tag{31}$$

where $\Delta\boldsymbol{v}^\star_k$ is the teacher's increment and decisions within 1 km of the target, which decide the corridor entry, carry the weight $\omega_k = 3$ (otherwise 1). The first round fits 100 epochs and each later round 30 epochs from the previous network; a final pass of 60 epochs with a cosine learning-rate decay from 3×10^{-4} to 10^{-5} fits the full dataset of about 240 000 decisions. The imitated policy predicts the teacher's burn decision in 89.8 % of the validation decisions, with a root-mean-square increment error of 0.13 m/s.

3.5. Reinforcement learning

The imitated policy is improved with recurrent PPO [28] in the exact CW simulator with the fault model (3) randomized per mission (80 % of missions faulty, each with a random nonempty subset of fault types and a random onset) and with navigation errors whose level f is drawn per mission uniformly from $[0, 4\,\%]$. The reward of a decision interval is

$$\begin{aligned} r = {} & 10\,\frac{\Delta\rho}{10\,\mathrm{km}} - 3\,\Delta v - \frac{\Delta t_d}{1\,\mathrm{h}} \\ & - 0.02 \sum \frac{\|\boldsymbol{u} - \boldsymbol{u}_{\mathrm{prop}}\|}{u_{\max}} + 10\,\mathbb{I}_{\mathrm{success}} \\ & - 20\,\mathbb{I}_{\mathrm{failure}} - 10\,\mathbb{I}_{\mathrm{timeout}}, \end{aligned} \tag{32}$$

where $\Delta\rho$ is the progress toward the 100 m sphere, Δv the propellant of the interval in m/s, Δt_d its duration and the sum runs over the shield's interventions. Fuel and time enter with the ratio of the performance index (10). With the probability ratio $q_k = \pi_\theta(a_k \mid \boldsymbol{o}_{0:k}) / \pi_{\theta_{\mathrm{old}}}(a_k \mid \boldsymbol{o}_{0:k})$ of the masked latent likelihood (14) and the critic V_ϕ, the policy and critic losses of an update are

$$\begin{aligned} L_{\mathrm{PPO}} &= -\mathbb{E}_k[\,\min\left(q_k \hat{A}_k,\ \mathrm{clip}(q_k, 1-c, 1+c)\,\hat{A}_k\right)], \\ \hat{A}_k &= \sum_{l \ge 0} (\gamma\lambda_{\mathrm{GAE}})^l\, \delta_{k+l}, \qquad \delta_k = r_k + \gamma V_{k+1} - V_k, \\ L_V &= \tfrac{1}{2}\,\mathbb{E}_k[(V_\phi(\boldsymbol{o}_k, \boldsymbol{p}_k) - \hat{A}_k - V_k)^2], \end{aligned} \tag{33}$$

with the advantages $\hat{A}_k$ of generalized advantage estimation [34], normalized over each update, the critic values V_k of the rollout, zero after the end of a mission, and the privileged critic input $\boldsymbol{p}_k$ described in Section 3.2. Following the demonstration-augmented policy-gradient strategy of retaining a decaying auxiliary objective [29], the actor minimizes the clipped PPO loss plus the imitation loss (31) on minibatches of the aggregated dataset:

$$L_{\mathrm{actor}} = L_{\mathrm{PPO}} + \lambda_j L_{\mathrm{BC}}, \qquad \lambda_j = 0.1 \cdot 0.98^{\,j}, \tag{34}$$

where j counts the actor updates. The decaying imitation term retains the teacher's guidance while PPO optimizes the mission return. The critic trains alone for the first 20 updates. The target KL divergence is 0.02; further actor updates stop when the sampled approximate KL exceeds 0.03, while critic updates continue for all epochs. The initial latent standard deviations are 0.3 for the gate and 0.02 for the increment, controlling exploration of burn timing and magnitude separately. Each update collects 16 decisions from each of 512 parallel missions and runs four epochs of eight minibatches of 16-decision sequences, with discount $\gamma = 0.999$ per decision, $\lambda_{\mathrm{GAE}} = 0.95$, clip ratio $c = 0.2$ and learning rates 10^{-4} and 10^{-3} for actor and critic.

Each run comprises three stages of 304, 304 and 305 updates (2.2×10^9 control intervals in total). Stage two starts from the final network of stage one, and stage three from the checkpoint selected in stage two. Both restarts use a new optimizer and five critic-only warm-up updates, with the imitation weight continuing its decay. Stage three broadens the mission mixture to include 20 % exact-navigation missions and 15 % starts near a constraint. The latter use the stress-test distributions (Section 4.2) with different seeds. Stages one and two use the continuous $f \in [0, 4\,\%]$ distribution together with the 0.5 m uncertainty floor. Checkpoints are saved every ten updates and ranked on 192 development missions from the stage's training mixture, first by success rate minus twice the violation rate and then by J. These development seeds are disjoint from the test seeds. Four exploration/learning-rate settings, the continuation to stage two, and three stage-three mixtures were compared on separate development missions spanning the navigation-error levels and stress conditions of Section 4.2.

4. Experimental setup

4.1. Methods

The comparisons assess fuel–time cost, improvement beyond imitation and constraint outcomes. All methods fly the same initial states, fault realizations and navigation-error processes. Unless a safety ablation specifies otherwise, they also use the same recoverability shield and navigation margin. Table A.1 collects the numerical settings.

The classical baselines use the common handover condition of 100 m range and speed up to 1 m/s. They span fixed-time transfers, online transfer-time selection, drift approaches and continuous feedback:

- *Fixed-time impulsive targeting* (Imp-T3, Imp-T4, Imp-T5): the executor of Section 3.1 driven by CW re-targeting every 300 s toward an arrival at the entry point fixed at the mission start (3 h, 4 h and 5 h, rear-plane checked), followed by the drift phase.

- *Certainty-equivalent impulsive planners* (Section 3.4): Imp-CE; with the corridor re-entry rule (Imp-CE-axis); with, in addition, the per-axis efficiency estimate (Imp-CE-adapt); and with, in addition, the dead band of two or three standard deviations of the velocity error (Imp-CE-adapt-2σ and Imp-CE-adapt-3σ, the teacher). On the development missions the dead band of three standard deviations gave the lowest performance index of all certainty-equivalent variants at every navigation-error level from 1 to 3 %, nominally and under combined faults; dead bands of one to two standard deviations, fixed dead bands of 0.05 to 0.5 m/s and re-planning intervals of 15 to 60 min for Imp-CE-adapt and Drift-CE were less economical.

- *Drift designs* (Drift-T2 to Drift-T6, Drift-CE): CW targeting of the entry point with closed-loop burns and a coast-only disturbance observer, a burn into a drift toward the aim point at 0.1 m/s to 0.8 m/s with lateral centering and a return to the axis when the chaser approaches the corridor sphere off the axis; the transfer time is fixed (2 h to 6 h) or chosen by the certainty-equivalent rule.

- *Glideslope guidance* [6]: two-impulse transfer to 1 km on the $-y$ axis, then six inbound glideslope pulses decelerating from 0.6 m/s to 0.2 m/s.

- *Artificial potential function* (APF): a guidance design based on potential-field rendezvous [7], implemented here as an attractive velocity field toward the aim point with rear-plane repulsion, a lateral waypoint for the corridor entry, a speed limit of 0.8 m/s and a disturbance observer.

- *Waypoint DOB-LQR*: LQR on the exact CW model with a disturbance observer, aimed at the entry point and then at a hold point inside the handover sphere; and a two-phase design (CW targeting to the entry, then DOB-LQR).

The learned comparisons include the proposed method (Imp-RL, three seeds), its imitated policy before RL (Imp-BC), and the same guidance trained with exact navigation (Imp-RL-exact, three seeds). Imp-RL-exact imitates Imp-CE and is improved in 200 PPO updates of 128 parallel missions with an increment noise of 0.004 and an actor learning rate of 3×10^{-5}. The safety mechanisms and the alternatives to imitation are compared within this exact-navigation training: PPO without imitation (Imp-RL-scratch), initialized with the gate closed and an increment noise of 0.1; the training with the shield removed (Imp-RL-none), with the one-step filter (Imp-RL-1step), and with a Lagrangian penalty on violations and no shield (Imp-RL-Lag, three seeds); and end-to-end PPO that outputs the full acceleration every 10 s, initialized by behavior cloning of the waypoint DOB-LQR, with the same four safety mechanisms (E2E-none, E2E-1step, E2E-Lag, E2E-shield). A variant trained on a wider fault envelope, Imp-RL-W, is introduced in Section 5.5. Checkpoints for all learned methods are selected using the criterion in Section 3.5.

4.2. Scenarios and protocol

The main comparison varies the navigation-error level in nominal and combined-fault missions: every method flies with exact navigation and at $f = 0.5$, 1, 2, 3 and 5 %. Four further fault scenarios isolate efficiency loss, misalignment and bias, and combine the extended OOD ranges. These four scenarios are evaluated with exact navigation and at $f = 1$, 2 and 3 % for the learned guidance, its imitated policy, the certainty-equivalent planners Imp-CE-adapt, Imp-CE-adapt-2σ and Imp-CE-adapt-3σ, Drift-CE and the feedback guidance laws glideslope, APF and waypoint DOB-LQR. Five stress tests evaluate starts near a constraint, with exact navigation and at $f = 2$ %: outside the cone at 230 m to 600 m moving toward the corridor sphere (nominal and combined faults), near the axis at 130 m to 250 m closing at 0.6 m/s to 1.2 m/s (nominal and combined), and 20 m to 150 m behind the rear plane drifting forward. A fixed seed defines each scenario's initial states, hidden faults and unit navigation-error processes, and the navigation levels reuse those missions. Imp-RL and Imp-RL-exact use three training seeds; Imp-BC and the single-seed ablations use one. Each scenario contains 200 missions per method and training seed; each stress test contains 100. Section 5.5 extends these missions to further fault magnitudes and types.

Native Tudat propagation with SRP provides the primary evaluation, with a matched replay in the exact CW model. A mission ends at successful handover, the 12 h time limit, an audited constraint violation, or shield exhaustion. Upon exhaustion, the full backup is applied for one 1 s interval, that interval is audited, and the mission is recorded as a failure. Constraint violations and exhaustion events are recorded separately. Success rates carry Wilson 95 % intervals, and fuel and time are averaged over successful missions. Paired comparisons use the missions completed by both methods. The reported 95 % intervals resample the pooled successful mission–seed pairs with replacement; classical runs are matched to each learned-policy seed for the same mission.

Development evaluations use 200 missions per scenario and navigation level with seeds distinct from the test set. They set the four-standard-deviation navigation margin, planner dead bands, exploration settings and training mixtures. The two-stage policy was evaluated before stage three was introduced. The added stage targets two behaviours observed in the development and test missions: increased cost with exact navigation and prolonged holding near the corridor sphere in the 2 % stress tests. Its mixture was selected on development missions, and the final comparison reuses the earlier test missions to quantify the change.

5. Results

Unless identified as CW replay or development data, results use native Tudat test propagation. Costs are means over successful missions; paired differences use missions completed by both methods. Bootstrap intervals resample pooled seed–mission pairs, matching each learned seed to the same reference mission.

5.1. Performance under navigation errors

At navigation errors of 0.5 to 5 %, Imp-RL has the lowest mean performance index and propellant expenditure among the evaluated methods, both nominally and under combined faults (Fig. 2). Table 1 gives paired cost differences, and Supplementary Appendix A contains the complete comparison.

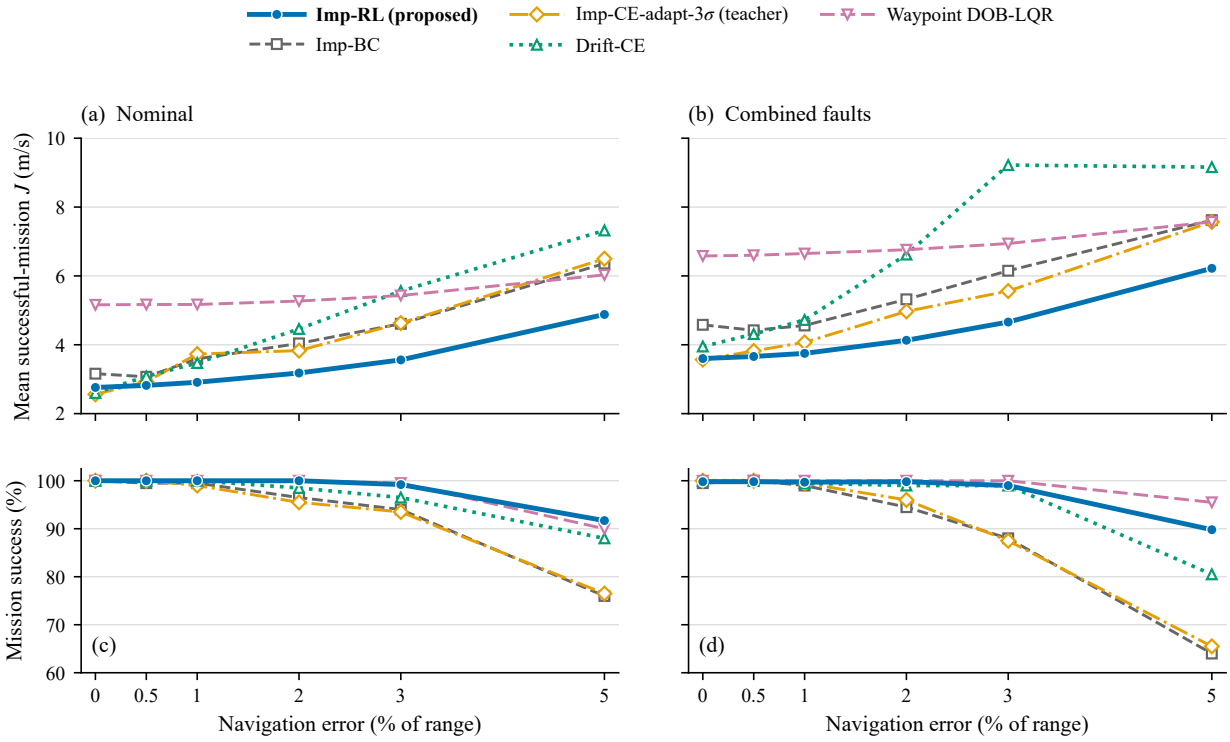


Figure 2: Navigation-error sensitivity: nominal missions (left) and combined faults (right), with successful-mission mean J above and success rate below.

Imp-RL lowers the teacher's cost on jointly successful missions by 3.8–25.6 % nominally and 4.1–17.6 % under combined faults. Both the teacher and the 2σ planner have higher paired costs at every tested error level, with pooled-pair bootstrap 95 % intervals below zero (Table 1). Imp-CE-adapt without a dead band shows a much larger increase in nominal propellant, from 1.41 m/s with exact navigation to 11.04 m/s at 5 %; at 2 %, its 5.30 m/s compares with 2.64 m/s for the teacher. Across nominal and combined-fault missions at 0.5–5 %, its paired cost reductions relative to the evaluated feedback guidance laws range from 17.1 to 68.4 %.

The 1 % nominal comparison includes three unusually costly teacher missions: shield overrides last 1.8–3.2 h, flight times reach 7.8–11.2 h, and propellant expenditure is 20–49 m/s. Their inclusion gives a paired cost reduction of 22.3 % for Imp-RL; excluding these three missions gives 8.2 %.

Table 1: Paired difference of the performance index of the proposed guidance to the strongest classical designs (%, with pooled-pair bootstrap 95 % interval), over the missions both complete, native Tudat. Columns identify the teacher (Imp-CE-adapt-3σ), the 2σ planner, Drift-CE and waypoint DOB-LQR.

Scenario	Error	Teacher	2σ planner	Drift-CE	DOB-LQR
Nominal	exact	+7.9 [+7.3, +8.6]	+7.9 [+7.3, +8.6]	+5.9 [+5.4, +6.5]	−46.5 [−47.0, −46.0]
	0.5 %	−3.8 [−7.1, −1.4]	−5.5 [−6.4, −4.7]	−8.4 [−9.1, −7.9]	−45.5 [−46.0, −44.9]
	1 %	−22.3 [−33.9, −12.5]	−12.3 [−13.2, −11.4]	−16.3 [−17.0, −15.7]	−43.8 [−44.4, −43.3]
	2 %	−17.3 [−18.6, −16.0]	−22.4 [−23.6, −21.3]	−28.8 [−32.0, −26.1]	−39.7 [−40.3, −39.0]
	3 %	−23.2 [−27.1, −20.2]	−29.6 [−33.4, −26.6]	−35.8 [−38.8, −33.0]	−34.4 [−35.3, −33.6]
	5 %	−25.6 [−27.8, −23.5]	−29.6 [−31.3, −27.7]	−33.8 [−35.5, −32.0]	−18.7 [−20.4, −17.0]
Combined	exact	+0.7 [−0.5, +2.0]	+0.7 [−0.5, +1.9]	−8.8 [−9.9, −7.6]	−45.3 [−46.4, −44.2]
	0.5 %	−4.1 [−5.2, −3.0]	−5.6 [−6.6, −4.6]	−15.0 [−16.1, −13.9]	−44.5 [−45.7, −43.3]
	1 %	−7.9 [−9.0, −6.8]	−12.1 [−14.0, −10.5]	−20.6 [−21.7, −19.5]	−43.5 [−44.7, −42.4]
	2 %	−17.1 [−20.6, −14.1]	−18.6 [−20.7, −16.6]	−37.7 [−50.5, −27.5]	−38.9 [−40.2, −37.7]
	3 %	−15.9 [−19.0, −12.9]	−22.3 [−24.3, −20.3]	−49.5 [−60.2, −39.9]	−32.7 [−34.4, −30.7]
	5 %	−17.6 [−20.3, −14.9]	−23.2 [−25.7, −20.9]	−32.6 [−34.9, −30.3]	−17.1 [−19.2, −15.1]

Imp-RL retains higher completion than its teacher at the largest navigation errors. At 5 %, it completes 91.7 % of nominal missions and 89.8 % under combined faults, compared with 76.5 % and 65.5 % for the teacher. APF completes all missions and glideslope completes 96.0 % at this level, with costs 2.1–3.2 times those of Imp-RL. The remaining completion and paired-cost comparisons are reported in Table 1 and Supplementary Appendix A.

The fuel–time comparison shows that Imp-RL combines lower mean propellant expenditure with shorter mean flights than its teacher over successful missions (Fig. 3). The grey steps trace the classical Pareto front, and the shaded region contains fuel–time combinations dominated by classical designs. At 1–3 % navigation error, its successful-mission means are 1.85–2.43 m/s nominally and 2.70–3.51 m/s under combined faults. Relative to the teacher, it saves 8.5–27.7 % propellant and arrives 0.21–0.48 h earlier. Classical methods that arrive as early or earlier at these levels use 1.3–4.3 times its propellant.

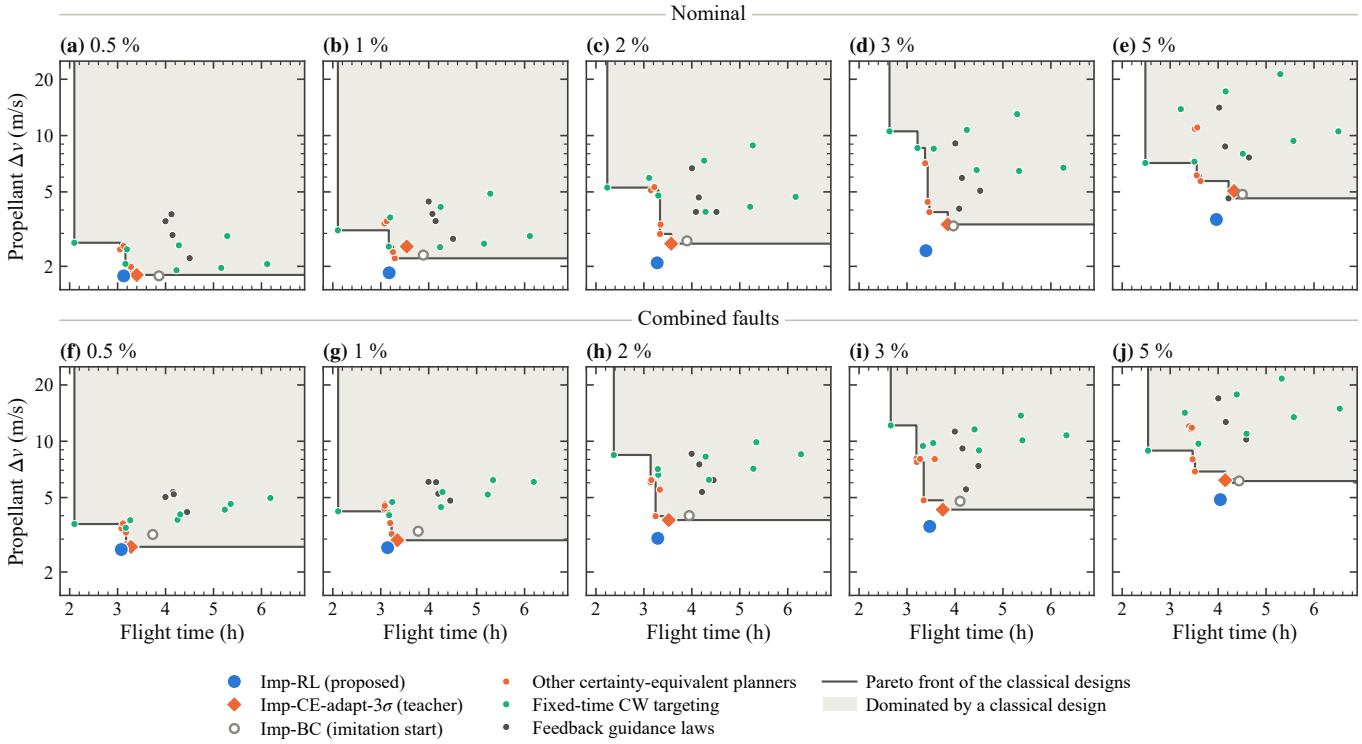


Figure 3: Fuel–time trade-offs at 0.5–5 % navigation error, with successful-mission means for nominal (top) and combined-fault cases (bottom).

With exact navigation, the CE planners provide the lowest nominal cost; Imp-CE achieves 2.51 m/s. Relative to Imp-CE-adapt-3σ, Imp-RL arrives 0.49 h earlier with 26 % more propellant and a 7.9 % higher paired cost. Under combined faults, its mean cost is 3.60 m/s and its paired difference from this planner is +0.7 % (95 % interval −0.5 to +2.0 %), with a flight 0.40 h shorter. Imp-RL also outperforms Imp-RL-exact in the noisy runs: at 2 % nominal error, their mean costs are 3.18 m/s and 6.19 m/s, respectively, while Imp-CE gives 6.15 m/s.

The same pattern appears when efficiency loss, misalignment and bias are tested separately. At 1–3 % navigation error, Imp-RL has the lowest mean cost in every single-fault scenario and completes 99.7–100 % of missions. Its paired cost reduction relative to the teacher is 8.3–19.0 % for efficiency loss, 7.5–26.4 % for misalignment and 15.0–22.1 % for bias. Under the extended OOD fault magnitudes it completes 90.7–93.5 % of missions, with a 3.6–7.0 % lower paired cost than the teacher. The complete single-fault results are in Supplementary Appendix A.1; Supplementary Appendix B presents the exact-navigation comparisons and the fault-response diagnostic.

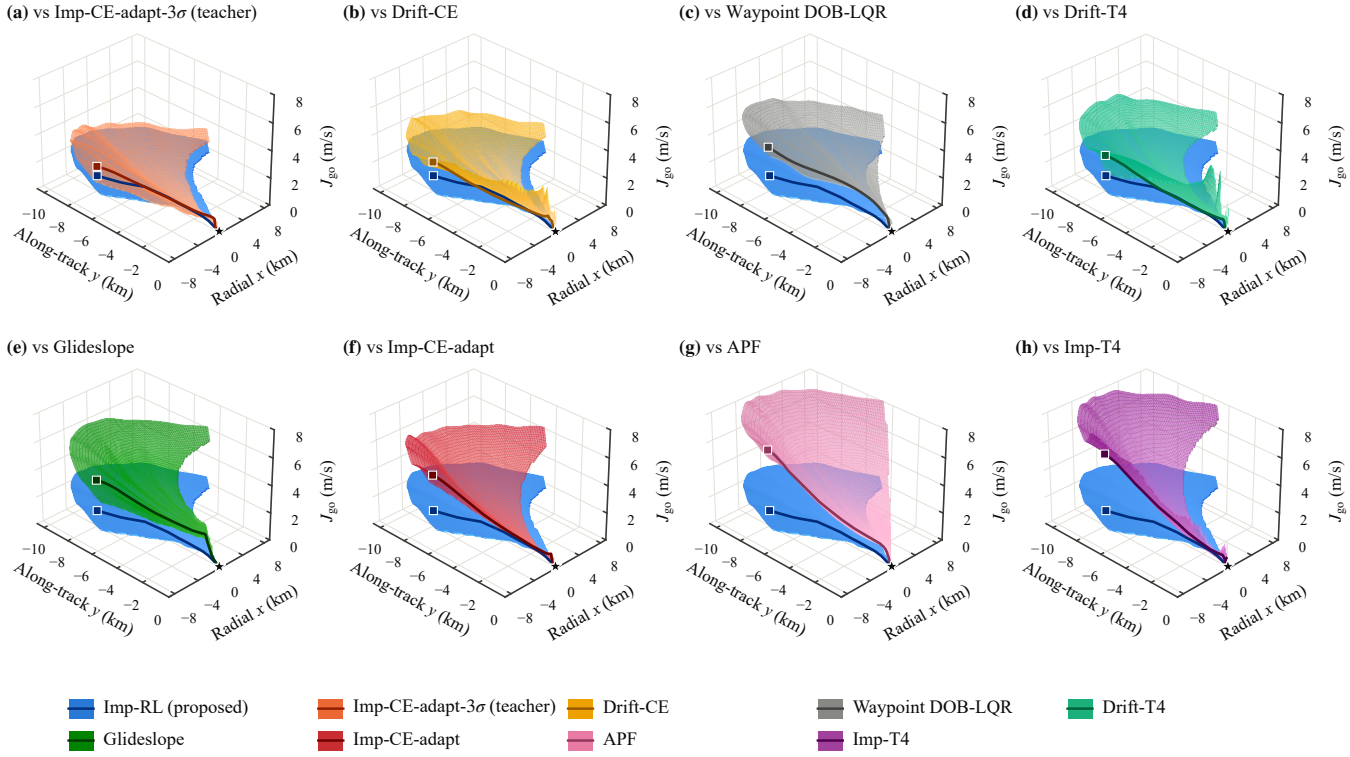


Figure 4: Position-conditioned remaining cost along 300 nominal CW missions at 2 % navigation error: Imp-RL (blue) and the classical guidance law named in each panel.

The remaining-cost profiles in Fig. 4 place most of the difference from impulsive planners beyond 3 km. Across 300 nominal CW approaches at 2 % navigation error, Imp-RL has the lowest mean remaining cost in the sampled bands beyond 1 km: 0.90 m/s at 1–3 km, 1.55 m/s at 3–6 km, and 2.05 m/s beyond 6 km. The teacher, Imp-CE-adapt and Imp-T4 differ from it by at most 0.11 m/s at 1–3 km; glideslope, APF and waypoint DOB-LQR have 0.77–1.45 m/s more remaining cost there. Beyond 6 km, the teacher gives 2.45 m/s, Drift-CE 2.89 m/s and the other designs 3.54–5.71 m/s. Imp-RL completes all 300 missions with mean total cost 3.17 m/s, compared with 287 and 3.81 m/s for the teacher, and 267–300 and 4.41–8.55 m/s for the other designs.

Each surface averages the remaining propellant and weighted flight time along that method's successful trajectories, excluding missions whose total cost exceeds three times its successful-mission median. The profiles therefore reflect the positions and histories visited by each method. Dark traces project example trajectories from their first plotted positions (squares) to handover (stars). Supplementary Appendix E gives the sampling and surface-construction details.

5.2. Learning burn timing and velocity increments

Reinforcement learning improves both cost and completion over the imitated policy. At 0.5–5 % navigation error, Imp-RL reduces Imp-BC's cost on jointly successful missions by 8.1–23.6 % nominally and 17.0–24.8 % under combined faults, while shortening flight time by 0.5–0.7 h. At 5 %, completion rises from 76.0 to 91.7 % nominally and from 64.0 to 89.8 % under combined faults (Tables A.2 and A.3).

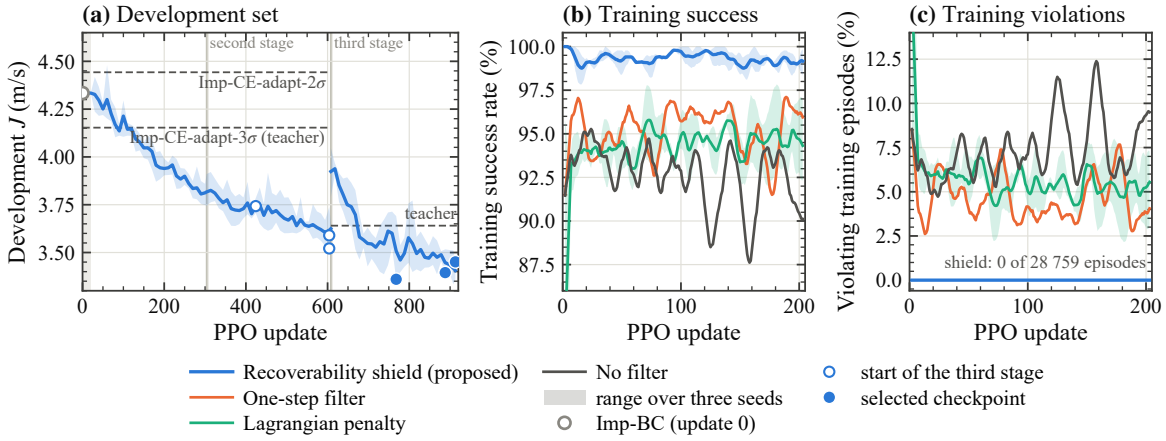


Figure 5: Training evolution: (a) development cost; (b, c) exact-navigation success and violation rates. Bands span three training seeds.

Development-set cost falls through the three training stages (Fig. 5a). The selected stage-2 checkpoints reduce mean J from 4.33 m/s after imitation to 3.52–3.74 m/s. Across the three seeds, stage three reduces mean J from 3.79–4.05 to 3.36–3.45 m/s on development missions drawn from its training mixture. Development stress-test completion for rear-plane starts at 2 % navigation error rises from 46.7 to 72.8 %. Supplementary Appendix D reports the full stage comparisons and exploration settings.

The mission in Fig. 6 illustrates the reduction in burn count and cost. Chosen near the orbital plane with propellant close to both methods' medians, this nominal 2 % navigation-error case starts with a 319 m position error and a reported $\sqrt{3}\sigma_p$ of 357 m, shown by the dashed curve in panel (c). Imp-CE-adapt fires 39 burns using 5.61 m/s; its 3σ teacher variant fires 9 burns using 2.47 m/s and arrives in 4.00 h. Imp-RL fires 7 burns using 1.91 m/s and arrives in 3.32 h. Its initial burn is 0.89 m/s, compared with the teacher's 1.03 m/s, followed by corrections at 2.08 and 2.75 h and four small corridor-approach burns. The navigation error decreases with range, and the CW replay reproduces both methods' Tudat results for this mission.

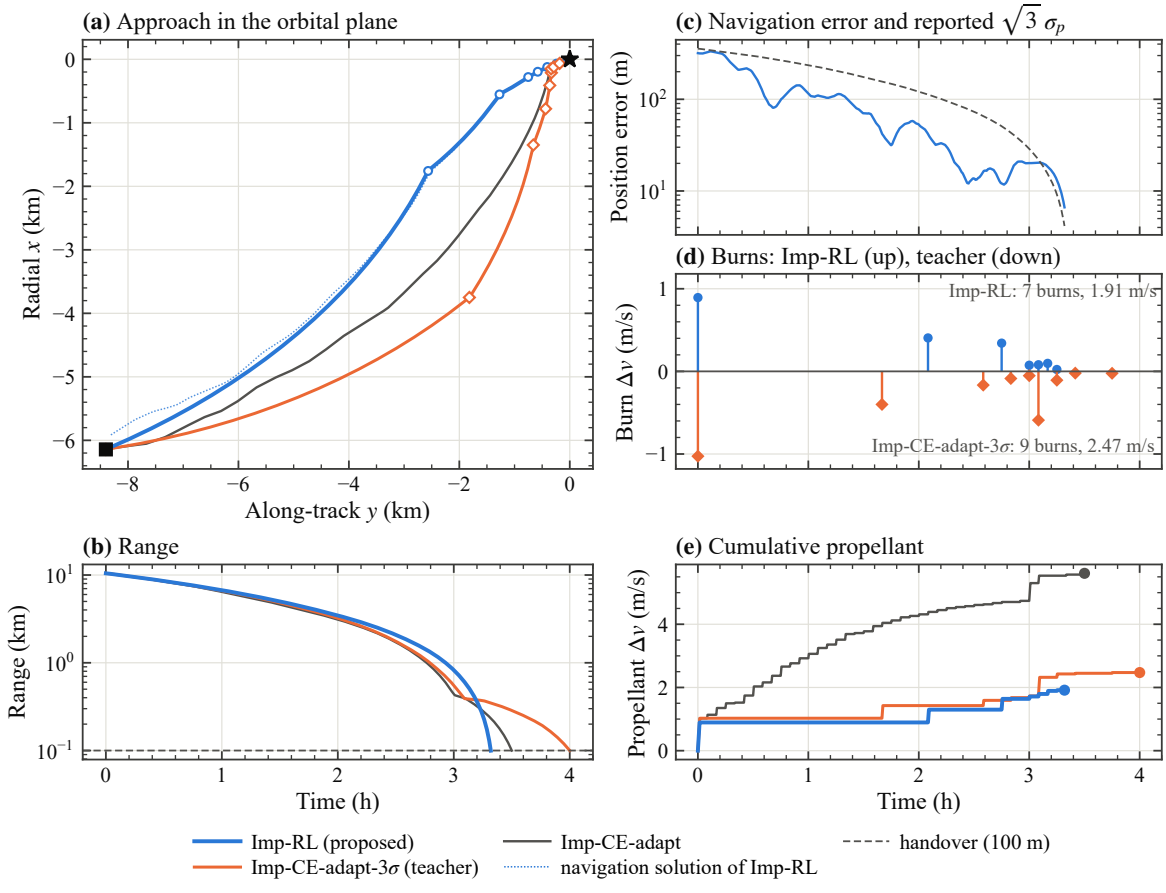


Figure 6: Representative nominal CW replay at 2 % navigation error: (a) trajectory; (b)–(e) range, position error, burn increments and cumulative propellant.

The burn-gate response is direction dependent in the velocity-estimate sweep (Fig. 7). The probe uses a coasting decision 4.4 km from the target in one of the 300 nominal CW missions underlying Fig. 4, with 2 % navigation error and $\sigma_v = 0.077$ m/s; the recurrent state and other observation inputs are fixed while velocity is offset along and across the line of sight, normalized by σ_v. The cross locates the true velocity relative to the estimate. The gate requests coasting inside its zero contour and a burn outside it, with execution also subject to the increment threshold of Section 3.1. Across 9255 coasting decisions from the 300 missions, a 24-direction scan gives a median nearest zero-gate offset of $4.6\,\sigma_v$ (interquartile range 3.0–6.1 σ_v). Median offsets along and across the line of sight are 7.9 and $4.9\,\sigma_v$; the teacher's increment threshold provides a reference magnitude of $3\sqrt{3}\sigma_v = 5.2\sigma_v$. For position offsets along the line of sight, 75 % of decisions retain a closed gate within $30\,\sigma_p$.

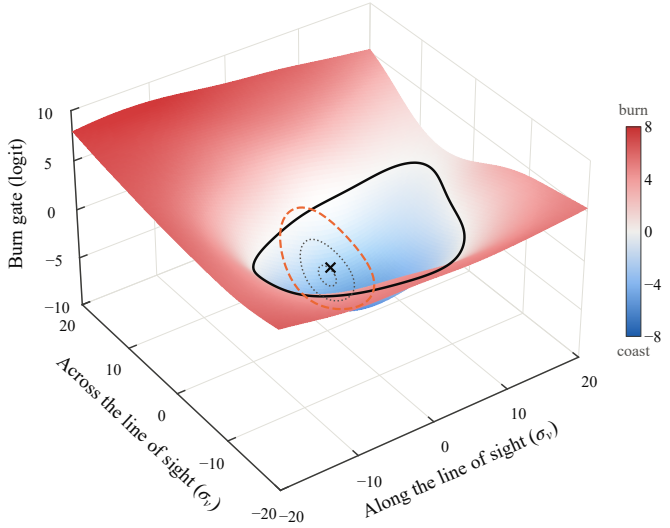


Figure 7: Velocity-offset response of the Imp-RL burn gate at 2 % navigation error. Black marks the zero-gate contour; orange marks the teacher's increment threshold.

With exact navigation, RL shortens flight time by 0.12–0.36 h in every scenario, with paired cost changes from −6.7 to +1.8 % (Table B.3). PPO from scratch never reaches handover within the training budget. In a pilot initialized from an earlier imitated policy, removing the imitation term in (34) lowers development cost from 4.66 to 3.87 before it returns to 4.39; the regularized run from the same start reaches 3.40.

5.3. *Constraint audit and safety ablations*

The true-trajectory audit records zero constraint violations for every shielded method through the termination of its test missions. For Imp-RL, this comprises 16 800 runs across the six fault scenarios and navigation-error levels, 3000 runs beyond the training envelope and 3000 stress runs with exact navigation or 2 % error (Table 2). Its 582 010 training episodes contain 5 violations, or 0.00086 %: four at the corridor boundary and one at the rear plane, including four episodes with faults. The shield uses a four-standard-deviation position margin throughout these runs.

The exact-navigation ablations show fewer violations with the recoverability shield (Table 2). Imp-RL-exact records none in 3600 test missions, 1500 stress missions and 28 759 training episodes. Matched training produces violation rates of 7.3 % without a filter, 4.6 % with the one-step filter and 5.2–5.8 % with a Lagrangian penalty (Fig. 5c). Test violations are concentrated under faults and near constraints: the one-step variant violates in 64 of 200 OOD missions, and the unshielded and one-step variants violate in 8–94 % of individual stress tests. Most crossings occur at the corridor, with additional rear-plane crossings. Shielded end-to-end PPO also records zero violations, with 21–41 % mission completion; its unshielded training violates in 87–99 % of episodes.

The shielded trajectories pass close to the constraints (Fig. 8b–d). Panels (b, c) accumulate mission fractions by minimum corridor margin in the OOD and stress tests, respectively; panel (d) gives the minimum rear-plane margin in OOD. In the OOD and stress tests, respectively, 30 % and 57 % of missions come within 1.1 m of the corridor boundary. Across all 5100 exact-navigation test and stress missions, the minimum corridor and rear-plane margins are 0.50 m and 0.73 m.

Table 2: Constraint audit in native Tudat. Violation columns give violating/tested missions; training entries give the percentage of violating episodes. Learned methods pool three seeds where available. Imp-RL pools the navigation-error levels specified in Section 4.2; the remaining rows use exact navigation.

Policy and safety mechanism	Training	In distribution		OOD	Stress
	viol. (%)	viol.	success (%)	viol.	viol.
Imp-RL (proposed), recoverability shield	0.001	0/14400	99.1	0/2400	0/3000
Imp-RL-exact, recoverability shield	0.0	0/3000	99.4	0/600	0/1500
Imp-RL-exact, one-step filter	4.6	35/1000	96.2	64/200	263/500
Imp-RL-exact, Lagrangian penalty	5.5	120/3000	96.0	198/600	–
Imp-RL-exact, no filter	7.3	46/1000	95.4	69/200	292/500
E2E, recoverability shield	0.0	0/1000	28.2	0/200	0/500
E2E, one-step filter	92.9	1814/3000	6.1	382/600	846/1500
E2E, Lagrangian penalty	34.4	1/3000	0.0	7/600	399/1500
E2E, no filter	96.7	1352/3000	1.1	324/600	828/1500

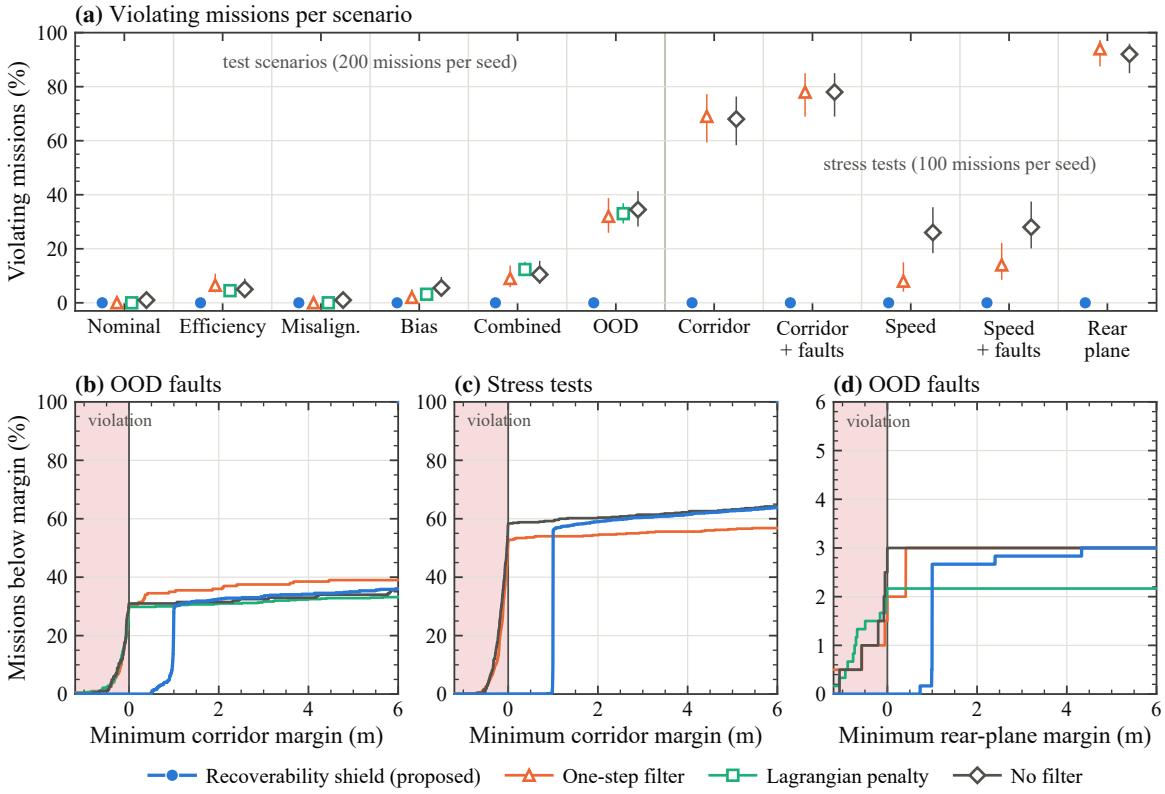


Figure 8: Exact-navigation constraint audit: (a) violation rates with Wilson 95 % intervals; (b)–(d) cumulative distributions of minimum corridor and rear-plane margins.

5.4. Exact-navigation trajectories and performance distributions

Exact-navigation runs resolve the transfer geometry and mission-to-mission variability. Imp-RL-exact, trained with exact navigation to imitate Imp-CE, is evaluated through trajectory replays and performance distributions alongside Imp-RL and the classical designs.

In the CW ensemble of Fig. 9, all 20 000 nominal missions and 19 533 of 20 000 combined-fault missions reach handover. No constraint violation is recorded in either ensemble. These runs use Imp-RL-exact seed 1 and mission seeds distinct from the test set. The nominal median burn count is 19, with 80 % of missions firing 12–29 burns. The median first burn is 0.80 m/s, and the median first-burn share of mission propellant is 46 %; subsequent burns have median magnitude 0.031 m/s, with 41 % occurring in the last kilometre. Panel (a) shows later burns beyond 1 km for one mission in twenty; the black example is near the orbital plane and closest to median propellant. Nominal trajectories reach off-axis angles up to 86° beyond 1 km. Of the 3989 missions exceeding 60°, 3985 start below the target orbit and drift forward beneath it, accounting for 40 % of below-orbit starts. Overall, 89 % enter the last 500 m from below the target orbit (Fig. 9b).

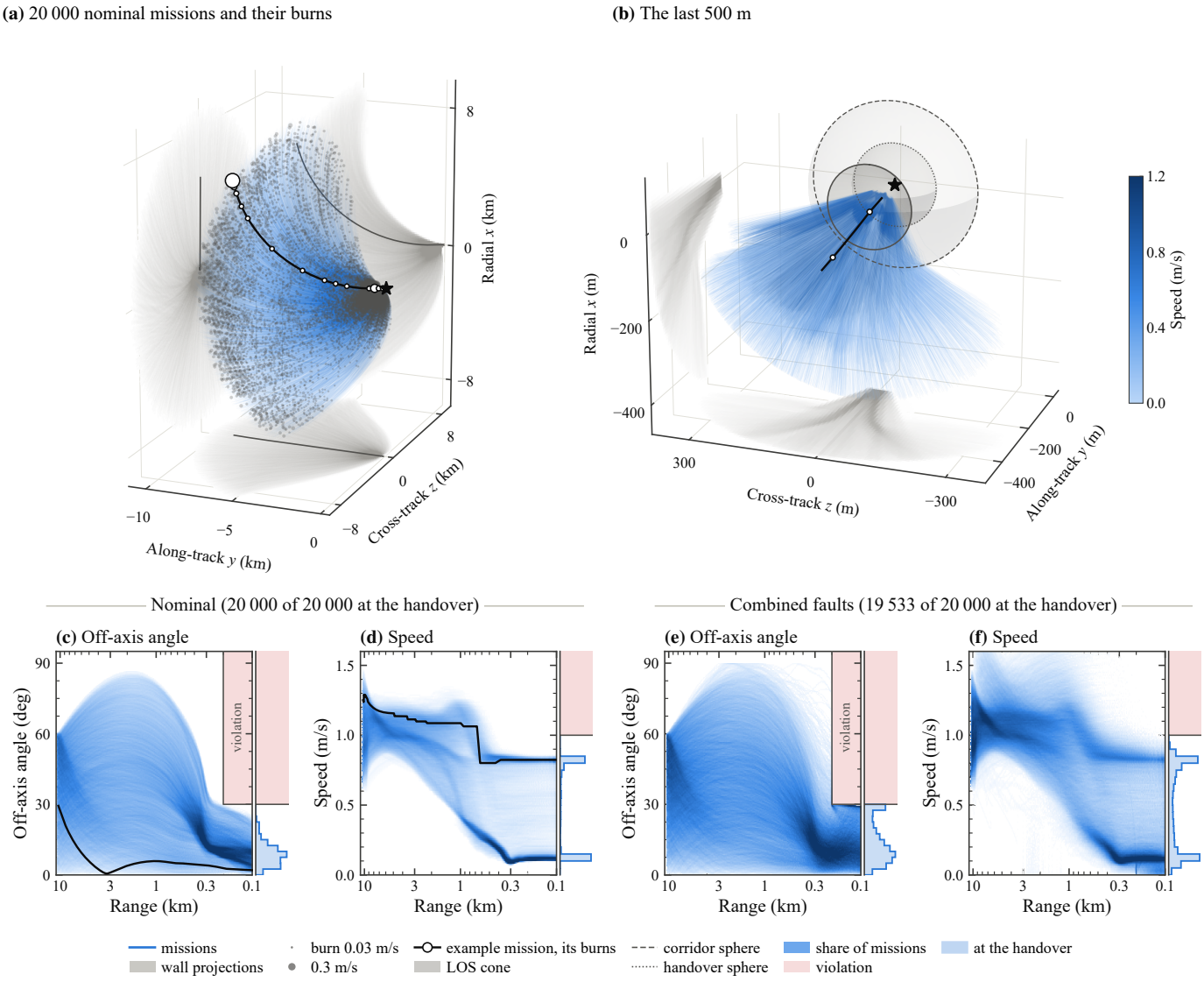


Figure 9: CW trajectory ensembles for Imp-RL-exact: (a, b) transfer and terminal geometry; (c)–(f) angle and speed distributions over successful nominal and combined-fault missions.

The trajectories converge toward the corridor and form two handover-speed groups (Fig. 9c–f). From 350 m inward, 90 % of nominal missions lie within 30° of the approach axis. At the 10 s recording interval, maximum angles inside the corridor are 29.6° nominally and 29.7° under combined faults. The nominal handover angle has median 8.5° and maximum 25.5°; the faulty maximum is 29.4°. Median nominal speed is 0.93 m/s between 3 and 1 km, before separating toward handover speeds near 0.8 and 0.1 m/s, the teacher's drift-speed bounds. Of above-orbit starts, 85 % brake near 650 m and retain about 0.8 m/s; 88 % of below-orbit starts approach the slow group. Median arrivals are 2.8 h and 4.1 h, respectively. Maximum handover speeds are 0.85 m/s nominally and 0.98 m/s under faults, within the 1 m/s limit. Faulty missions have a median of 22 burns and wider near-field distributions: 174 come within 50 m of the rear plane, with a minimum margin of 1.0 m. Of the 467 combined-fault timeouts, 92 % hold just outside the corridor sphere at 201–209 m.

Among the methods in Fig. 10, Imp-RL has the earliest median arrival in all three scenarios. Its flight-time interquartile range is also narrower than those of Imp-RL-exact and the CE planners. Its median flight times are 2.94 h nominally, 2.86 h under combined faults and 2.93 h in OOD, with median propellant of 1.77 m/s, 2.39 m/s and 3.19 m/s, respectively. Imp-RL-exact also arrives earlier than the shown classical methods under combined and OOD faults: its median times of 3.04 h and 3.01 h compare with 3.09–3.10 h for the CE planners and Drift-T3 and 4.13–4.16 h for Drift-T4 and glideslope. Its median propellant in these scenarios, 2.50 m/s and 3.49 m/s, lies below the drift and glideslope designs and above Imp-CE and Imp-CE-adapt. Nominally, the CE and drift designs use less median propellant, 1.30–1.54 m/s compared with 1.74 m/s for Imp-RL-exact. Dashed lines provide the Imp-RL medians as a reference. Success percentages are printed below the flight-time panels; $+n$ counts observations outside the plotted limits.

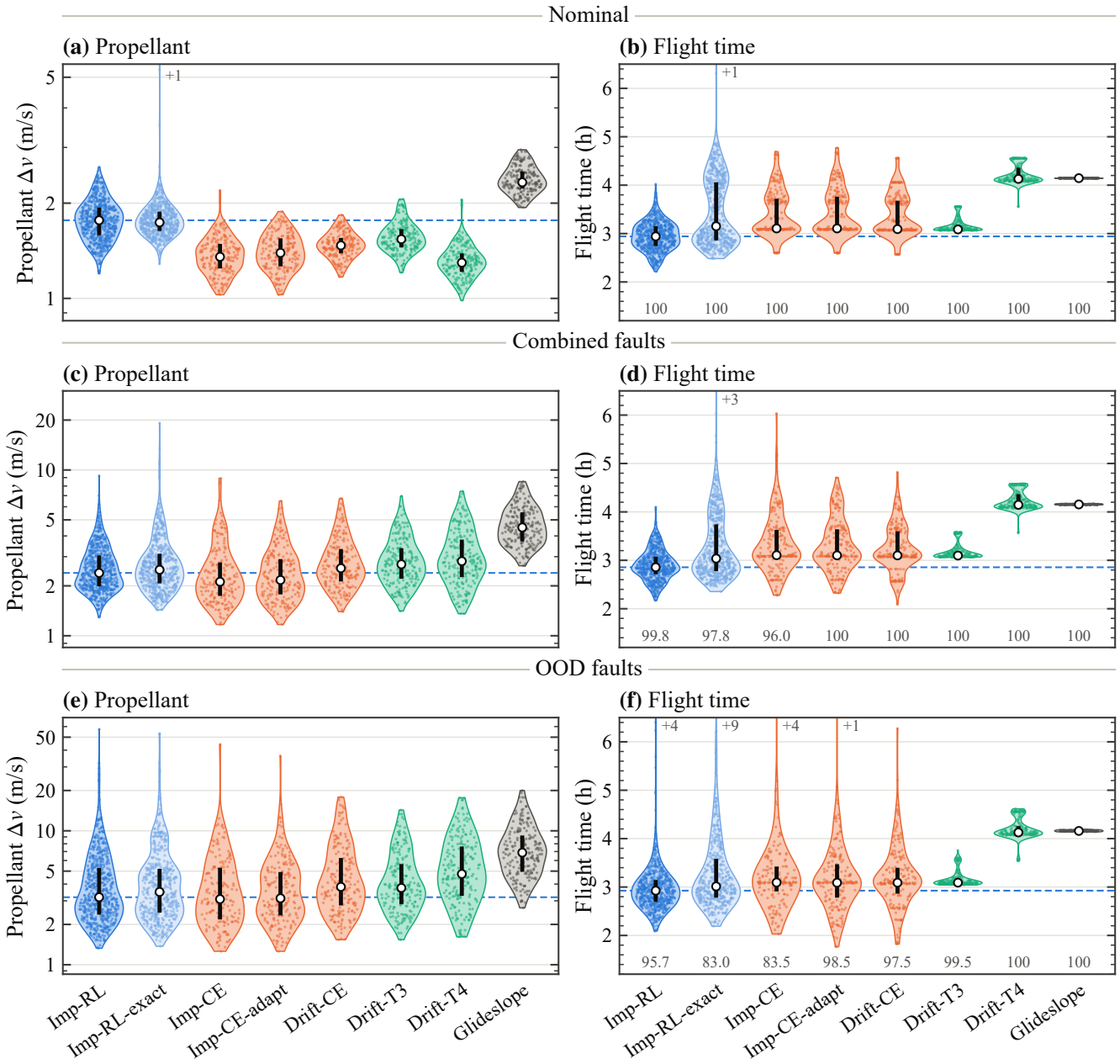


Figure 10: Propellant and flight-time distributions for successful exact-navigation missions. White circles and black bars mark medians and interquartile ranges.

5.5. Beyond the training envelope

The extended tests change the magnitude or time dependence of the actuator faults. OOD-2 uses efficiencies down to 0.2, misalignment up to 10° and bias up to 5×10^{-4} m/s^2. The Thruster case assigns an efficiency to each firing direction; Degradation decreases efficiency over time; Outage introduces intermittent loss of thrust; and Unmodelled combines these changes. Imp-RL completes 99.7, 100 and 95.8 % of the Thruster, Degradation and Outage missions, and 89.8 % of Unmodelled missions (Fig. 11). In OOD-2 it completes 88.5 %, compared with 96.5 % for Imp-CE-adapt and 63.7 % for Imp-RL-exact. The variant trained on the wider fault envelope, Imp-RL-W, completes 93.5 % of OOD-2 missions.

The stress tests place the initial state next to the corridor, rear plane or speed boundary. With exact navigation, Imp-RL completes 100 % of the nominal corridor and rear-plane missions and 99.3 % of the combined-fault corridor missions. At 2 % navigation error, completion is 97.3, 82.0 and 95.3 %, respectively. Some missions exhaust the admissible-command set at the first step; among its 900 corridor and rear-plane runs at this error level, ten further failures occur after the first step. No constraint violations are recorded in these extended tests. Supplementary Appendix C gives the full scenario definitions, performance comparisons and failure counts.

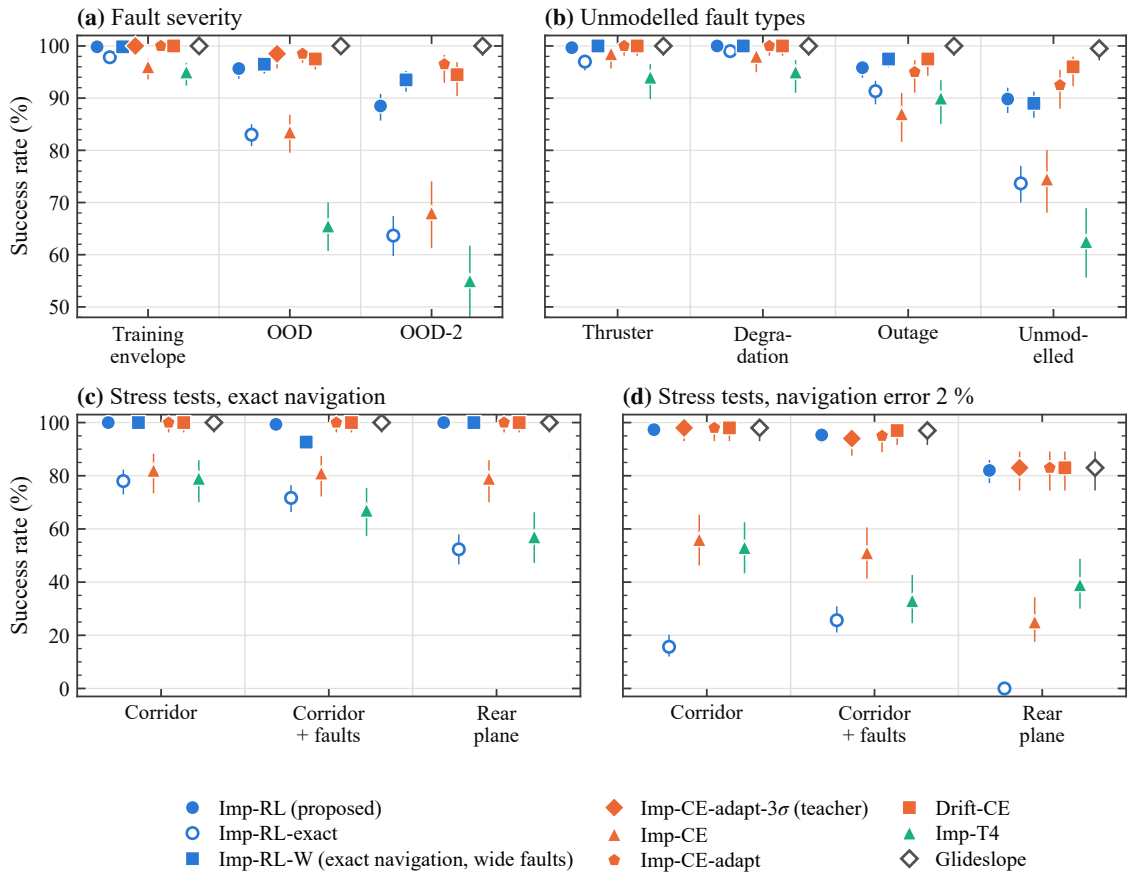


Figure 11: Mission completion under (a) fault severity, (b) unmodelled faults and (c, d) stress conditions, with Wilson 95 % intervals. Navigation is exact in (a)–(c), with 2 % error in (d).

5.6. Model mismatch: CW versus Tudat

CW replay and Tudat propagation give the same outcome in 99.9 % of Imp-RL's 22 800 missions. Across methods and scenarios, mean performance indices differ by 0.4 % on average; the CW audit records no violations across 111 200 runs. For the 120 paired comparisons between Imp-RL and ten other methods, the median discrepancy is 0.05 percentage points and 90 % lie within 0.4 points. The largest difference is the 1 % nominal comparison against the teacher: −16.8 % in CW and −22.3 % in Tudat. One teacher mission reaches handover in 11.2 h in Tudat and times out in CW. The differential perturbations in these GEO approaches at 10 km are of order 10^{-8} m/s^2.

6. Discussion

Navigation uncertainty changes the value of a correction. Repeated certainty-equivalent targeting reacts to changes in the estimated state, including changes caused by the navigation error. A magnitude dead band reduces this expenditure, but can also delay useful corrections. This tradeoff helps explain why the teacher spends less propellant than unrestricted retargeting yet loses completion at larger errors. The learned policy responds differently by direction: at fixed recurrent state, its coast region extends farther along the line of sight than across it (Fig. 7). Equally sized velocity-estimate changes can therefore lead to different burn decisions, allowing corrections in one direction while suppressing them in another. In the example mission of Fig. 6, a smaller initial increment, two later corrections and small corridor trims yield both lower expenditure and earlier arrival than the dead-band teacher.

The comparison with Imp-BC links this behaviour to optimization over the complete approach. Imitation supplies a usable targeting strategy at the states visited by the student; its loss rewards agreement with the teacher at each decision. PPO instead evaluates increments and coasts through their contribution to subsequent propellant expenditure, flight time and mission completion. The lower mission cost and shorter flight after PPO (Fig. 3) are consistent with revising the sequence of manoeuvres across the transfer. Increment exploration permits such changes, while the decaying imitation term keeps the targeting strategy available during optimization. In the pilot comparison, cost continues to fall with imitation regularization after the unregularized run loses its initial gain.

The shield checks whether a command leaves room to brake before reaching a boundary. A command can pass a one-second constraint check while leaving a stopping distance that intersects the corridor boundary or rear plane. Recoverability checking evaluates that subsequent braking segment, with the assumed deceleration and uncertainty margins of Section 3.3. The matched ablations concentrate violations at these boundaries, especially under faults and near-constraint starts. In the corresponding shielded runs, the minimum-margin distributions show trajectories approaching within roughly a metre of the corridor while retaining positive audited margins (Fig. 8).

The prescribed navigation-error process isolates the guidance response to uncertain estimates. Coupling the policy to an angles-only navigation filter would extend the timing problem to information gathering, connecting the present correction policy to observability-aware far-range planning [15]. A manoeuvre would then change both the relative trajectory and the observability of range, so its value would include the effect of improved subsequent estimates on later corrections. The next question is when improved estimates justify an extra burn and how they change the preferred coast intervals.

7. Conclusion

Recurrent impulsive guidance learns when to correct an uncertain rendezvous trajectory and how much velocity increment to apply, with recoverability checking at the thrust-command level. In native Tudat evaluation of the GEO approach from approximately 10 km to 100 m, it achieves the lowest mean fuel–time cost among the evaluated methods in nominal and combined-fault missions at navigation errors of 0.5–5 % of range. Cost is 4–26 % below that of its adaptive teacher on paired successful missions. At 5 % error, nominal completion rises from the teacher's 76.5 % to 91.7 %. The true-trajectory audit records zero constraint violations through termination in all 22 800 test missions.

References

[1] W. Fehse, Automated Rendezvous and Docking of Spacecraft, no. 16 in Cambridge Aerospace Series, Cambridge University Press, 2003. doi:10.1017/CBO9780511543388.

[2] S. D'Amico, J.-S. Ardaens, G. Gaias, H. Benninghoff, B. Schlepp, J. L. Jørgensen, Noncooperative rendezvous using angles-only optical navigation: System design and flight results, Journal of Guidance, Control, and Dynamics 36 (6) (2013) 1576–1595. doi:10.2514/1.59236.

[3] J.-S. Ardaens, G. Gaias, Flight demonstration of spaceborne real-time angles-only navigation to a noncooperative target in low Earth orbit, Acta Astronautica 153 (2018) 367–382. doi:10.1016/j.actaastro.2018.01.044.

[4] G. Gaias, J.-S. Ardaens, Flight demonstration of autonomous noncooperative rendezvous in low Earth orbit, Journal of Guidance, Control, and Dynamics 41 (6) (2018) 1337–1354. doi:10.2514/1.G003239.

[5] 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.

[6] H. B. Hablani, M. L. Tapper, D. J. Dana-Bashian, Guidance and relative navigation for autonomous rendezvous in a circular orbit, Journal of Guidance, Control, and Dynamics 25 (3) (2002) 553–562. doi:10.2514/2.4916.

[7] I. Lopez, C. R. McInnes, Autonomous rendezvous using artificial potential function guidance, Journal of Guidance, Control, and Dynamics 18 (2) (1995) 237–241. doi:10.2514/3.21375.

[8] A. Weiss, M. Baldwin, R. S. Erwin, I. Kolmanovsky, Model predictive control for spacecraft rendezvous and docking: Strategies for handling constraints and case studies, IEEE Transactions on Control Systems Technology 23 (4) (2015) 1638–1647. doi:10.1109/TCST.2014.2379639.

[9] W.-H. Chen, J. Yang, L. Guo, S. Li, Disturbance-observer-based control and related methods—an overview, IEEE Transactions on Industrial Electronics 63 (2) (2016) 1083–1095. doi:10.1109/TIE.2015.2478397.

[10] 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.

[11] L. Federici, B. Benedikter, A. Zavoli, Deep learning techniques for autonomous spacecraft guidance during proximity operations, Journal of Spacecraft and Rockets 58 (6) (2021) 1774–1785. doi:10.2514/1.A35076.

[12] 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.

[13] B. Gaudet, R. Linares, R. Furfaro, Adaptive guidance and integrated navigation with reinforcement meta-learning, Acta Astronautica 169 (2020) 180–190. doi:10.1016/j.actaastro.2020.01.007.

[14] 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.

[15] M. C. Wijayatunga, R. Armellin, H. Holt, Robust trajectory design and guidance for far-range rendezvous using reinforcement learning with safety and observability considerations, Aerospace Science and Technology 159 (2025) 109996. doi:10.1016/j.ast.2025.109996.

[16] M. Stephenson, H. Schaub, Two approaches to autonomous spacecraft inspection: Shielded reinforcement learning versus probabilistic roadmaps, The Journal of the Astronautical Sciences 73 (5) (2026) 89. doi:10.1007/s40295-026-00631-w.

[17] 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

[18] A. Ray, J. Achiam, D. Amodei, Benchmarking safe exploration in deep reinforcement learning, Tech. rep., OpenAI (2019).
URL https://cdn.openai.com/safexp-short.pdf

[19] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, U. Topcu, Safe reinforcement learning via shielding, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, pp. 2669–2678. doi:10.1609/aaai.v32i1.11797.

[20] T. Gurriet, M. Mote, A. Singletary, P. Nilsson, E. Feron, A. D. Ames, A scalable safety critical control framework for nonlinear systems, IEEE Access 8 (2020) 187249–187275. doi:10.1109/ACCESS.2020.3025248.

[21] 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.

[22] 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.

[23] 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.

[24] A. Banerjee, K. Rahmani, J. Biswas, I. Dillig, Dynamic model predictive shielding for provably safe reinforcement learning, in: Advances in Neural Information Processing Systems, Vol. 37, Curran Associates, Inc., 2024, pp. 100131–100159. doi:10.52202/079017-3177.

[25] D. P. Nguyen, K.-C. Hsu, W. Yu, J. Tan, J. F. Fisac, Gameplay filters: Robust zero-shot safety through adversarial imagination, in: Proceedings of The 8th Conference on Robot Learning, Vol. 270 of Proceedings of Machine Learning Research, PMLR, 2025, pp. 387–407.
URL https://proceedings.mlr.press/v270/nguyen25a.html

[26] M. Wijayatunga, N. Wallace, S. Sukkarieh, R. Armellin, Learning safety-guaranteed, non-greedy control barrier functions using reinforcement learning, in: 2026 IEEE Aerospace Conference, IEEE, 2026, pp. 1–14. doi:10.1109/aero66936.2026.11520115.

[27] S. Ross, G. J. Gordon, J. A. Bagnell, A reduction of imitation learning and structured prediction to no-regret online learning, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Vol. 15 of Proceedings of Machine Learning Research, PMLR, 2011, pp. 627–635.
URL https://proceedings.mlr.press/v15/ross11a.html

[28] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017). doi:10.48550/arXiv.1707.06347.
URL https://arxiv.org/abs/1707.06347

[29] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, S. Levine, Learning complex dexterous manipulation with deep reinforcement learning and demonstrations, in: Proceedings of Robotics: Science and Systems, Pittsburgh, Pennsylvania, 2018. doi:10.15607/RSS.2018.XIV.049.

[30] 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, in: Europlanet Science Congress 2022, Vol. 16 of EPSC Abstracts, 2022, ePSC2022-253. doi:10.5194/epsc2022-253.
URL https://meetingorganizer.copernicus.org/EPSC2022/EPSC2022-253.html

[31] M. Hausknecht, P. Stone, Deep reinforcement learning in parameterized action space, in: International Conference on Learning Representations (ICLR), 2016.
URL https://www.cs.utexas.edu/~AustinVilla/details/ICLR16-hausknecht.html

[32] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, pp. 1724–1734. doi:10.3115/v1/D14-1179.

[33] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, P. Abbeel, Asymmetric actor critic for image-based robot learning, in: Proceedings of Robotics: Science and Systems, Pittsburgh, Pennsylvania, 2018. doi:10.15607/RSS.2018.XIV.008.

[34] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, High-dimensional continuous control using generalized advantage estimation, in: International Conference on Learning Representations (ICLR), 2016.
URL https://arxiv.org/abs/1506.02438