(Preprint) AAS 25-532

ENHANCING CISLUNAR OPERATIONS BY INTEGRATING REINFORCEMENT LEARNING INTO CLASSICAL RELATIVE GUIDANCE METHODS

Carlo Zambaldo[*], Luca Giorcelli[†], and Mauro Massari[‡]

The growing interest in cislunar exploration demands guidance systems capable of handling the nonlinear dynamics of Near-Rectilinear Halo Orbits (NRHOs). Ensuring autonomy and minimizing human intervention are essential in such deep space missions, where communication delays impose strict self-sufficiency. This work investigates the integration of Reinforcement Learning (RL) into classical relative guidance strategies to enhance efficiency, flexibility and robustness during proximity operations without compromising safety and interpretability of the system. The proposed algorithm employs a hybrid approach: Approximating Sequence of Riccati Equations (ASRE) is used for trajectory optimization, Artificial Potential Field (APF) ensures path constraint handling, and Sliding Mode Control (SMC) provides robust control. The RL agent operates at a high level, selecting among predefined guidance configurations based on system observations. The framework is validated through extensive simulations in a CR3BP environment, focusing on rendezvous and docking in different scenarios. Results show that the RL-aided system outperforms traditional APF+SMC configurations, with significant improvements in propellant efficiency, Time Of Flight (TOF) and decision adaptability, while maintaining operational safety even under partial system failure. The proposed structure offers a scalable and modular solution to future autonomous guidance challenges in complex environments.

INTRODUCTION

The growing ambition for cislunar exploration fuels the demand for robust autonomous algorithms capable of performing rendezvous, docking and proximity operations in highly non-linear orbital environments. In particular, spacecrafts operating in Near-Rectilinear Halo Orbits (NRHOs) encounter third-body perturbations and other disturbances, requiring autonomous systems to enhance adaptability and minimize human intervention, particularly when communication delays impose self-sufficiency. Traditional guidance and control strategies, such as Proportional Navigation (PN),[1] Optimal Feedback Guidance Laws,[2] Artificial Potential Field (APF),[1] Zero-Effort-Miss / Zero-Effort-Velocity (ZEM/ZEV),[1] Sliding Mode Control (SMC)[3] and Model Predictive Control (MPC)[3] have proven reliability in LEO scenarios. Several studies are also dedicated to the relative motion and rendezvous problem in cislunar space. For instance, Galullo et al. suggested using SDRE to compute the trajectory considering a safe approaching cone acting as path constraint.[4] Yet, these methods can suffer from limited real-time adaptability, reduced flexibility and a lack of autonomy, usually requiring human-supervision. To overcome these limitations, recent studies have

[*]Research Fellow, Department of Aerospace Science and Technology, Politecnico di Milano, carlo.zambaldo@polimi.it
[†]PhD candidate, Department of Aerospace Science and Technology, Politecnico di Milano, luca.giorcelli@polimi.it
[‡]Associate Professor, Department of Aerospace Science and Technology, Politecnico di Milano, mauro.massari@polimi.it

explored the usage of Machine Learning (ML) techniques to solve directly the guidance problem.[5] These purely data-driven approaches can exhibit remarkable adaptiveness and autonomy, but may suffer from the lack of interpretability and strong analytical guarantees. Consequently, they can compromise safety if the learned policy encounters an unseen condition, since they directly dictate spacecraft control authority. The core objective of this research is to demonstrate how RL can be effectively integrated with established guidance algorithms to preserve system reliability while enhancing efficiency, adaptability and autonomy. Specifically, the guidance architecture presented here employs a high-level RL agent to exploit classical guidance methods, leveraging well-tested frameworks, such as ASRE, APF and SMC, ensuring collision avoidance and robust tracking while optimizing the trajectory and enhancing autonomy.

PROBLEM STATEMENT

To obtain a representative scenario, the L2 southern 9:2 lunar synodic resonant NRHO – the same orbit of the Lunar Gateway – was adopted as the operational trajectory for the target. The chaser spacecraft, considered as a logistic module, was assumed to be in proximity of the target, after the NRHO insertion and rendezvous maneuvers, which should position it within $10\,\mathrm{km}$ of the target – inside the Rendezvous Sphere but outside the Approach Sphere (AS) – with a moderately low initial velocity.

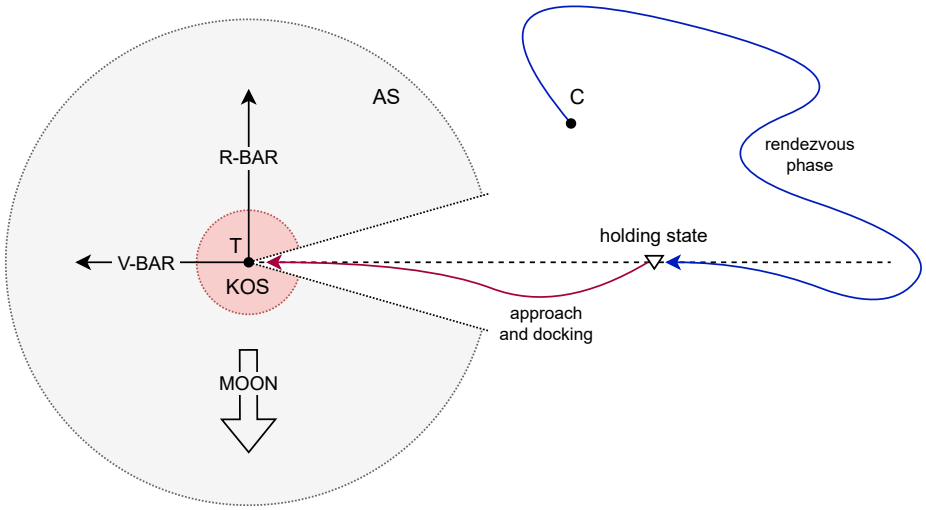


Figure 1: Proximity Operations Constraints and Phases

The mission involves performing relative maneuvers to achieve a predefined goal. As shown in Figure 1, to be consistent with standard safety zones[6] and to align the guidance design with conventional mission operations, two distinct phases were identified:

- The *Rendezvous Phase* (or phase 1), where the chaser spacecraft performs a sequence of maneuvers to approach a safe configuration (namely the holding state).
- The *Final Approach and Docking Phase* (or phase 2), where the spacecraft approaches the target for docking.

Model Setup

Figure 2 depicts the framework used to simulate the problem. Once the initial conditions[7] are set, the *Physical Environment* block propagates the dynamics of both the target and the chaser. The two states are then processed by the *OBNavigation*, which computes the target state in a moon-centered synodic coordinate frame and the chaser relative state in the target's Local Vertical Local Horizontal (LVLH) frame. These values are passed to the *OBGuidance* block, the core focus of this study, which determines the guidance command for the chaser spacecraft. Finally, the resulting control action is fed into an ideal controller (*OBControl*) that binds it to a predefined range and outputs the command in the synodic reference frame. This modular design allows for an easy integration of custom algorithms and expansion to other mission scenarios.

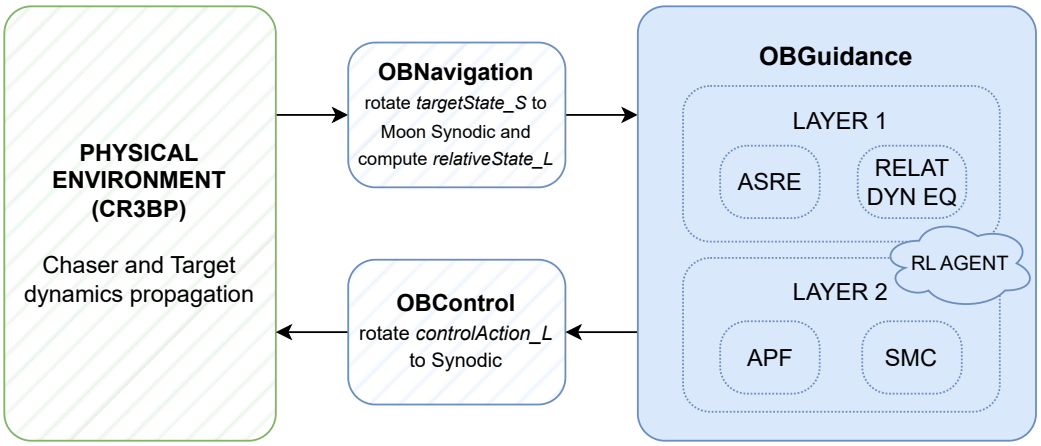


Figure 2: Simulation Framework

Spacecrafts Data

The physical quantities used for modeling the spacecrafts are those in Table 1. The areas were retrieved graphically from a Gateway blueprint*, the masses from NASA and ESA, the thrust parameters were assumed similar to those of R-4D-11, which are installed in ESA's Automated Transfer Vehicle. Lastly, the reflectivity data was obtained from Reference 8. Please observe that the target spacecraft was assumed to be cooperative and passive (i.e. with no orbital manoeuvring capabilities), and its orbit was assumed to be known. To be conservative, only one thruster was considered for the chaser.

Table 1: Specifications of Chaser and Target

Parameter	Symbol	Units	Chaser	Target
Spacecraft Mass	m	kg	15000	40000
SRP Area	A	m^2	18	110
Specular Reflectivity	ρ_s	-	0.5	0.5
Diffuse Reflectivity	ρ_d	-	0.1	0.1
Specific Impulse	I_s	s	270	n/a
Maximum Thrust	T_{max}	N	490	n/a

*https://www.esa.int/ESA_Multimedia/Images/2023/03/Gateway_blueprint

Environment Model

In this work, the dynamics were propagated using two Earth-Moon-spacecraft Circular Restricted Three-Body Problem (CR3BP) models in the classical synodic reference frame, reported in Figure 3. To validate the methodology, Solar Radiation Pressure was introduced as the primary environmental disturbance. In addition, random noise was later included, without significantly affecting the results.

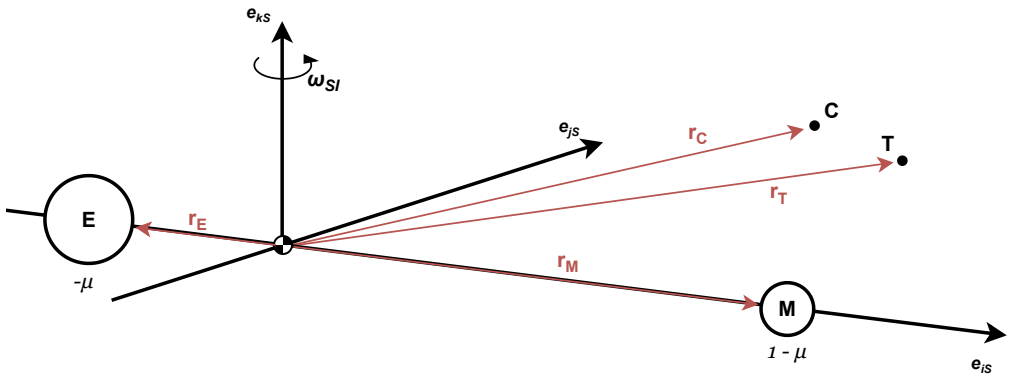


Figure 3: Classic Synodic Reference Frame

Relative Dynamics Model

A relative dynamics propagation model was used to perform onboard optimization via ASRE. Specifically, the Linearized Non-Linear Relative Equations of motion (LNLRE) in the LVLH frame were used.[9] This was done due to the inadequacy of two-body equations in restricted three-body contexts,[9] especially near the aposelene arc, where third-body effects are significant.[10] The target LVLH frame, shown in Figure 4, was derived from a Moon-centered synodic frame.

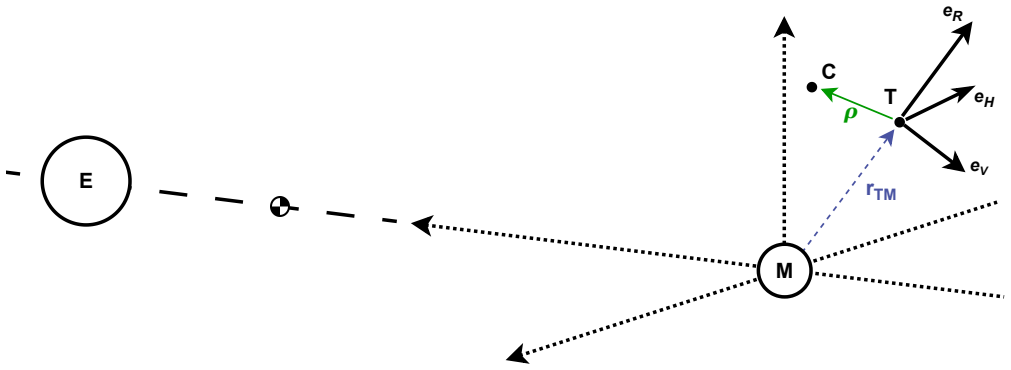


Figure 4: LVLH Reference Frame

GUIDANCE ALGORITHM

To address the problem at hand, the proposed guidance scheme consists of two nested loops, shown in Figure 5. The inner layer (L2) operates at a frequency of 5 Hz, continuously adjusting control inputs based on real-time feedback to ensure precise trajectory corrections. The outer layer (L1) is activated by an RL agent, which updates its action every 100 guidance steps (i.e. with a frequency of 0.05 Hz). The agent executes its action only at the first step of each cycle, setting the "SKIP" action for the remaining 99 steps, to avoid unnecessary L1 recomputations. Overall, the algorithm leverages:

1. **ASRE**, providing an optimized reference trajectory to guide the chaser.
2. **APF**, which enforces real-time collision avoidance by generating artificial repulsive forces defined in the target LVLH.
3. **SMC**, that combines the various contributions, attenuating disturbances and guaranteeing finite-time convergence of the tracking error.
4. The **RL agent**, which eventually regulates the system, observing states and choosing the best guidance configuration to optimize overall performance.

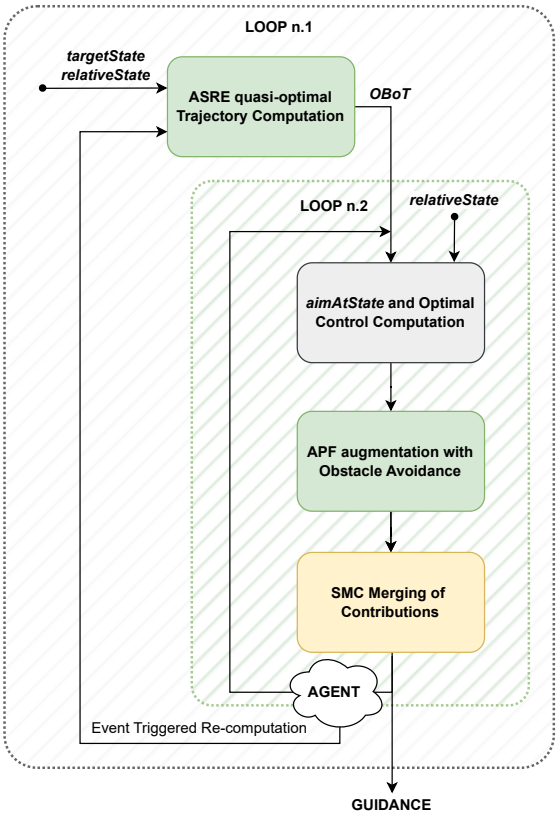


Figure 5: Guidance Algorithm Flowchart

Figure 6 illustrates the algorithm flowchart: once the agent selects an action, two configurations are possible: nominal mode and safe mode. The "SKIP" action maintains the previous mode. With the "COMPUTE" action, the algorithm calculates an optimal trajectory via ASRE, while "DELETE" removes the available optimal trajectory. In the inner layer (L2) the attractive field is computed and merged with the APF repulsive field. Finally, the obtained sliding surface is used in the SMC framework, also incorporating optimal control when available. This layered approach balances precision, adaptability, and computational efficiency, ensuring reliability in case of RL or L1 failures.

First Layer

According to Figure 6, whenever the agent's action is set to "COMPUTE", ASRE is used to determine an optimal trajectory and optimal control, following the steps presented in Reference 8. These values are then stored into the onboard computer and used for all the following GNC steps. Being computed at discrete time steps t_k, the optimal trajectory and the optimal control will henceforth be denoted as $\boldsymbol{\xi}(t_k)$ and $\boldsymbol{\chi}(t_k)$, respectively. Intuitively, the "DELETE" action removes the stored $\boldsymbol{\xi}(t_k)$ and $\boldsymbol{\chi}(t_k)$, while the "SKIP" action keeps the configuration from the previous steps.
The TOF required by ASRE was computed according to the pseudo-empirical law reported in Equation (1), where $\boldsymbol{\delta} = \boldsymbol{\rho} - \boldsymbol{\rho}_\star$, $\boldsymbol{\rho}_\star$ is the aimed relative position, x_c a conversion factor, and $\boldsymbol{e}_V$ as

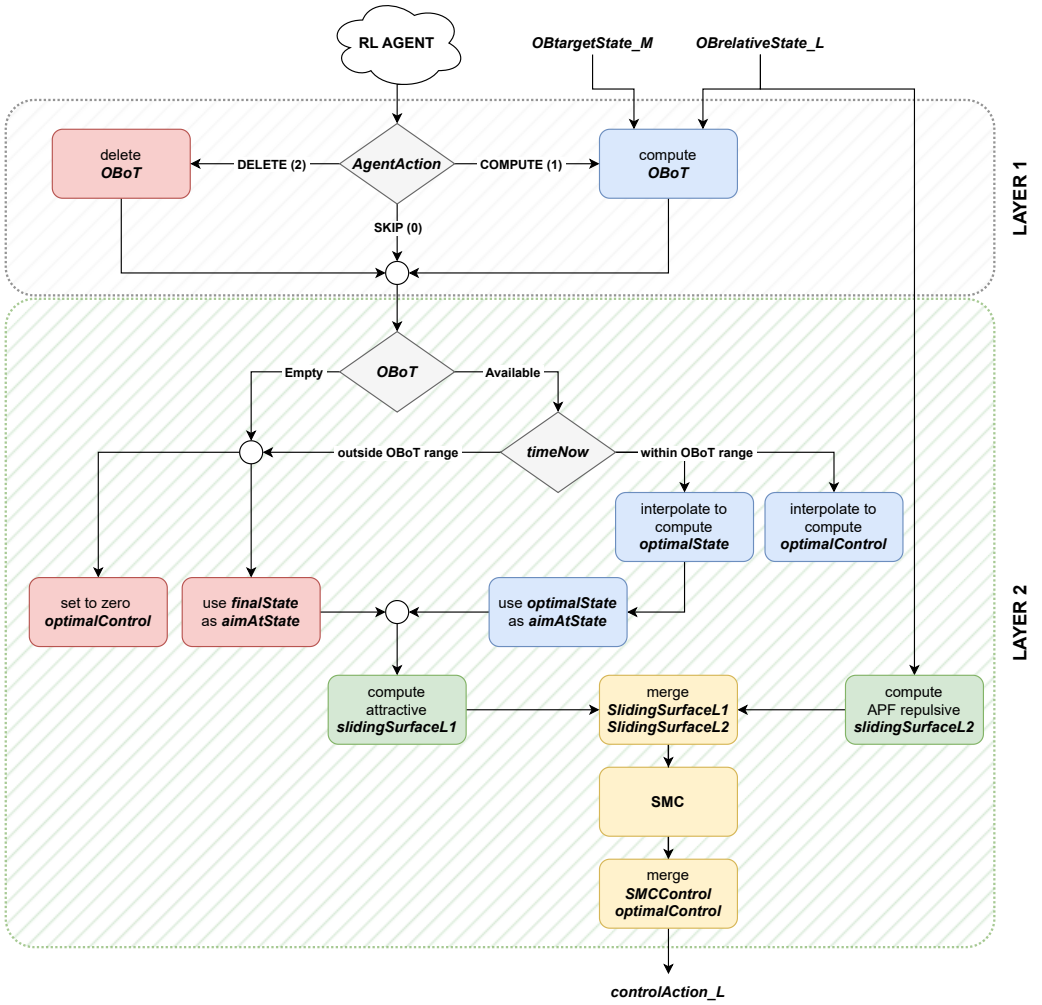


Figure 6: Functioning flowchart diagram: the RL agent triggers the logic in Layer 1 defined in Figure 4.

$$\Delta t = t_f - t_i = \frac{\|\boldsymbol{\delta}\|}{4 \cdot 10^{-4}} \cdot \left[1.1 - \tanh\left(\frac{\|\boldsymbol{\delta}\| \cdot x_c}{5}\right)\right] \cdot \left[2 + \frac{\boldsymbol{\delta} \cdot \boldsymbol{e}_V}{\|\boldsymbol{\delta}\|}\right] \tag{1}$$

Second Layer

Following L1, the second layer starts by defining a sliding surface using the precomputed optimal trajectory (if available) and the APF repulsive field, as:

$$\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}_{L1}(t) + \boldsymbol{\sigma}_{L2}(t) \tag{2}$$

Eventually, the SMC is used to merge the contributions as in Equation (3), similarly to the Optimal Sliding Guidance described in the work of D. R. Wibben and R. Furfaro.[11]

$$\boxed{\boldsymbol{u} = \boldsymbol{u}_{opt} - U_{max} \cdot \tanh(\boldsymbol{\sigma})} \tag{3}$$

When $\boldsymbol{\xi}(t_k)$ and $\boldsymbol{\chi}(t_k)$ are available, it is possible to interpolate them at the current time t to determine the optimal state $\boldsymbol{x}_{opt}(t)$ and optimal control $\boldsymbol{u}_{opt}(t)$ as:

$$\boldsymbol{x}_{opt}(t) = \text{interp}\left(\boldsymbol{\xi}(t_k), t\right) \quad \text{and} \quad \boldsymbol{u}_{opt}(t) = \text{interp}\left(\boldsymbol{\chi}(t_k), t\right)$$

$\boldsymbol{x}_{opt}$ is then used as the "*aimAtState*" $\boldsymbol{x}_{\rho}^{\star}$, and the optimal control $\boldsymbol{u}_{opt}$, is used to augment the SMC output as seen in Equation (3). Once $\boldsymbol{x}_{\rho}^{\star} = [\boldsymbol{\rho}_{\star}, \dot{\boldsymbol{\rho}}_{\star}]^{\top}$ is set, the attractive component of the sliding surface is computed by adding the terms as in Equation (4), where $\boldsymbol{\sigma}_{L1,pos} = (\boldsymbol{\rho} - \boldsymbol{\rho}_{\star})$ and $\boldsymbol{\sigma}_{L1,vel} = (\dot{\boldsymbol{\rho}} - \dot{\boldsymbol{\rho}}_{\star})$, while $\boldsymbol{\kappa}_{L1}^{pos}$ and $\boldsymbol{\kappa}_{L1}^{vel}$ are the gains.

$$\boldsymbol{\sigma}_{L1} = \boldsymbol{\kappa}_{L1}^{vel}\boldsymbol{\sigma}_{L1,vel} + \boldsymbol{\kappa}_{L1}^{pos}\boldsymbol{\sigma}_{L1,pos} \tag{4}$$

Observe that in case the optimal values are unavailable or "expired", the algorithm switches $\boldsymbol{x}_{\rho}^{\star}$ to the final state $\boldsymbol{x}_f$, and $\boldsymbol{u}_{opt}$ is set to zero, effectively leading to a simple APF+SMC. This design guarantees that even if L1 fails, the APF+SMC guidance remains robust and collision-free, thus without compromising safety.

Concerning the second component of Equation (2), it is computed as $\boldsymbol{\sigma}_{L2} = \nabla U_{APF}$, where ∇U_{APF} depends on the mission phase.
During the *Rendezvous Phase* (or phase 1), a simple Safe Sphere (SS) of radius 2 km was used, to ensure a safe margin around the 1 km-radius Approach Sphere, described by Equation (5), where R_{SS} is such that none of the safety constraints is violated.

$$h(\boldsymbol{\rho}) = \boldsymbol{\rho}^{\top}\boldsymbol{\rho} - R_{\text{SS}}^2 \tag{5}$$

The gradient of h is simply $\nabla h(\boldsymbol{\rho}) = 2\boldsymbol{\rho}$. Thus, the repulsive field is obtained via Equation (6), where $\Delta\boldsymbol{\rho} = \boldsymbol{\rho} - \boldsymbol{\rho}_{hold}$, $\boldsymbol{\rho}_{hold}$ is the holding state, and the coefficients used are reported in Table 2. Note that $\Delta\boldsymbol{\rho}$ is used in the definition in order to have a null repulsive field in correspondence to the final state.

$$\nabla U_{APF} = \begin{cases} -\mathbf{K}_{SS}^{in} \cdot \dfrac{\boldsymbol{\rho}}{\|\boldsymbol{\rho}\|}, & \text{if } \|\boldsymbol{\rho}\|^2 \leq R_{\text{SS}}^2, \\ \mathbf{K}_{SS}^{out} \cdot \left(\dfrac{\Delta\boldsymbol{\rho}}{h(\boldsymbol{\rho})^2} - \dfrac{(\Delta\boldsymbol{\rho}^{\top}\Delta\boldsymbol{\rho}) \cdot \nabla h(\boldsymbol{\rho})}{h(\boldsymbol{\rho})^3} \right), & \text{otherwise.} \end{cases} \tag{6}$$

Table 2: APF coefficients, Rendezvous Phase

$\mathbf{K}_{SS}^{in}$	$\mathbf{K}_{SS}^{out}$	R_{SS} [km]	$[\,\boldsymbol{\rho}_{hold}\,]_{\mathcal{L}}$ [km]
$\begin{bmatrix} 2\cdot10^3 & 0 & 0 \\ 0 & 5\cdot10^2 & 0 \\ 0 & 0 & 2\cdot10^3 \end{bmatrix}$	$\begin{bmatrix} 5\cdot10^5 & 0 & 0 \\ 0 & 6\cdot10^6 & 0 \\ 0 & 0 & 5\cdot10^5 \end{bmatrix}$	2	$[0, -4, 0,]^{\top}$

For what concerns the *Final Approach and Docking Phase* (or phase 2), a semi-cubical parabola, outlined in Equation (7), was considered[12] allowing for improved precision in the closing maneuvers of docking and greater margin for far-range procedures.

$$h(\boldsymbol{\rho}) = \rho_R^2 + a_{cone}^2 \left(\rho_V - b_{cone}\right)^3 + \rho_H^2 \tag{7}$$

The repulsive term is computed by means of Equation (8), whose coefficients are those in Table 3.

$$\nabla U_{APF} = \begin{cases} \mathbf{K}_{C}^{out} \cdot \dfrac{\boldsymbol{\rho}}{\|\boldsymbol{\rho}\|}, & \text{if } h(\boldsymbol{\rho}) \geq 0, \\ \mathbf{K}_{C}^{in} \cdot \left(\dfrac{\boldsymbol{\rho}}{h(\boldsymbol{\rho})^2} - \dfrac{(\boldsymbol{\rho}^{\top}\boldsymbol{\rho}) \cdot \nabla h(\boldsymbol{\rho})}{h(\boldsymbol{\rho})^3} \right), & \text{otherwise.} \end{cases} \tag{8}$$

Table 3: APF coefficients, Approach and Docking Phase

$\mathbf{K}_C^{in}$	$\mathbf{K}_C^{out}$	a_{cone}	b_{cone}
$\begin{bmatrix} 5\cdot10^{-3} & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 5\cdot10^{-3} \end{bmatrix} + \frac{\lvert\rho_V\rvert^3}{10^9} \begin{bmatrix} 300 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 300 \end{bmatrix}$	$\begin{bmatrix} 10 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 10 \end{bmatrix}$	0.02	10

Reinforcement Learning Agent

The observation space of the agent consists of the on-board estimated relative state ($\boldsymbol{x}_\rho$) expressed in the LVLH frame, the mean control action of the last 100 GNC steps ($\bar{\boldsymbol{u}}$), and the age of the on-board optimal trajectory (called "*OBoTAge*"), which specifies whether the trajectory is available and quantifies the time elapsed since its last computation. In turn, the agent action space is represented by the discrete "*AgentAction*", which specifies the trigger to: maintain the current state (SKIP or 0), update the optimal trajectory (COMPUTE or 1) or delete it (DELETE or 2).
The default Stable-Baselines3 (SB3) *MlpPolicy* was employed for the implementation of two Multilayer Perceptrons (MLPs), one for the actor and the other for the critic, both featuring two hidden layers of 64 neurons each and using the *tanh* activation function. By adopting an Actor-Critic method like PPO, the agent effectively learns to exploit the guidance configurations.

TRAINING AND SIMULATIONS

Reward Function Definition

The framework employed for training the agent is depicted in Figure 7. In particular, the reward is a fundamental component of RL, serving as the primary mechanism to guide the agent behavior toward achieving its objectives.

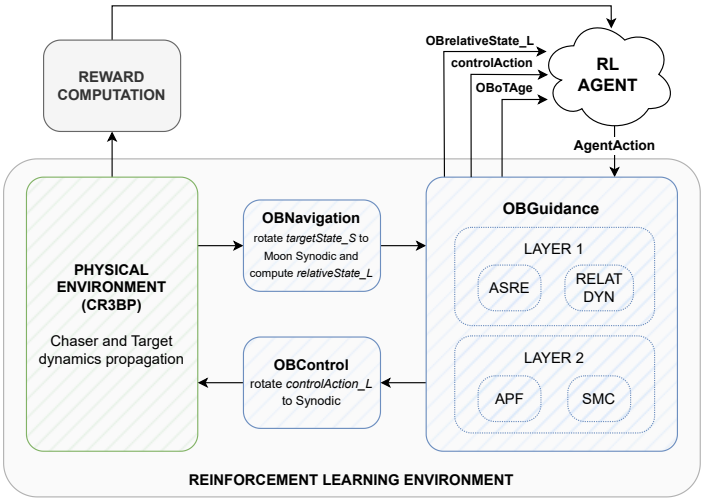


Figure 7: Training Framework

In this specific context, the purpose of the reward function is to encode mission requirements and operational constraints, encouraging the agent to make decisions that balance safety, efficiency, and computational resource management. More precisely, the reward function is designed to: encourage safety and precision (especially for docking), promote actions that minimize fuel consumption, ensure efficient use of onboard resources (penalizing useless behaviors), discourage unsafe maneuvers (such as violating collision avoidance constraints) and reduce the required TOF. To do so, an ad-hoc reward function $\mathcal{R}(s, a)$ was designed for both phases. Indeed, since two constraints were used for different operational scenarios, two agents had to be trained and tuned to different requirements.

The general shape of the reward function can be written in general form:

$$\mathcal{R}(s,a) = \mathcal{R}_{\text{action}} + \mathcal{R}_{\text{precision}} + \mathcal{R}_{\text{simtime}} + \mathcal{R}_{\text{control}} + \mathcal{R}_{\text{end}} \tag{9}$$

where most terms are bounded and within the range $[-1,+1]$. Moreover, the SB3's built-in *VecNormalize* method was used to have a faster and more stable training, with better gradient updates and reduced instability in policy updates.

Simulations

Finally, to evaluate the performance and robustness of the proposed strategy, different testing scenarios were defined, considering the spacecraft data in Table 1. These aim to assess the guidance capabilities under varying conditions: the objective is to verify that the agent successfully learns and converges to an effective guidance strategy, exploiting the available configurations to reach the final goal. In this context, the safe mode scenario specifically serves as a baseline to prove the fail-safety of the system and to show that the proposed strategy can indeed enhance it. Moreover, the aim is to verify that, if a failure was detected and assuming that a FDIR system was in place, the backup APF+SMC guidance can maintain operational integrity. However, since no full FDIR is implemented, this test does not constitute a full validation of the safety itself.

Generation of the Initial Conditions

Four different regions were identified along the target's orbit: aposelene, approaching aposelene, leaving aposelene, and periselene, see Figure 8. Generally, rendezvous operations are preferably conducted in the aposelene region due to favorable dynamics.[10] However, all four regions were considered for testing purposes. A progressive validation of the algorithm was obtained by starting from basic fail-safe behavior up to advanced RL-guided maneuvers in increasingly dynamic environments. Thus, for each region and mission phase, results are evaluated under both safe-mode and nominal conditions. To simulate the two phases, different methods were adopted to generate the initial conditions, tailored to the constraints, and a fixed seed was used to ensure reproducibility and fair comparison across all tests.

In Phase 1, positions were randomly sampled in spherical coordinates within the annular region $[2,9]\,\mathrm{km}$ and then converted to Cartesian. Relative velocities ranged from -5 to $+5\,\mathrm{m/s}$.

In Phase 2, initial positions were limited to a cone-shaped volume around the target, ensuring compliance with the radius constraint. Velocities were limited to $[-2,+2]\,\mathrm{m/s}$.

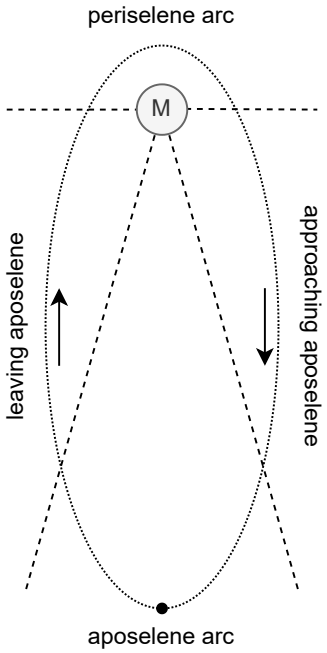


Figure 8: Target's orbit highlighting the identified regions for defining mission scenarios

Performance Metrics

Once the set of n_{pop} initial conditions, namely the population, was generated, the performance could be evaluated using several key parameters. These are:

- **Final Position and Velocity Statistics**: standard deviation and mean of the final spacecraft position and velocity were computed for $\boldsymbol{e}_R$, $\boldsymbol{e}_V$, and $\boldsymbol{e}_H$ axes.
- **Mean Control Action**: simply computed from the overall control history, integrating $\boldsymbol{u}(t)$ over time for each simulation and averaging across the population.
- **Mean Execution Time**: represents the mean elapsed time required for a Guidance Navigation and Control (GNC) step to be completed. Since the simulations were run on a high-performance computer, the execution time was then converted to an equivalent time, considering the specifications of a GR740 quad-core fault-tolerant LEON4FT SPARC V8 processor, via the simple proportion:

$$\bar{t}_{\text{GNC}}^{\text{exec}}\big|_{eq} = \bar{t}_{\text{GNC}}^{\text{exec}} \cdot \frac{2.65\,\text{GHz}}{250\,\text{MHz}} \cdot \frac{24\,\text{cores}}{4\,\text{cores}} \tag{10}$$

- **Mean Cost**: the total maneuver cost expressed in terms of propellant mass used Δm, or by using the total change in velocity ΔV to generalise:

$$\Delta m = \frac{1}{n_{pop}} \sum_{i=1}^{n_{pop}} \sum_{k=0}^{n_{sim}(i)} \frac{\|\mathbf{u}_i(t_k)\| \cdot dt}{I_{\text{sp}} \cdot g_0} \tag{11}$$

where $dt = f_{GNC}^{-1}$ is the GNC time step, and all the terms are converted to SI units. In general, notice that in Equation (11) the length of an episode n_{sim} varies for each simulation.

Simulation Stopping Criteria

The phase-specific termination conditions used are reported in Table 4. These serve as a stopping condition for the simulations and allow to infer the actual performances of the guidance scheme.

Table 4: Termination Conditions for the Two Phases, compatible with docking standards[13]

Phase ID	Outcome	Termination Condition
1	Success	Relative position error: $\|\boldsymbol{\rho} - \boldsymbol{\rho}_f\| < 200\,\text{m}$ Relative velocity error: $\|\dot{\boldsymbol{\rho}} - \dot{\boldsymbol{\rho}}_f\| < 0.5\,\text{m/s}$
	Fail	In this phase, a crash is highly improbable as the chaser remains distant from the target. In any case, if $\|\boldsymbol{\rho}\| < 200\,\text{m}$ the algorithm stops since the Keep-Out Sphere (KOS) is violated.
2	Success	Convergence on $\boldsymbol{e}_V$: $-5\,\text{mm} \leq \rho_V - \rho_{V_f} \leq 0\,\text{mm}$ Convergence on $\boldsymbol{e}_R$ and $\boldsymbol{e}_H$: $\sqrt{\rho_R^2 + \rho_H^2} \leq 10\,\text{cm}$ Docking velocity: $\|\dot{\rho}_R\| \leq 0.04\,\text{m/s}$ and $\|\dot{\boldsymbol{\rho}}_H\| \leq 0.04\,\text{m/s}$ while $\|\dot{\rho}_V\| \leq 0.1\,\text{m/s}$
	Fail	If the previous conditions are not met, the docking attempt fails (crash). Also if the chaser moves in front of the target without having satisfied the position and velocity constraints the outcome is a crash.

TESTING SCENARIO: RENDEZVOUS PHASE

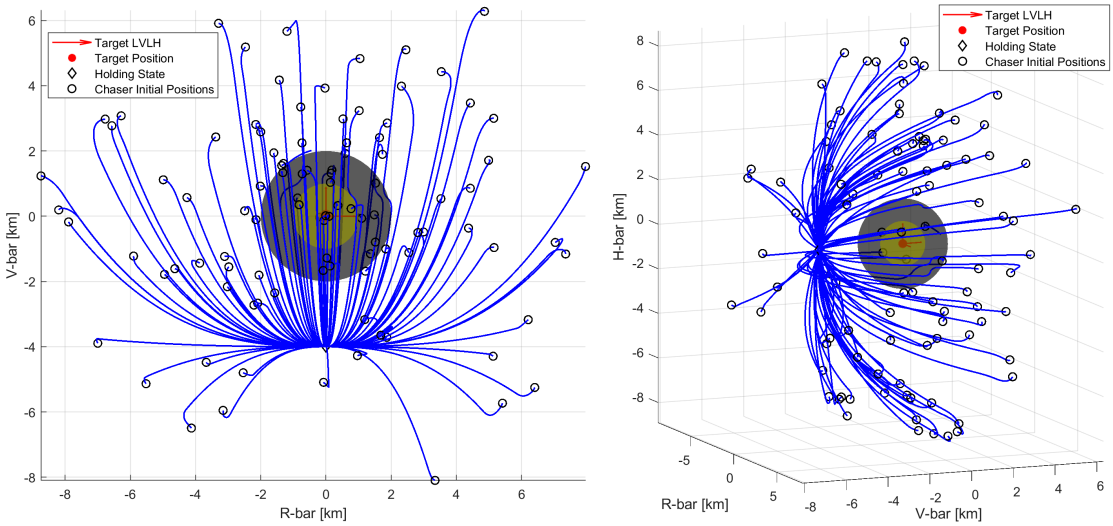
The first phase to be analyzed is the rendezvous to a safe holding state. 100 initial conditions were generated for both the safe and nominal modes, with a initial relative distance of 6.54 ± 1.80 [km] and velocity of 4.54 ± 1.30 [m/s].

Aposelene Region

The first region to be studied is the aposelene one, where Figure 9 reports the 100 trajectories obtained in the safe mode scenario. As shown in Table 5, both modes converge to a satisfactory result, with the nominal mode presenting higher fuel efficiency and a lower TOF. Indeed, the total propellant mass used drops from 2.63 ± 0.73 [kg] in safe mode, to 0.70 ± 0.54 [kg] in nominal mode, proving that ASRE indeed optimizes the trajectory, and almost halves the TOF. This result is expected, proving that the agent learns to efficiently exploit the available guidance scheme to maximize its reward. It was also observed that 1% of the safe mode simulations did not converge in time. This was due to the spacecraft getting stuck in a local minimum of the APF and not due to a constraint violation nor a collision. In any case, this situation is solved when the APF+SMC guidance is augmented with ASRE and the RL agent, proving that the agent avoids local minimums. The equivalent execution time is fully compatible with the 1 Hz GNC requirement for both modes, being in the order of tens of milliseconds. This result is ascribable to the average re-computations of the first layer, being 1.71 per episode. Concerning the final state precision the effectiveness of the method was confirmed. A low precision underlined by this metric is partially due to the stopping criterion used for this phase, defined as in Table 4.

Table 5: Phase 1, aposelene scenario: overall performances for 100 simulations

Phase 1 Mode	ΔV [m/s]	$\bar{t}^{\text{exec}}_{\text{GNC}}\big\vert_{eq}$ [ms]	TOF [min]	Success
Safe Mode	13.16 ± 3.63	10.77 ± 1.85	211.83 ± 14.38	99%
Nominal Mode	3.48 ± 2.69	22.74 ± 63.31	118.88 ± 46.20	100%

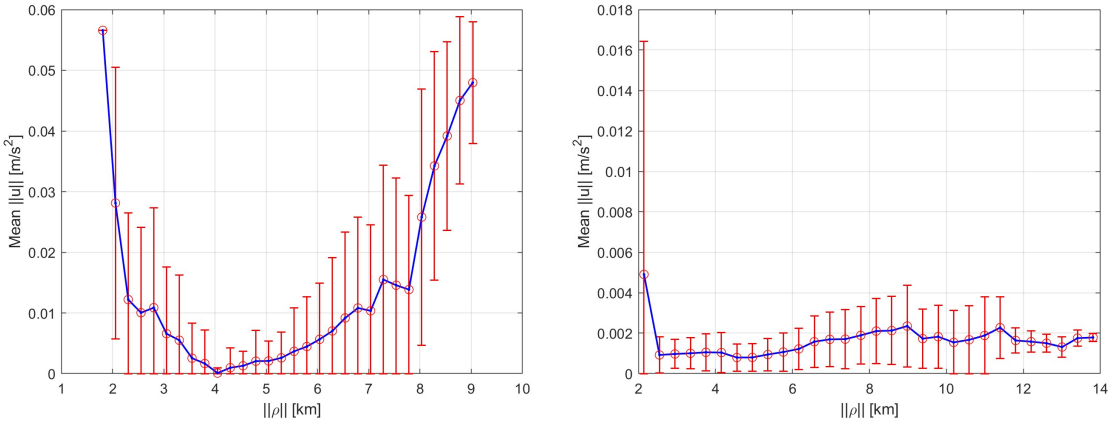


(a) Top view: V-R plane

(b) 3D View

Figure 9: Safe Mode trajectories for rendezvous phase

Another interesting result can be observed in Figure 10. This image shows the average control action computed in 30 "bins" of $\|\boldsymbol{\rho}\|$. In particular, the nominal configuration allows to abate the control action required not only globally, but also locally. This is due to the fact that the attractive field is generated from ASRE, hence the sliding surface smoothly guides the spacecraft to the final state. It is also interesting to notice that in Figure 10a the minimum is reached for $\|\boldsymbol{\rho}\| = 4\,\mathrm{km}$, which is where the holding state is. Moreover, the same image shows a violation of the soft constraint region (i.e. $\|\boldsymbol{\rho}\| \leq 2\,\mathrm{km}$), leading to the APF enforcing maximum thrust away from the target. This condition was considered by design and indeed used as a buffer, since the constraint was defined taking into account a margin of 100% with respect to the true AS.



(a) Safe mode scenario

(b) Nominal scenario

Figure 10: Phase 1: mean control action and standard deviation as a function of the relative position

Approaching Aposelene Region

Similar results were obtained for the approaching aposelene region. However, a very slight increase in the overall maneuver cost was observed, as reported in Table 6.

Table 6: Phase 1, approaching aposelene scenario: overall performances for 100 simulations

Phase 1	ΔV [m/s]	$\bar{t}_{\mathrm{GNC}}^{\mathrm{exec}}\big\vert_{eq}$ [ms]	TOF [min]	Success
Safe Mode	13.01 ± 4.15	14.21 ± 3.58	211.66 ± 14.81	100%
Nominal Mode	3.51 ± 2.74	25.02 ± 80.61	118.73 ± 45.72	100%

Leaving Aposelene Region

As for the previous two phases also when leaving the aposelene the guidance performances were satisfactory, as reported in Table 7. In particular, the higher propellant usage can be attributed to the faster dynamics of the environment.

Table 7: Phase 1, leaving aposelene scenario: overall performances for 100 simulations

Phase 1	ΔV [m/s]	$\bar{t}_{\mathrm{GNC}}^{\mathrm{exec}}\big\vert_{eq}$ [ms]	TOF [min]	Success
Safe Mode	13.01 ± 4.15	14.36 ± 3.58	211.65 ± 14.80	100%
Nominal Mode	3.50 ± 2.74	24.97 ± 79.48	119.19 ± 46.69	100%

Periselene Region

Lastly, the periselene region was considered. The values in Table 8 show an increase in propellant consumption of up to around $2\,\mathrm{m/s}$ and a general loss in performance.
It is worth pointing out that for this specific scenario the agent had an average number of L1 re-computations of around 5.48 per episode, in contrast to the previous ones of $[1.71, 1.77]$. This underlines the ability of the agent to adapt to the fast dynamics of the environment, even though with a degraded performance. Also the higher GNC execution time for nominal mode suggests that the L1 optimization problem becomes stiffer for the ODE solver.

Table 8: Phase 1, periselene region scenario: overall performances for 100 simulations

Phase 1	ΔV [m/s]	$\bar{t}^{\text{exec}}_{\text{GNC}}\big\vert_{eq}$ [ms]	TOF [min]	Success
Safe Mode	13.35 ± 4.17	13.81 ± 2.82	211.76 ± 14.68	100%
Nominal Mode	5.40 ± 3.43	28.09 ± 337.10	122.40 ± 50.09	100%

TESTING SCENARIO: APPROACH AND DOCKING PHASE

In this section, the approach and docking scenario is presented. As previously done, 100 random initial states for the chaser were generated, with a mean initial distance of 3.03 ± 0.57 [km] and a mean initial velocity of 1.95 ± 0.58 [m/s].

Aposelene Region

As before, the first region to be analyzed is the aposelene and the results are reported in Table 9. The algorithm required a mean control cost of around 1.82 ± 0.46 [kg] for safe mode, compared to 0.42 ± 0.50 [kg] for nominal operations. On the downside, this jeopardized precision, slightly degrading performance, as shown in Table 10 and graphically in Figure 11. Possibly, the cause lies in the suboptimal tuning of the attractive surface gains, combined with the reward function insufficient emphasis on precision during docking. This issue is predicted to be attenuated by properly tuning

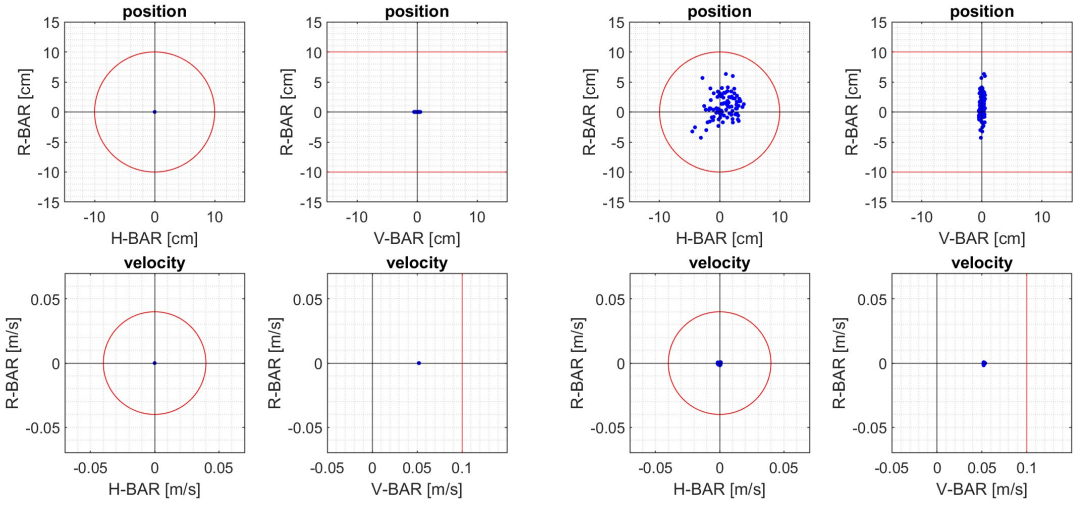
Table 9: Phase 2, aposelene scenario: overall performances for 100 simulations

Phase 2 Mode	ΔV [m/s]	$\bar{t}^{\text{exec}}_{\text{GNC}}\big\vert_{eq}$ [ms]	TOF [min]	Success
Safe Mode	9.08 ± 2.29	12.33 ± 0.79	167.53 ± 11.27	100%
Nominal Mode	2.10 ± 2.50	22.54 ± 165.82	136.48 ± 17.16	100%

Table 10: Approach and Docking results: mean and standard deviation for the reached final state.

Component	**Units**	**Safe Mode**	**Nominal Mode**
ρ_R	[cm]	0.00 ± 0.00	0.94 ± 1.95
ρ_V	[cm]	0.00 ± 0.31	0.03 ± 0.27
ρ_H	[cm]	0.01 ± 0.00	0.71 ± 1.79
$\dot{\rho}_R$	[cm/s]	0.00 ± 0.00	-0.02 ± 0.05
$\dot{\rho}_V$	[cm/s]	5.21 ± 0.01	5.21 ± 0.03
$\dot{\rho}_H$	[cm/s]	0.00 ± 0.00	-0.02 ± 0.05

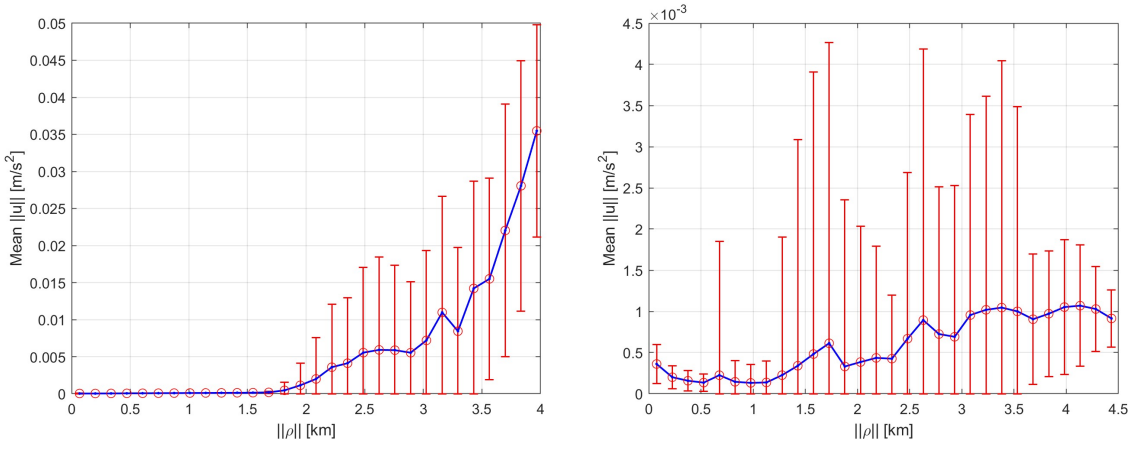
the reward or by manually enforcing a safe mode for the final phase. Moreover, it was noticed that by slightly increasing the APF $\boldsymbol{\kappa}_{L1}^{pos}$ values, it is possible to balance velocity and position precision at the expense of collision avoidance capabilities.
In Figure 12 the control action was correlated with the norm of the relative distance. Specifically, for the second mission phase, the docking port (i.e. the final state) is located at $\|\boldsymbol{\rho}\| = 0\,\mathrm{km}$. Furthermore, notice that the chaser initial positions are in the range of approximately $[1.7, 4.5]$ km. This zone is indeed characterized by a higher control action for the safe mode, due to the fact that the algorithm is prone to reducing the docking error from the early stages, as demonstrated in Figure 13. This feature leads to the main advantage of safe mode: higher precision and collision avoidance capability. Moreover, Figure 12 emphasizes how the augmented guidance scheme distributes the control action along the path, while reducing the overall propellant consumption.



(a) Safe mode: docking precision

(b) Nominal mode: docking precision

Figure 11: Phase 2: Docking precision for nominal and safe mode scenarios. No outliers were identified. Red lines are the docking constraints, according to Table 4.



(a) Safe mode scenario

(b) Nominal scenario

Figure 12: Phase 2: mean control action and standard deviation as a function of the relative position

This behavior is expected, since the attractive field gently drives the spacecraft to the final state, considering an optimal path. On the other hand, the safe mode forces the spacecraft to reach and eventually slide along a fixed sliding surface, as shown in fig. 13, which may be suboptimal and surely imposes a higher control action. At this stage, it is worth remarking that ASRE does not consider path constraints. Hence, the trajectory is optimal as long as it does not violate (and is far enough from) any constraint. Indeed, if a limit is exceeded, the APF strongly degrades the performances by imposing the maximum thrust available to bring the spacecraft back to a safe configuration. However, this last condition should be studied more in detail, introducing a FDIR strategy to make the algorithm more robust.

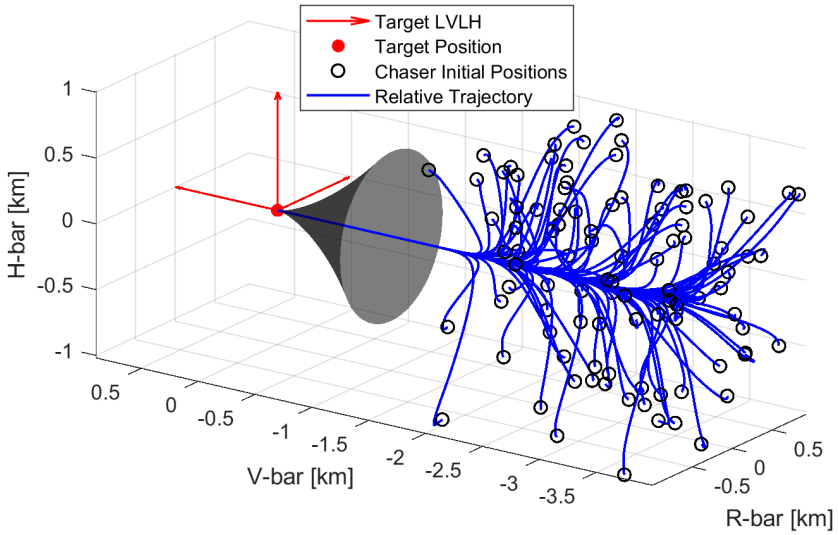


Figure 13: Safe Mode trajectories for approach and docking phase

Approaching Aposelene Region

Following the same steps as before, the approaching aposelene region was considered. As expected, the results show a slight degradation of the overall performances with respect to the previous scenario, due to the higher non-linearities. Still, the augmented guidance scheme allows for a reduction in propellant usage and TOF, at the expense of lower docking precision, Figure 14a.

Table 11: Phase 2, approaching aposelene scenario: overall performances for 100 simulations

Phase 2	ΔV [m/s]	$\bar{t}_{\text{GNC}}^{\text{exec}}\big\|_{eq}$ [ms]	TOF [min]	Success
Safe Mode	9.08 ± 2.29	13.21 ± 3.39	167.53 ± 11.27	100%
Nominal Mode	5.96 ± 2.56	23.14 ± 240.11	147.32 ± 20.51	100%

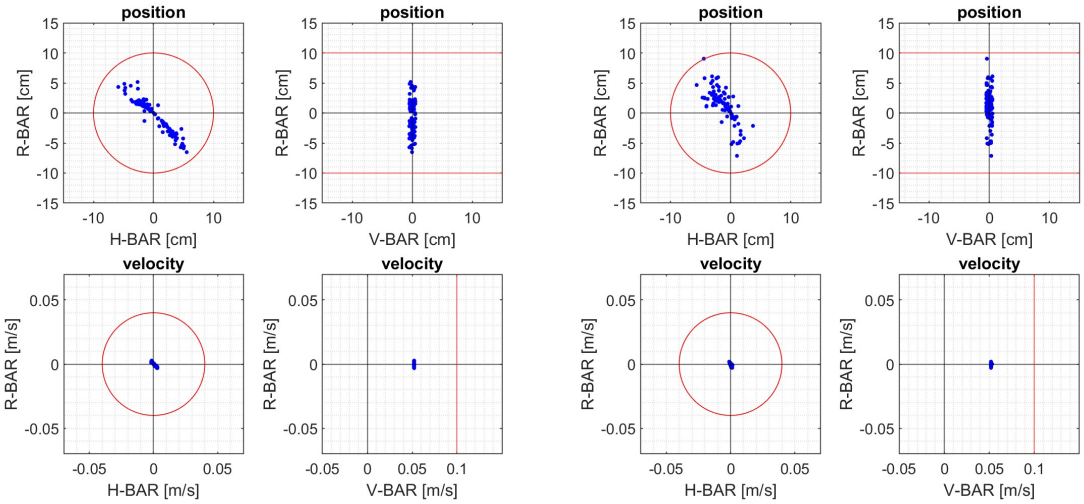
Leaving Aposelene Region

Similar results are obtained in the leaving aposelene region, shown in Table 12. In this scenario 1% of the trajectories violated the docking condition. This can be seen in Figure 14b, where one of the 100 points slightly violates the position constraint. This is due to the fact that as the chaser approaches the target, the environment dynamics becomes faster, and thus more critical. As seen for the first phase, the simulations highlight the ability of the agent to properly select its action, in

order to reduce the propellant usage and TOF. However, at the same time a further degradation in precision was observed due to the higher non-linearities, as reported in Figure 14a.

Table 12: Phase 2, leaving aposelene scenario: overall performances for 100 simulations

Phase 2	ΔV [m/s]	$\bar{t}^{\text{exec}}_{\text{GNC}}\big\vert_{eq}$ [ms]	TOF [min]	Success
Safe Mode	9.08 ± 2.29	13.54 ± 3.58	167.53 ± 11.27	100%
Nominal Mode	3.22 ± 2.33	22.79 ± 192.16	144.92 ± 22.23	99%



(a) Approaching aposelene region

(b) Leaving aposelene region

Figure 14: Phase 2: Docking precision for the approaching aposelene and leaving aposelene scenarios, both in nominal mode. Red lines represent the docking constraints, according to Table 4.

Periselene Region

To conclude the analysis, the periselene arc was considered. Unfortunately, while for safe mode the success rate was more than satisfactory, the nominal mode demonstrated unacceptable results.

Table 13: Phase 2, periselene region scenario: overall performances for 100 simulations

Phase 2	ΔV [m/s]	$\bar{t}^{\text{exec}}_{\text{GNC}}\big\vert_{eq}$ [ms]	TOF [min]	Success
Safe Mode	10.55 ± 2.49	12.93 ± 2.52	137.83 ± 8.61	100%
Nominal Mode	n/a	n/a	n/a	2%

It is essential to point out that, as the chaser approaches the faster dynamics region, i.e. the periselene arc, the setup for ASRE appears less suitable for optimizing the trajectory. More precisely, the equation for the computation of the TOF, Equation (1), does not consider the change in dynamics. Indeed, in this region, one would expect a faster docking, hence a smaller TOF. This can also be proven by comparing the obtained TOF for the safe modes in Table 13 with the ones in Table 9. Therefore, after a subsequent review of Equation (1), i.e. by reducing the TOF, the simulations were repeated and the success rate increased to 57%. This is still not satisfactory and clearly demonstrates that the inclusion of the mission constraints directly inside the first guidance layer would improve the performance of the algorithm in this region and enhance its robustness and reliability.

CONCLUSIONS

On the whole, the augmented guidance demonstrated good performances in both terms of propellant consumption and reduced TOF. Moreover, the latter can be further reduced by enforcing a shorter TOF into ASRE optimization algorithm, at the expense of a higher propellant consumption. The equivalent execution times were satisfactory for most of the scenarios, with some exceptions for the periselene region, due to a higher variance. Indeed, this region very likely produces a stiffer problem to be solved inside ASRE, occasionally leading to longer execution time. Furthermore, even though docking precision was reduced, it is predicted that by either enforcing constraints inside L1 or by increasing the influence of $\boldsymbol{\sigma}_{L1}$ inside $\boldsymbol{\sigma}$, it is possible to obtain a better performance. This work investigated the integration of RL into advanced relative guidance frameworks to enhance autonomy, adaptability, and efficiency in spacecraft proximity operations, particularly in NRHO. The proposed hybrid framework combines ASRE for trajectory planning, APF for collision avoidance, and SMC for robustness, with an RL agent acting as a high-level decision-maker rather than a direct controller. This structure allows the system to optimize performance (e.g., propellant usage, TOF) while ensuring safety and interpretability. This demonstrates the potential of RL to augment classical guidance systems, promoting safer and more autonomous space operations without sacrificing reliability or transparency.

Simulations within the CR3BP validated the framework, highlighting strong performance in rendezvous tasks, especially in the dynamically favorable aposelene region. The RL-aided system outperformed traditional APF+SMC, effectively handling constraints and avoiding local minima, while remaining reliable thanks to a fallback safe mode that bypasses the RL component if needed. The agent showed adaptability across mission phases and proved capable of enhancing resource management, despite minor precision trade-offs. These results indicate that RL can safely complement classical guidance systems, paving the way for more autonomous and efficient space missions.

REFERENCES

[1] E. Capello, F. Dabbene, G. Guglieri, and E. Punta, ""Flyable" Guidance and Control Algorithms for Orbital Rendezvous Maneuver," *SICE Journal of Control, Measurement, and System Integration*, Vol. 11, Jan. 2018, pp. 14–24, 10.9746/jcmsi.11.14.

[2] R. Chai, A. Tsourdos, A. Savvaris, S. Chai, Y. Xia, and C. L. Philip Chen, "Review of advanced guidance and control algorithms for space/aerospace vehicles," *Progress in Aerospace Sciences*, Vol. 122, Apr. 2021, p. 100696, 10.1016/j.paerosci.2021.100696.

[3] Vincenzo Pesce, Andrea Colagrossi, and Stefano Silvestrini, *Modern Spacecraft Guidance, Navigation, and Control*. Elsevier, Nov. 2022.

[4] M. Galullo, G. Bucchioni, G. Franzini, and M. Innocenti, "Closed Loop Guidance During Close Range Rendezvous in a Three Body Problem," *The Journal of the Astronautical Sciences*, Vol. 69, Feb. 2022, pp. 28–50, 10.1007/s40295-021-00289-6.

[5] G. Fereoli, "Meta-reinforcement learning for spacecraft proximity operations guidance and control in Cislunar Space," Master's thesis, Politecnico di Milano, Dec. 2023. Accepted: 2024-03-08T10:34:12Z.

[6] S. Lizy-Destrez, B. Le Bihan, A. Campolo, and S. Manglativi, "Safety Analysis for Near Rectilinear Orbit Close Approach Rendezvous in the Circular Restricted Three-Body Problem," *The 68th annual International Astronautical Congress (IAC 2017)*, Adelaide, Australia, Sept. 2017.

[7] E. M Zimovan Spreen, *Trajectory design and targeting for applications to the exploration program in cislunar space*. PhD thesis, Purdue University, 2021.

[8] F. Topputo, M. Miani, and F. Bernelli-Zazzera, "Optimal Selection of the Coefficient Matrix in State-Dependent Control Methods," *Journal of Guidance, Control and Dynamics*, Vol. 38, No. 5, 2015, pp. 861–873, 10.2514/1.G000136.

[9] G. Franzini and M. Innocenti, "Relative Motion Dynamics in the Restricted Three-Body Problem," *Journal of Spacecraft and Rockets*, Vol. 56, Apr. 2019, pp. 1–16, 10.2514/1.A34390.

[10] L. Bucci, M. Lavagna, and F. Renk, "Relative dynamics analysis and rendezvous techniques for Lunar Near Rectilinear Halo Orbits," *68th International Astronautical Congress (IAC 2017)*, 2017.

[11] D. R. Wibben and R. Furfaro, "Optimal sliding guidance algorithm for Mars powered descent phase," *Advances in Space Research*, Vol. 57, Feb. 2016, pp. 948–961, 10.1016/j.asr.2015.12.006.

[12] L. Giorcelli, "Autonomous rendezvous and docking in highly elliptical orbits with Artificial Potential Functions and Sliding Mode Control," Master's thesis, Politecnico di Milano, Oct. 2023.

[13] N. Jody Minor, "International Deep Space Interoperability Standards," 2022.

ACRONYMS

APF	Artificial Potential Field
AS	Approach Sphere
ASRE	Approximating Sequence of Riccati Equations
CR3BP	Circular Restricted Three-Body Problem
FDIR	Fault Detection, Isolation, and Recovery
GNC	Guidance Navigation and Control
KOS	Keep-Out Sphere
LEO	Low Earth Orbit
LVLH	Local Vertical Local Horizontal
ML	Machine Learning
MLP	Multilayer Perceptron
MPC	Model Predictive Control
NRHO	Near-Rectilinear Halo Orbit
ODE	Ordinary Differential Equation
PN	Proportional Navigation
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
SB3	Stable-Baselines3
SDRE	State-Dependent Riccati Equation
SI	International System of Units
SMC	Sliding Mode Control
SS	Safe Sphere
TOF	Time Of Flight
ZEM/ZEV	Zero-Effort-Miss / Zero-Effort-Velocity