# A 6-DOF Deep Reinforcement Learning Architecture for Autonomous Debris Capture and Orbit Transfer

1st Federico Falconi
*Dept. of Computer, Control and Management Engineering*
*University of Rome "La Sapienza"*
Rome, Italy
falconi.2223679@studenti.uniroma1.it

2nd Kirolos Romany Anwar Kamel
*Dept. of Computer, Control and Management Engineering*
*University of Rome "La Sapienza"*
Rome, Italy
kamel@diag.uniroma1.it

3rd Danilo Menegatti
*Dept. of Computer, Control and Management Engineering*
*University of Rome "La Sapienza"*
Rome, Italy
menegatti@diag.uniroma1.it

4th Francesco Delli Priscoli
*Dept. of Computer, Control and Management Engineering*
*University of Rome "La Sapienza"*
Rome, Italy
dellipriscoli@diag.uniroma1.it

***Abstract*—While Deep Reinforcement Learning (DRL) offers robust real-time control for Active Debris Removal (ADR), current literature often assumes idealized dynamics and isolated mission segments. This paper presents an end-to-end 6 Degrees of Freedom (6-DOF) Guidance, Navigation, and Control (GNC) architecture using Proximal Policy Optimization (PPO) for a complete four-phase ADR mission. Bypassing standard linear approximations, the custom environment utilizes an absolute Keplerian N-Body Earth-Centered Inertial (ECI) propagator. It enforces operational realism through continuous Tsiolkovsky propellant dynamics, unmitigated kinematic handovers between phases, and instantaneous Huygens-Steiner inertia updates upon target capture. Trained via Curriculum Learning and astrodynamic energy rewards, the agent successfully executes fuel-optimized orbital transfers, performs dynamic obstacle avoidance, stabilizes proximity operations despite handover drift, and autonomously adapts to the severely degraded mass dynamics of the composite orbital vehicle.**

***Index Terms*—Active Debris Removal, Deep Reinforcement Learning, Proximal Policy Optimization, 6-DOF GNC, Keplerian propagator, Curriculum Learning, Obstacle Avoidance, astrodynamics.**

## I. Introduction

The exponential growth of space activities has led to severe congestion in Low Earth Orbit (LEO), bringing the orbital environment perilously close to the collisional cascading effect known as the Kessler Syndrome [1]. ESA's DISCOS database currently tracks tens of thousands of objects larger than 10 cm, alongside millions of smaller lethal fragments traveling at hypervelocity [2]. Space agencies are urgently transitioning to active remediation through Active Debris Removal (ADR) missions. This paradigm shift is exemplified by the ESA-commissioned ClearSpace-1 [3], a landmark mission tasked with the autonomous capture and de-orbiting of the uncooperative VESPA payload adapter from a 700 km altitude. Unlike cooperative docking, ADR targets lack telemetry and navigational aids. Although modeled here as a passive object, this complete lack of cooperation makes the physical approach inherently hazardous.

The Rendezvous and Proximity Operations (RPO) phase demands Guidance, Navigation, and Control (GNC) systems capable of real-time decision-making under strict collision avoidance and pointing constraints [4]. Classical approaches pair Hill-Clohessy-Wiltshire (HCW) linearized relative dynamics with Model Predictive Control (MPC) [5], [14]. While space-proven for cooperative docking, Non-Linear MPC (NMPC) relies on solving intensive non-linear optimization problems online. For 6-DOF coupled rotational-translational dynamics with obstacle avoidance, the computational footprint of NMPC requires optimization times ranging from 50 to over 200 milliseconds per step on terrestrial CPUs [15], scaling up dangerously when ported to standard radiation-hardened spaceborne microprocessors [16].

Deep Reinforcement Learning (DRL) shifts this massive optimization burden entirely offline. Once a neural network converges, onboard inference is reduced to a deterministic forward pass. To quantitatively validate the feasibility of this approach, the inference time of the trained PPO actor network was profiled: on an Apple M2 processor, the policy requires an average of 0.18 milliseconds per step. This two-orders-of-magnitude reduction compared to typical NMPC solvers makes DRL highly suitable for current space-grade hardware. Foundational works have established Proximal Policy Optimization (PPO) as the reference algorithm for spacecraft proximity operations [6], effectively handling reward shaping for fuel efficiency and line-of-sight compliance [7]. Furthermore,

This research was supported by the Department of Computer, Control and Management Engineering (DIAG), Sapienza University of Rome.

DRL agents naturally exhibit robust generalization against environmental uncertainties, often outperforming deterministic MPC architectures in highly uncooperative scenarios [8].

Yet, a significant technological gap persists concerning the Sim-to-Real translation. Most DRL models for spaceflight conventionally treat mission phases as isolated environments, artificially re-initializing the agent in a perfectly stable hover at each task's start, and almost exclusively assume constant mass [9]. In reality, continuous missions imply compounding kinematic drift, and grasping an uncooperative target instantaneously alters the center of gravity and inertia tensor. To explicitly move beyond the concepts of basic PPO control and reward shaping, this paper quantitatively isolates its novel contributions through a continuous, unmitigated mission formulation:

1) **Continuous Multi-Phase Formulation:** The elimination of artificial state resets between mission segments. The agent inherits residual velocities directly from the preceding phase, proving its capacity to stabilize a $+0.5\,\frac{\mathrm{m}}{\mathrm{s}}$ unmitigated handover perturbation.
2) **Instantaneous Inertia Adaptation:** The dynamic modification of the inertia tensor via the Huygens-Steiner theorem. We validate the policy's ability to seamlessly manage a $+400\,\mathrm{kg{\cdot}m^2}$ inertia surge upon physical capture without explicit control law retuning.
3) **Astrodynamic-Informed De-orbiting:** The resolution of Cartesian "ghost-target" reward stagnation by substituting standard Euclidean distance tracking with an orbital mechanics-based energy metric.
4) **Absolute N-Body Dynamics:** The replacement of linearized HCW proximity approximations with an absolute ECI Keplerian propagator, enforcing true spaceflight non-linearities and continuous Tsiolkovsky mass depletion.

## II. System Model

### A. Absolute Keplerian N-Body Propagator

To capture spaceflight non-linearities, the physics engine operates in the absolute ECI frame rather than on HCW approximations. The translational motion of the spacecraft is governed by the restricted two-body problem:

$$\ddot{\vec{r}} + \frac{\mu}{\|\vec{r}\|^3}\,\vec{r} = \frac{\vec{F}_{ctrl}}{m}, \tag{1}$$

where $\mu = 3.986 \times 10^{14}\,\frac{\mathrm{m}^3}{\mathrm{s}^2}$, $\vec{r}$ is the 3D position vector in ECI, and $\vec{F}_{ctrl}$ the active control force. During coasting ($\vec{F}_{ctrl} = \mathbf{0}$), specific angular momentum and mechanical energy are conserved, naturally confining all bodies to Keplerian ellipses.

A critical challenge is the dimensional mismatch: ECI radial distances exceed 7,000 km while the proximity workspace spans $\sim$10 m. A Direction Cosine Matrix (DCM) $Q_{LVLH}^{ECI}$, recomputed each timestep from the target's absolute state, resolves this with a dual-frame architecture: a macro-scale N-Body engine propagates all bodies in ECI; the DCM rotates the PPO agent's LVLH thrust commands into ECI before integration, then reprojects the resulting state into normalized LVLH observations. Coriolis forces, gravity gradient drift, and orbital precession emerge organically without any explicit programming.

### B. 6-DOF Attitude Kinematics and Coupled GNC

The spacecraft is modeled as a rigid body. Rotational dynamics follow Euler's equations:

$$\dot{\vec{\omega}} = I^{-1}[\vec{\tau}_{body} - \vec{\omega} \times (I\vec{\omega})]\,, \tag{2}$$

where $I$ is the diagonal inertia tensor and $\vec{\tau}_{body}$ the applied torques. Attitude is tracked with unit quaternions [10] to avoid gimbal lock:

$$\dot{q} = \tfrac{1}{2}\,\Omega(\vec{\omega})\,q. \tag{3}$$

A coupled GNC architecture forces the single PPO agent to manage both orbital translation and attitude simultaneously. Observations are expressed in the body frame (simulating onboard LIDAR and star tracker outputs), yielding a 16-dimensional telemetry vector: body-relative positions and velocities to target and obstacle, angular rates $\vec{\omega}$, and quaternion $q$.

### C. Time-Variant Mass and Tsiolkovsky Integration

Thrust is intrinsically linked to propellant consumption at every timestep:

$$dm = \frac{\sum |F_{transl}| + \sum |\tau_{rot}|/d}{I_{sp} \cdot g_0}\,\Delta t, \tag{4}$$

where $F_{transl}$ and $\tau_{rot}$ are the commanded translational forces and rotational torques across all axes, respectively. The constant $d = 1$ m represents the assumed thruster moment arm, $I_{sp} = 220$ s is the specific impulse of the hydrazine propellant, $g_0 = 9.81\,\frac{\mathrm{m}}{\mathrm{s}^2}$ is the standard gravitational acceleration, and $\Delta t = 1.0$ s is the integration timestep. As propellant depletes (600 kg wet to 450 kg dry), the inertia tensor scales proportionally and the spacecraft becomes progressively more agile. A dedicated fuel-gauge observation ($m_{curr}/m_{init}$) allows the neural network to adapt pulse durations dynamically. The reward:

$$R_{fuel} = -0.5 \cdot dm \tag{5}$$

directly optimizes $\Delta V$ in physical units.

## III. Multi-Phase ADR Mission Architecture

The ADR mission is modeled as a four-phase finite-state machine within a single Gymnasium environment. Crucially, the simulation enforces continuous episode progression without phase resets. Rather than artificially re-initializing the spacecraft at idealized coordinates, the agent inherits the exact physical state from the terminal step of the preceding phase. This choice exposes the PPO agent to realistic kinematic boundary conditions, forcing it to immediately counteract unscripted handover perturbations instead of starting from a stable hover. Consequently, the policy must learn a robust control strategy to seamlessly manage these non-ideal transitional dynamics.

### A. Markov Decision Process Formulation

The simulation is modeled as a Markov Decision Process (MDP). The final architecture is based exclusively on a stochastic PPO policy, selected for its ability to manage hybrid action spaces and its stability against physical discontinuities. *State Space ($\mathcal{S}$).* The continuous state space $\mathcal{S} \in \mathbb{R}^{16}$ is expressed in the body frame, comprising: relative position $\vec{r}_{rel}$ and velocity $\vec{v}_{rel}$ to the target, angular velocity $\vec{\omega}$, attitude quaternion $q$, and a normalized fuel-gauge scalar ($m_{curr}/m_{init}$). Phase-specific augmentations are dynamically applied to this core vector. *Action Space ($\mathcal{A}$).* The action space adapts to optimize exploration, transitioning from constrained discrete sets for macro-orbital transfers to a fully unconstrained continuous mapping for 6-DOF proximity operations. *Reward Function ($\mathcal{R}$).* The reward alters its structure across phases, driving the agent through fundamentally different physical tasks.

**Algorithm 1** 4-Phase PPO Training

**Input:** Networks $\theta, \phi$, Base inertia $I_{base}$, Debris mass $m_d$, Arm offset $d_{nose}$

**for** iteration $= 1, 2, \ldots, N$ **do**
  Reset to Phase 0: initialize state $\mathcal{S}_0$ and $I = I_{base}$
  **for** step $t = 0, 1, \ldots, T$ **do**
    Execute $\mathcal{A}_t \sim \pi_\theta(\cdot|\mathcal{S}_t)$, observe $\mathcal{R}_t$ and $\mathcal{S}_{t+1}$
    **if** *Capture Triggered* (Phase 1 $\rightarrow$ Phase 2) **then**
      **Huygens-Steiner Update:** $I \leftarrow I_{base} + m_d d_{nose}^2$
      Preserve $\mathcal{S}_{t+1}$ kinematics (No episodic reset)
    **end if**
    Store transition $(\mathcal{S}_t, \mathcal{A}_t, \mathcal{R}_t, \mathcal{S}_{t+1})$ in buffer $\mathcal{B}$
  **end for**
  Compute Advantages $\hat{A}_t$ via GAE
  Optimize actor $L^{CLIP}(\theta)$ and critic MSE($\phi$) over $K$ epochs
  Clear buffer $\mathcal{B}$
**end for**

### B. Phase 0: Far Rendezvous and Hohmann Transfer

The chaser departs a circular 600 km parking orbit to autonomously intercept the target's orbit at 610 km. This 10 km altitude gap is navigated via a Hohmann transfer. While analytically an optimal two-impulse maneuver requiring exclusively tangential $\Delta V$, it presents a sparse-reward exploration challenge for a DRL agent, which must discover the exact discrete burn sequence and ballistic drift entirely from scratch. *State and Action Space.* To provide the neural network with direct astrodynamic context without relying on internal recurrent memory, the normalized radial velocity $v_{rad}$ was explicitly added to the state vector. In Keplerian motion, radial velocity dictates the orbital phase: a strictly positive $v_{rad}$ identifies the ascending arc, signaling the necessity to coast, whereas the condition $v_{rad} \approx 0$ geometrically identifies the apogee, acting as a temporal trigger for the final circularization burn. Early experiments utilizing an unconstrained 3D continuous action space revealed systematic propellant waste. Recognizing that only prograde and retrograde thrusting efficiently alters orbital energy and Semi-Major Axis (SMA), the action space for this phase was constrained to a 1D discrete set: $\mathcal{A}_{discrete} = \{\text{Prograde}, \text{Coast}, \text{Retrograde}\}$. This pruning drastically reduced the exploration dimensionality from $3^6$ to just $3^1$, eliminating radial exploration noise.

*Reward Function.* The Hohmann transfer requires an extended (≈580 s) ballistic coasting arc between the initial perigee impulse and the final circularization burn at apogee. Because DRL agents inherently struggle with prolonged inaction, an "invisible expert" logic was integrated into the reward function. This mechanism applies a severe penalization of $r_t = -15.0$ for any engine activation during the optimal coasting window, while issuing dense micro-rewards of $r_t = +1.0$ to reinforce the engine-off condition, thereby coercing the agent into learning the required patience.

*Phase Dynamics and Success.* The simulation transitions to Phase 1 only when the chaser successfully enters a strict kinematic funnel, defined by the simultaneous satisfaction of four boundary conditions:

$$\begin{cases} |\Delta a| < 5 \text{ km} \\ e < 0.01 \\ \|\vec{r}_{rel}\| < 5000 \text{ m} \\ \|\vec{v}_{rel}\| < 5 \frac{\text{m}}{\text{s}} \end{cases} \tag{6}$$

where $|\Delta a|$ is the SMA absolute error, $e$ is the orbital eccentricity, $\|\vec{r}_{rel}\|$ and $\|\vec{v}_{rel}\|$ are the relative Euclidean distance and velocity magnitudes. To ensure a highly stable handover, a precision refinement stage tightens the tolerances to:

$$|\Delta a| < 2.5 \text{ km} \quad e < 0.005 \tag{7}$$

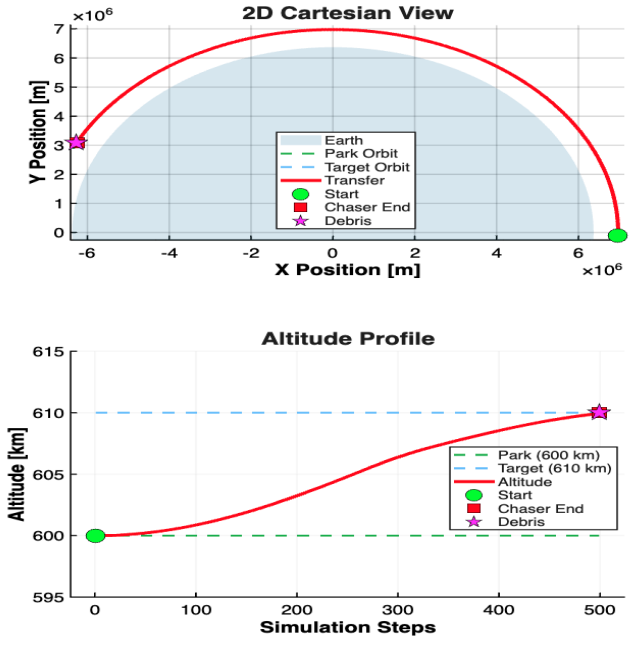


Fig. 1. Phase 0 outbound Hohmann transfer. Top: 2D Cartesian view of the 600–610 km transfer ellipse. Bottom: Altitude profile validating the extended ballistic coasting arc between the perigee and apogee impulses.

### C. Phase 1: Close Approach and Phase 2: Capture and Retreat

Following orbital insertion, the mission transitions to a coupled 6-DOF proximity operation culminating in the physical capture of the uncooperative debris and the subsequent stabilization of the composite system.

*State and Action Space.* The state is augmented with a central obstacle's relative coordinates for Artificial Potential Field (APF) keep-out zones. Requiring simultaneous position and attitude control, the policy outputs a continuous 6-DOF vector $\mathcal{A}_{cont} \in [-1, 1]^6$, mapped linearly to the RCS translational forces and rotational torques. This architecture allows the network to learn complex cross-coupling effects between translational burns and attitude drift.

*Phase 1 Dynamics: Handover and Obstacle Avoidance.* The chaser inherits a residual prograde excess of $+0.5\,\frac{\mathrm{m}}{\mathrm{s}}$ at a distance of 8 m from the target. This perturbation guarantees a catastrophic collision in 16 s and induces significant radial drift via HCW dynamics. The agent must execute an aggressive retrograde braking maneuver upon initialization while simultaneously computing a 6-DOF Collision Avoidance Maneuver (CAM) to bypass the $r_{ost} = 2\,\mathrm{m}$ APF keep-out zone and while maintaining strict line-of-sight pointing toward the target.

*Phase 1 to Phase 2 Transition (Capture Trigger).* The simulation switches from Phase 1 to Phase 2 when all soft-docking conditions are satisfied:

$$\begin{cases} r_{rel} \leq 0.5\,\mathrm{m} \\ \|\vec{v}_{rel}\| < 0.2\,\frac{\mathrm{m}}{\mathrm{s}} \\ \|\vec{\omega}\| < 0.05\,\frac{\mathrm{rad}}{\mathrm{s}} \\ \text{pointing} > 0.9 \end{cases} \tag{8}$$

Upon this trigger, the environment executes an instantaneous physical update: the 100 kg debris mass is rigidly attached at a 2 m longitudinal offset. The system inertia tensor is immediately recalculated via the Huygens-Steiner Parallel Axis Theorem:

$$I_{new} = I_{base} + m_d\, d_{nose}^2, \tag{9}$$

adding $\Delta I = [400, 10, 400]$ kg·m$^2$ to the pitch and yaw axes. This transition occurs mid-episode without resetting the agent, imposing an immediate kinematic discontinuity that the policy must manage in a continuous control loop.

*Phase 2 Dynamics: Sluggish Control Adaptation.* The updated inertia radically alters vehicle response. Euler's rotational equations:

$$\dot{\vec{\omega}} = I^{-1}\vec{\tau} \tag{10}$$

dictate that angular acceleration drops by a factor of $9\times$ on the transverse axes, transforming the agile chaser into an asymmetrically loaded composite vehicle. Unlike classical deterministic controllers the DRL agent must autonomously adapt to the new control sensitivity. During this phase, the GNC objective inverts: the new target becomes the initial staging point, requiring prolonged, sustained torque pulses to maneuver the unbalanced orbital tug without inducing tumbling.

*Reward Function (Milestone Shaping).* To mitigate the risk of gradient vanishing over the 1000-step dual-phase episode, a breadcrumb strategy uses progressive milestone rewards. A $+50.0$ reward is issued exactly at the moment of the Capture Trigger, followed by a final $+100.0$ reward upon successfully returning to the staging point. This strategy provides a stable gradient signal across the extreme physical discontinuity of the capture event, ensuring the policy reliably masters the entire transitional sequence.

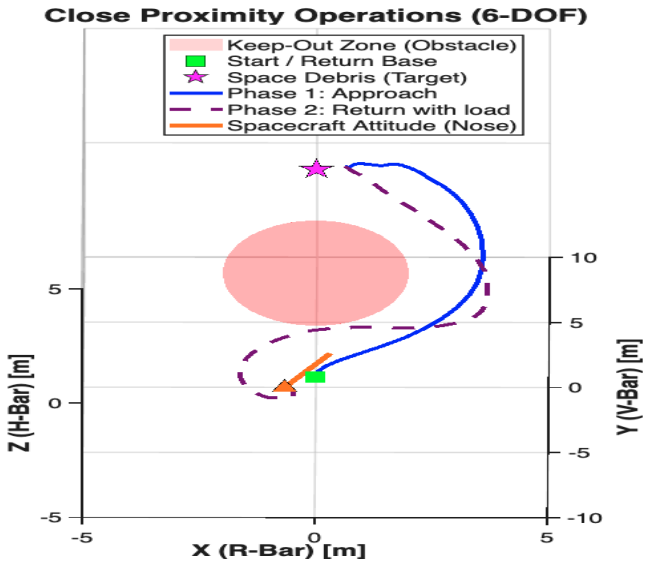


Fig. 2. 6-DOF trajectory in the relative LVLH frame during Close Proximity Operations. The blue curve (Phase 1) demonstrates the initial CAM maneuver to bypass the large central bstacle's APF Keep-Out Zone. The purple dashed curve (Phase 2) illustrates the stabilized return path of the composite vehicle, providing visual evidence of the agent's real-time adaptation to the Huygens-Steiner inertia shift.

### D. Phase 3: Payload De-orbiting

The final objective is to safely de-orbiting the composite vehicle from the 610 km capture altitude back to the 600 km disposal zone. In this phase, the agent transitions back to macro-scale orbital maneuvering, challenged by the increased mass ($\approx$620 kg) and the asymmetric inertia tensor.

Unlike prior phases, de-orbiting does not target a specific spatial coordinate or a physical object. Instead, the objective is global altitude reduction to the closest point on the target orbital shell to minimize energy expenditure.

*State and Action Space.* The state reverts to the core 16-dimensional telemetry vector, as proximity LIDAR data is no longer relevant for descent. The action space remains continuous 6-DOF, requiring sustained translational thrust to counteract the heavily loaded tug's sluggishness.

*Reward Function (Astrodynamic Energy).* Standard DRL approaches relying on naive Cartesian distance formulations for target tracking fail in this phase. If the 600 km destination is modeled as a fixed physical coordinate below the spacecraft,

the Euclidean distance from a 610 km circular orbit to that point remains a constant ≈10 km throughout the entire orbital revolution. This provides zero learning gradient, causing the agent to converge to a trivial local optimum: shutting down all thrusters to avoid fuel penalties and perpetually drifting at 610 km.

This "ghost-target" stagnation was resolved by using a purely orbital mechanics-based energy reward. The normalized Semi-Major Axis (SMA) descent is tracked across a carefully shaped two-stage reward function:

$$Q = \mathrm{SMA}_{desc} - \big(e \cdot 100 \cdot \max(0,\, \mathrm{SMA}_{desc})\big), \tag{11}$$

where $\mathrm{SMA}_{desc} = 0$ at the initial 610 km orbit and reaches 1 at the 600 km target. In Stage 1, the agent receives a positive gradient for executing the initial retrograde burn at apogee that lowers the perigee. In Stage 2, as the vehicle approaches the target altitude, the eccentricity penalty ($e$) activates, dynamically forcing the agent to execute a second circularization burn at perigee. This structure coerces the neural network into autonomously discovering the theoretical optimality of a two-burn Hohmann return transfer.

*Phase Dynamics and Success.* Due to the system's sluggish rotational response, achieving precise orbital insertion is highly complex, requiring the policy to initiate maneuvers with significant lead time. Success requires the simultaneous satisfaction of the following tolerances:

$$\begin{cases} |\Delta h| < 250\,\mathrm{m} \\ e < 0.005 \end{cases} \tag{12}$$

where $|\Delta h|$ is the absolute altitude error relative to the 600 km circular disposal zone.

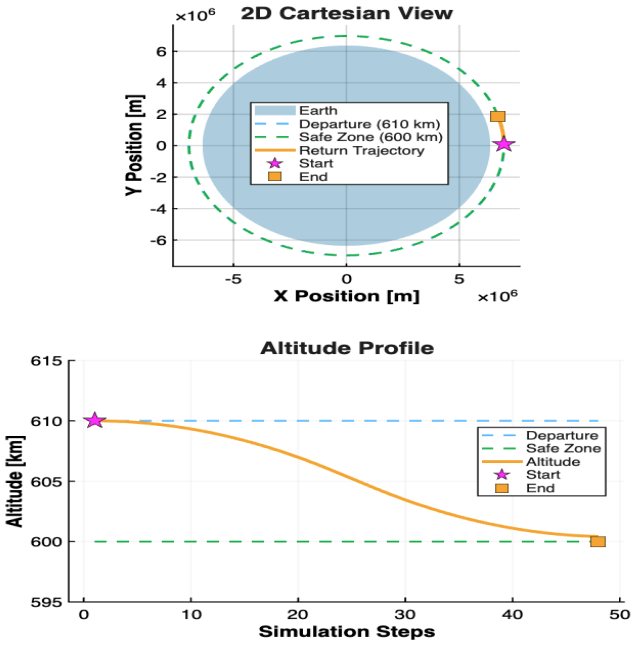


Fig. 3. Phase 3 de-orbiting maneuver of the composite vehicle. Top: 2D Cartesian view of the retrograde Hohmann return. Bottom: Altitude profile validating the 610–600 km descent managed by the energy reward.

## IV. Simulation Results

### A. Developmental Progression and Training Flow

The project followed a bottom-up developmental flow (Table I) to bridge simplified kinematics with mission realism. Stages 1–2 represent preliminary research where DDPG validated basic 2D relative motion. The subsequent transition to the final 4-phase architecture (Stages 3–6) adopted PPO to manage increased complexity. PPO natively supports *Multi-Discrete* actions for hybrid spaces, while its clipped surrogate objective:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\Big[\min\Big(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta),\ 1{-}\varepsilon,\ 1{+}\varepsilon)\hat{A}_t\Big)\Big] \tag{13}$$

ensures numerical stability against gradient shocks and inertia shifts during the capture event.

TABLE I
Developmental Stages toward the Final 4-Phase Architecture

| Stage | Environment | Alg. | Steps |
|---|---|---|---|
| 1 | 2D Planar Kinematics | DDPG | 50K |
| 2 | 2D CW Relative Orbit | DDPG | 80K |
| 3 | 3D Cross-track | PPO | 300K |
| 4 | Keplerian N-Body ECI | PPO | 750K |
| 5 | 6-DOF + Tsiolkovsky | PPO | 1.5M |
| 6 | End-to-End 4-Phase ADR | PPO | 23M |
| | **Total Development steps** | | **≈25.5M** |

Due to sparse rewards and extreme exploration demands within the End-to-End 4-Phase ADR, orbital Phases 0 and 3 were trained using a specific Curriculum Learning (CL) strategy for precision refinement. Mastering long-arc energy tracking and astrodynamic boundaries necessitated 10 million timesteps each. Conversely, 6-DOF proximity operations (Phases 1–2) converged more rapidly by leveraging the spatial awareness established during earlier developmental stages. The final mission policy was achieved after a cumulative trajectory of approximately 23 million timesteps.

### B. Hardware Parameters and Safety Constraints

TABLE II
Spacecraft Hardware and Mission Safety Constraints

| Parameter | Symbol | Value |
|---|---|---|
| Chaser Initial Wet Mass | $m_0$ | 600.0 kg |
| Chaser Dry Mass | $m_{dry}$ | 450.0 kg |
| Target Debris Mass | $m_d$ | 100.0 kg |
| Max RCS Thrust (transl.) | $F_{max}$ | 50.0 N |
| Max RCS Torque (rot.) | $\tau_{max}$ | 5.0 Nm |
| Specific Impulse | $I_{sp}$ | 220.0 s |
| Base Inertia Tensor | $I_{base}$ | [50, 50, 50] kg·m$^2$ |
| Debris Added Inertia | $\Delta I$ | [400, 10, 400] kg·m$^2$ |
| Capture Arm Offset | $d_{nose}$ | 2.0 m |
| Keep-Out Zone Radius | $r_{ost}$ | 2.0 m |
| Capture Radius Tolerance | $r_{goal}$ | 0.5 m |
| Max Docking Velocity | $v_{max}$ | 0.2 $\frac{\mathrm{m}}{\mathrm{s}}$ |
| Max Tumbling Rate | $\omega_{max}$ | 0.05 $\frac{\mathrm{rad}}{\mathrm{s}}$ |
| Min Pointing Accuracy | — | 0.9 (cosine) |
| Max Episode Duration | $T_{max}$ | 1000 steps |

### C. Quantitative Results

To explicitly address the limitations of existing literature—relying on isolated mission segments, constant-mass, and static inertial assumptions—the quantitative evaluation strictly isolates the agent's performance during unmitigated phase transitions and abrupt dynamic shifts.

**Phase 0.** The agent converged on the two-burn Hohmann sequence without explicit programming. The $v_{rad}$ observation proved critical for autonomous burn timing. CL precision refinement consistently achieved:

$$\begin{cases} |\Delta a| < 2.5\,\mathrm{km} \\ e < 0.005 \end{cases} \tag{14}$$

providing a clean kinematic funnel for Phase 1.

**Phase 1.** Unlike standard approaches that artificially reinitialize the agent in a perfectly stable hover, our continuous formulation forces the policy to handle real residual velocities. Despite the $+0.5\,\frac{\mathrm{m}}{\mathrm{s}}$ handover perturbation, the trained policy executed an immediate retrograde correction and achieved full attitude stabilization:

$$\|\vec{\omega}\| < 0.05\,\frac{\mathrm{rad}}{\mathrm{s}} \tag{15}$$

bypassed the central obstacle via a CAM, and completed soft docking within tolerances. This quantitatively validates the agent's robustness to unscripted transitional dynamics.

**Phase 2.** Moving beyond ideal symmetric simulations, the policy successfully absorbed the abrupt physical discontinuity triggered by the capture event. Following the instantaneous inertia shock ($+400\ \mathrm{kg}\cdot\mathrm{m}^2$ in pitch/yaw), the vehicle's rotational agility dropped severely. The agent autonomously extended torque pulse durations to compensate for the degraded angular acceleration. The asymmetric composite vehicle was stabilized and the retreat completed without explicit controller retuning, demonstrating a clear advantage over deterministic controllers requiring offline parameter updates.

**Phase 3.** The astrodynamic energy reward (11) resolved the ghost-target stagnation typical of Cartesian formulations. The agent converged on the two-burn retrograde return, achieving disposal insertion with:

$$\begin{cases} |\Delta h| < 250\,\mathrm{m} \\ e < 0.005 \end{cases} \tag{16}$$

after CL refinement.

**Full mission.** Across the complete four-phase ADR profile, the agent consistently demonstrated fuel-optimized maneuverability. By successfully navigating the unmitigated kinematic handovers, the abrupt Huygens-Steiner inertia shift, and the continuous Tsiolkovsky propellant dynamics—where the vehicle mass is explicitly treated as a time-varying parameter driven by continuous fuel depletion rather than assuming an unrealistic constant baseline—the proposed architecture directly bridges the Sim-to-Real gap where traditional isolated-environment DRL agents would fail, completing the operation with zero collisions and strict constraint compliance throughout.

## V. Conclusion

This paper presented a high-fidelity 6-DOF DRL architecture for end-to-end autonomous Active Debris Removal. Departing from HCW linear approximations in favor of an absolute ECI Keplerian N-Body propagator exposes the agent to true spaceflight non-linearities. The continuous Tsiolkovsky integration, imperfect kinematic handovers, and instantaneous Huygens-Steiner inertia modification upon debris capture constitute a rigorous, non-linear simulation environment tailored for missions such as ClearSpace-1. Guided by astrodynamic energy rewards and a nine-stage Curriculum Learning strategy spanning 23 million timesteps, the PPO agent executed fuel-optimized Hohmann transfers, millimetric 6-DOF soft docking despite kinematic drift, and autonomous policy adaptation to the degraded dynamics of the 700 kg composite orbital tug—completing the full mission with highly optimized hydrazine consumption and zero constraint violations.

Future developments will address environmental perturbations ($J_2$, atmospheric drag, solar radiation pressure), tumbling debris models, and sensor noise matrices to bridge the Sim-to-Real gap. Multi-Agent RL cooperative swarms and Hierarchical RL for robotic arm coordination represent the natural evolution toward hardware-in-the-loop ADR validation.

## References


[1] D. J. Kessler and B. G. Cour-Palais, "Collision frequency of artificial satellites: The creation of a debris belt," *J. Geophys. Res.*, vol. 83, no. A6, pp. 2637–2646, 1978.

[2] European Space Agency, "DISCOS: Database and Information System Characterising Objects in Space," ESA Space Debris Office, 2024.

[3] European Space Agency, "ClearSpace-1: Active Debris Removal Mission," ESA Space Safety Programme, 2020.

[4] European Cooperation for Space Standardization, "Space engineering: Spacecraft attitude and orbit control," *ECSS-E-ST-60-30C*, 2013.

[5] W. H. Clohessy and R. S. Wiltshire, "Terminal guidance system for satellite rendezvous," *J. Aerosp. Sci.*, vol. 27, no. 9, pp. 653–658, 1960.

[6] J. Broida and R. Linares, "Spacecraft rendezvous guidance in cluttered environments via reinforcement learning," in *29th AAS/AIAA Space Flight Mechanics Meeting*, Ka'anapali, HI, 2019.

[7] K. Hovell and S. Ulrich, "Deep reinforcement learning for spacecraft proximity operations guidance," *J. Spacecr. Rockets*, vol. 58, no. 2, pp. 254–264, 2021.

[8] L. Brana, F. Capolupo, and M. Lavagna, "Deep reinforcement learning for autonomous spacecraft rendezvous with a non-cooperative target," in *Proc. 73rd Int. Astronautical Congress (IAC)*, Paris, 2022.

[9] B. Gaudet, R. Linares, and R. Furfaro, "Adaptive guidance and integrated navigation with reinforcement meta-learning," *Acta Astronautica*, vol. 169, pp. 180–190, 2020.

[10] M. J. Sidi, *Spacecraft Dynamics and Control*. Cambridge, UK: Cambridge University Press, 1997.

[11] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. ICLR*, San Juan, PR, 2016.

[12] J. Schulman *et al.*, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017.

[13] W. Fehse, *Automated Rendezvous and Docking of Spacecraft*. Cambridge, UK: Cambridge University Press, 2003.

[14] J. C. Bell, "Multi-phase Spacecraft Proximity Operations Using Model Predictive Control," M.S. thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH, 2021.

[15] A. Selim *et al.*, "Real-time optimal control of a 6-DOF state-constrained close-proximity rendezvous mission," *Acta Astronautica*, vol. 211, pp. 100-115, 2023.

[16] G. Behrendt *et al.*, "Time-Constrained Model Predictive Control for Autonomous Satellite Rendezvous, Proximity Operations, and Docking," *arXiv preprint arXiv:2501.13236*, 2025.