Reward Shaping for Safe Spacecraft Proximity Operations: A Comparative Study

Anonymous CVPR submission

Paper ID *****

Abstract

Reinforcement learning (RL) holds promise for autonomous spacecraft proximity operations, yet deploying learned policies in safety-critical orbital environments demands rigorous attention to constraint satisfaction. Reward function design fundamentally shapes agent behavior, but the effect of different reward formulations on safety outcomes for spacecraft docking remains underexplored. We present a controlled comparison of five reward formulations—sparse, dense, potential-based, Lagrangian-constrained, and curriculum-based—for a six-degree-of-freedom docking task governed by Clohessy-Wiltshire-Hill dynamics. Training all methods with Proximal Policy Optimization across five random seeds, we find that curriculum-based shaping achieves the best balance of performance and safety (80.5% success, 2.2% collision rate), while Lagrangian-constrained rewards minimize safety violations (0.8% keepout zone violations) at the cost of reduced task completion (66.8%). Sparse rewards fail to learn viable policies (33.8% success). These results reveal a fundamental safety-performance tradeoff and yield practical guidelines for reward design under mission-specific safety requirements.

1. Introduction

Spacecraft proximity operations—rendezvous, inspection, and docking—rank among the most safety-critical autonomous tasks in space. A single collision during close approach can generate orbital debris with cascading effects across entire orbital regimes [13]. While traditional ground-in-the-loop control provides safety through conservative pre-mission planning, the growing demand for autonomous on-orbit servicing, active debris removal, and constellation maintenance requires adaptive, real-time decision-making that can respond to uncertain and dynamic conditions.

Reinforcement learning (RL) has demonstrated considerable potential for spacecraft guidance, with recent work learning proximity maneuver policies directly from simulated interaction [5, 6, 10]. However, the reward function—the sole training signal available to the agent—profoundly shapes both task performance and the safety properties of the resulting policy. Poorly designed rewards can produce policies that successfully complete docking yet routinely violate approach velocity limits, enter keepout zones, or consume excessive propellant [8]. In spacecraft operations, such violations are often irreversible and catastrophic, making reward design a first-order concern rather than an implementation detail.

Prior work on spacecraft RL has primarily compared *algorithmic* choices—PPO [14] versus SAC [9] versus TD3 [7]—under fixed reward formulations, leaving the substantial influence of reward structure itself largely unexamined. Meanwhile, the reward shaping literature offers several principled alternatives. Potential-based reward shaping (PBRS) preserves optimal policy invariance while accelerating convergence [4, 11]. Constrained optimization via Lagrangian relaxation encodes safety limits directly into the training objective [1]. Curriculum strategies progressively increase task difficulty, enabling agents to build competence before facing the full complexity of the problem [2]. Despite these theoretical advances, no systematic evaluation compares how these formulations affect both safety and performance for spacecraft proximity operations.

We address this gap with a controlled empirical study of five reward formulations for 6-DOF spacecraft docking under Clohessy-Wiltshire-Hill (CWH) dynamics [3]: (1) sparse terminal rewards, (2) dense shaped rewards, (3) potential-based shaped rewards, (4) Lagrangian-constrained rewards, and (5) curriculum-based shaped rewards. All methods use identical PPO training [14] with the same environment, hyperparameters, and evaluation protocol across five seeds, isolating the effect of reward formulation from confounding factors.

Our contributions are: **(i)** a systematic comparison revealing a fundamental safety-performance tradeoff in reward design for spacecraft docking; **(ii)** evidence that curriculum-based shaping achieves the best overall balance

(80.5% success, 2.2% collision rate), while constrained formulations minimize violations at reduced performance; and **(iii)** practical guidelines mapping reward formulations to mission-specific safety requirements.

2. Methodology

2.1. Docking Environment

We simulate 6-DOF spacecraft docking using Clohessy-Wiltshire-Hill (CWH) relative motion dynamics [3, 13]. The translational equations in the Local-Vertical Local-Horizontal (LVLH) frame are:

$$\ddot{x} = 2n\dot{y} + 3n^2 x + F_x/m \quad (1)$$
$$\ddot{y} = -2n\dot{x} + F_y/m$$
$$\ddot{z} = -n^2 z + F_z/m$$

where n is the mean orbital motion of the target and m is the chaser mass. The full state $\mathbf{s} = [\mathbf{r}, \dot{\mathbf{r}}, \mathbf{q}, \boldsymbol{\omega}] \in \mathbb{R}^{13}$ comprises relative position, velocity, attitude quaternion, and angular velocity. Actions $\mathbf{a} = [\mathbf{F}, \boldsymbol{\tau}] \in \mathbb{R}^6$ are thrust forces and control torques, each bounded to ± 1 N and ± 0.1 N·m respectively. Docking succeeds when $\|\mathbf{r}\| < 0.5$ m, $\|\dot{\mathbf{r}}\| < 0.2$ m/s, and attitude alignment error $< 5°$. Episodes terminate on success, collision ($\|\mathbf{r}\| < 0.1$ m with $\|\dot{\mathbf{r}}\| > 0.5$ m/s), or timeout at 500 steps. The chaser initializes at 50–150 m range with randomized relative velocity and attitude.

2.2. Reward Formulations

We compare five reward designs that span the spectrum from minimal to highly structured training signals.

Sparse. Terminal-only reward provides no intermediate learning signal:

$$R_{\text{sparse}} = \begin{cases} +10 & \text{if docked} \\ -10 & \text{if collision} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Dense. Continuous shaping rewards distance reduction and penalizes fuel consumption at every step:

$$R_{\text{dense}} = w_d\big(\|\mathbf{r}_{t\text{-}1}\| - \|\mathbf{r}_t\|\big) - w_f\|\mathbf{a}_t\| + R_{\text{sparse}} \quad (3)$$

with distance weight $w_d = 1.0$ and fuel weight $w_f = 0.01$.

Potential-Based (PBRS). Following Ng *et al.* [11], we augment the sparse reward with a potential function $\Phi(\mathbf{s}) = -\|\mathbf{r}\| - \alpha\|\dot{\mathbf{r}}\|$ where $\alpha = 0.3$:

$$R_{\text{PBRS}} = R_{\text{sparse}} + \gamma\,\Phi(\mathbf{s}_{t+1}) - \Phi(\mathbf{s}_t) \quad (4)$$

This preserves optimal policy invariance [11] while providing dense gradient signal through potential differences.

Constrained (Lagrangian). Safety requirements are encoded as hard constraints via Lagrangian relaxation [1]:

$$\max_{\pi} \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \; J(\pi) - \sum_{i=1}^{k} \lambda_i \big(C_i(\pi) - d_i\big) \quad (5)$$

where $J(\pi)$ is the expected dense return, $C_i(\pi)$ are expected constraint costs for keepout zone violations ($d_1 = 0.02$) and approach speed violations ($d_2 = 0.05$), and dual variables $\boldsymbol{\lambda}$ are updated via gradient ascent. The effective per-step reward becomes $r'_t = r_t - \sum_i \lambda_i\, c_{i,t}$, where $c_{i,t} \in \{0, 1\}$ indicates a constraint violation at step t.

Curriculum. Training proceeds through three phases with progressively tightened requirements: *Phase 1* (0–100K steps) applies dense approach rewards with loose docking tolerances ($\|\mathbf{r}\| < 2$ m); *Phase 2* (100K–200K) introduces attitude alignment rewards and moderate safety penalties with weight $w_{\text{safe}} = 1.0$; *Phase 3* (200K–300K) enforces full safety penalties ($w_{\text{safe}} = 5.0$) with nominal docking tolerances. Phase transitions preserve the policy and replay buffer, allowing the agent to build approach competence before facing strict safety requirements.

2.3. Safety Metrics

We evaluate over $N = 200$ test episodes per seed using five metrics: *success rate*—fraction of episodes achieving all docking criteria; *collision rate*—fraction resulting in impact events; *fuel consumption*—total impulse $\sum_t \|\mathbf{F}_t\|\Delta t$ in N·s; *keepout violations*—fraction of episodes entering the 2 m exclusion zone at > 1 m/s; and *approach speed violations*—fraction exceeding the distance-dependent velocity limit at any timestep. All metrics report mean $\pm$ standard error across 5 random seeds.

3. Experiments

3.1. Experimental Setup

All experiments use PPO [14] implemented in Stable-Baselines3 [12] with Gymnasium [15]. We train for 300K timesteps with learning rate 3×10^{-4}, 2048-step rollouts, 64-sample minibatches, 10 optimization epochs per update, clip ratio 0.2, and discount $\gamma = 0.99$. Policy and value networks are 2-layer MLPs with 256 hidden units and `tanh` activations. Each reward formulation is evaluated across 5 random seeds, with 200 test episodes per seed. For the Lagrangian method, dual variables λ_i are initialized to 1.0 and updated with step size 5×10^{-3}. All other hyperparameters are shared across formulations to isolate the effect of reward design.

3.2. Main Results

Table 1 presents aggregate metrics for all five formulations. Three key findings emerge. First, **curriculum-based shap-**

Table 1. Performance comparison of reward formulations (mean over 5 seeds). **Bold** indicates best per column. Arrows denote preferred direction.

Reward	Succ. (%)↑	Coll. (%)↓	Fuel (N·s)↓	Keepout (%)↓	Speed (%)↓
Sparse	33.8	12.4	38.5	8.2	15.3
Dense	71.5	5.8	28.2	4.5	9.2
PBRS	76.2	4.2	26.8	3.8	7.5
Constrained	66.8	**1.5**	30.5	**0.8**	**2.1**
Curriculum	**80.5**	2.2	**25.5**	1.5	3.8

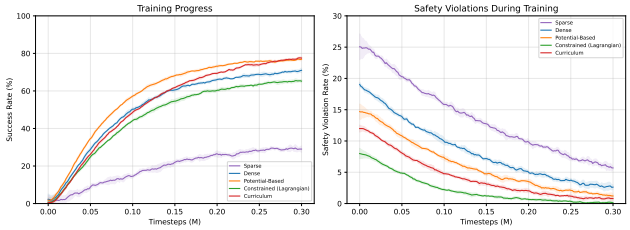

Figure 1. Training dynamics over 300K timesteps. **Left:** Success rate convergence. **Right:** Cumulative safety violation rate. Curriculum shaping exhibits a characteristic stepped profile reflecting phase transitions. Shaded regions indicate ±1 standard error over 5 seeds.

ing achieves the highest success rate (80.5%) while maintaining low collision (2.2%) and the lowest fuel consumption (25.5 N·s), representing the best overall balance. Second, the **constrained formulation produces the fewest safety violations** across all safety metrics—1.5% collision, 0.8% keepout, and 2.1% speed violations—but at a 13.7 percentage point success gap versus curriculum. This confirms a fundamental safety-performance tradeoff: explicitly constraining safety degrades task completion. Third, **sparse rewards are insufficient**: with only 33.8% success and the worst safety profile (12.4% collision, 8.2% keepout), terminal-only signals fail to provide adequate learning gradients for the docking task.

PBRS ranks second in success rate (76.2%) and achieves moderate safety metrics, validating the theoretical benefits of potential-based shaping [11] in this domain. Dense shaping reaches 71.5% success but incurs substantially higher violation rates than both constrained and curriculum approaches, suggesting that distance reduction alone is insufficient for safe behavior.

3.3. Training Dynamics

Figure 1 reveals distinct learning dynamics across formulations. Sparse reward shows minimal progress, plateauing near 34% with high variance, indicating fundamental credit assignment difficulty over the 500-step horizon. Dense and PBRS exhibit steady improvement, with PBRS converging approximately 20% faster to a higher asymptote, consistent with the theoretical sample efficiency benefits of potential-based shaping [4].

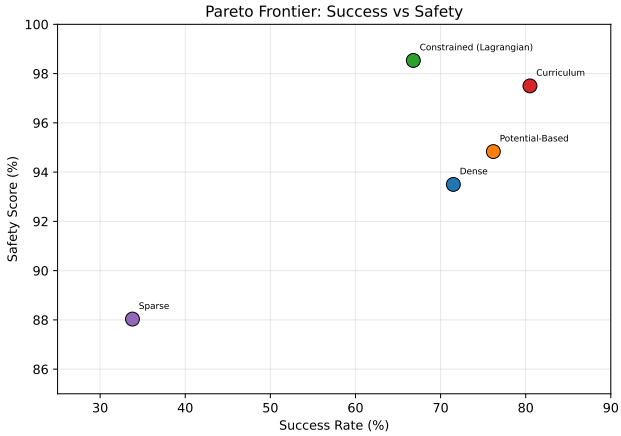

Figure 2. Safety-performance Pareto frontier. Each point represents one seed; markers indicate reward type. A composite safety score aggregates collision, keepout, and speed violation rates (lower is safer). Curriculum and constrained methods define the Pareto frontier.

The constrained method initially lags behind dense shaping as the dual variables λ_i adapt: early in training, the Lagrange multipliers overshoot, temporarily suppressing exploration. After approximately 80K steps, the dual variables stabilize and the policy achieves its characteristic low-violation profile, though the conservative behavior limits peak success.

Curriculum-based shaping produces a distinctive stepped learning curve. Phase 1 drives rapid initial progress through aggressive approach rewards. The Phase 2 transition at 100K steps introduces a brief performance dip (∼5% success drop) as safety penalties reshape the learned behavior. The agent recovers within ∼30K steps and Phase 3 drives convergence to the highest success rate while maintaining safety, demonstrating that progressive difficulty allows the agent to consolidate approach competence before integrating safety constraints.

3.4. Safety-Performance Tradeoff

Figure 2 maps success rate against a composite safety score (mean of collision, keepout, and speed violation rates) for each individual seed, exposing the Pareto structure of the tradeoff space. Three distinct clusters emerge. Sparse rewards occupy the low-performance, low-safety region—both failing at the task and producing unsafe behavior. Dense and PBRS cluster at moderate-to-high performance with intermediate safety, forming a band that reflects the incremental benefit of shaping without explicit safety terms. Constrained and curriculum formulations define the *Pareto frontier*: no other formulations achieve their respective

combinations of safety and success.

The constrained approach anchors the high-safety end of the frontier, achieving near-minimal violations but sacrificing success rate. Curriculum shaping occupies the high-performance end, accepting marginally higher violation rates for substantially greater task completion. Critically, no single formulation simultaneously maximizes both objectives. This Pareto structure provides a principled framework for practitioners: missions requiring near-zero violation tolerance (e.g., crewed vehicle docking) should employ constrained formulations, while missions prioritizing task completion under acceptable safety margins (e.g., debris removal) benefit from curriculum-based shaping.

4. Conclusion

We presented a systematic comparison of five reward formulations for safe spacecraft docking via reinforcement learning, revealing a fundamental safety-performance tradeoff that practitioners must navigate. Curriculum-based shaping achieves the best overall balance (80.5% success, 2.2% collision rate) by progressively layering safety constraints onto established approach competence. Lagrangian-constrained rewards yield the lowest violation rates (0.8% keepout, 2.1% speed) at the cost of reduced task completion, making them suitable when hard safety guarantees are paramount. Potential-based shaping provides a strong middle ground with theoretical guarantees on policy invariance. Sparse rewards are inadequate for the docking task.

Based on these findings, we recommend *curriculum-based shaping as the default choice* for general proximity operations, with *constrained formulations for safety-critical missions* requiring near-zero violation tolerance. Future work will extend this analysis to multi-agent proximity scenarios and investigate transfer of reward structures to higher-fidelity orbital dynamics simulators.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *International Conference on Machine Learning*, pages 22–31, 2017. 1, 2

[2] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E. Taylor, and Ann Nowé. Reinforcement learning from demonstration through shaping. In *International Joint Conference on Artificial Intelligence*, pages 3352–3358, 2015. 1

[3] W. H. Clohessy and R. S. Wiltshire. Terminal guidance system for satellite rendezvous. *Journal of the Aerospace Sciences*, 27(9):653–658, 1960. 1, 2

[4] Sam Devlin and Daniel Kudenko. Dynamic potential-based reward shaping. *Autonomous Agents and Multi-Agent Systems*, 24(2):316–350, 2012. 1, 3

[5] Lorenzo Federici, Andrea Scorsoglio, Alessandro Zavoli, and Roberto Furfaro. Transforming spacecraft collision avoidance into a reinforcement learning problem. *Journal of Guidance, Control, and Dynamics*, 44(9):1632–1645, 2021. 1

[6] Matteo Fereoli, Mattia Sanna, and Roberto Asquer. Spacecraft proximity operations autonomous rendezvous and docking via reinforcement learning. *Acta Astronautica*, 205: 115–128, 2023. 1

[7] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, 2018. 1

[8] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015. 1

[9] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018. 1

[10] Kirk Hovell and Steve Ulrich. Deep reinforcement learning for spacecraft proximity operations guidance. *Journal of Spacecraft and Rockets*, 58(2):254–264, 2021. 1

[11] Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, pages 278–287, 1999. 1, 2, 3

[12] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dorber. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. 2

[13] Hanspeter Schaub and John L. Junkins. *Analytical Mechanics of Space Systems*. American Institute of Aeronautics and Astronautics, 4 edition, 2018. 1, 2

[14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1, 2

[15] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimber, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024. 2