Deep Reinforcement Learning for Spacecraft Attitude Control During Atmospheric Re-Entry

Alexander Fabisch[1], Melvin Laux[1], Mariela De Lucas Álvarez[1], Edoardo Caroselli[2], Julian Theis[2]
[1]DFKI GmbH, Robotics Innovation Center: {first_name.last_name}@dfki.de
[2]Airbus Defence and Space GmbH: {first_name.last_name}@airbus.com

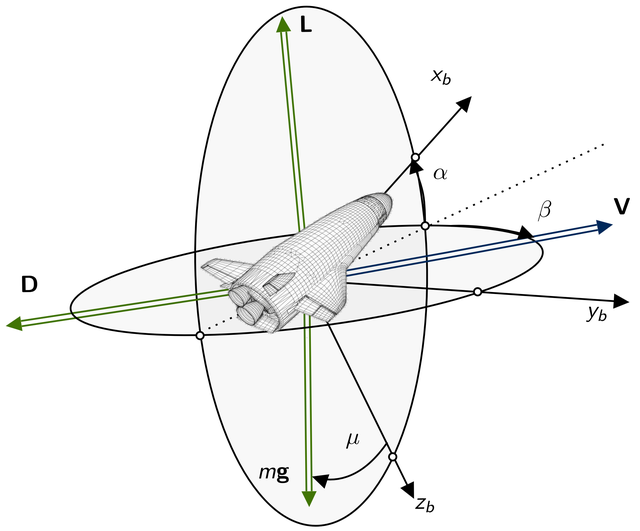
Motivation

- Aerodynamic conditions change rapidly in hypersonic atmospheric re-entry
- Deep RL could improve handling nonlinear dynamics, uncertainties, and failure cases
- Applicability of SotA deep RL is largely unexplored as spacecraft control is safety-critical

Contributions

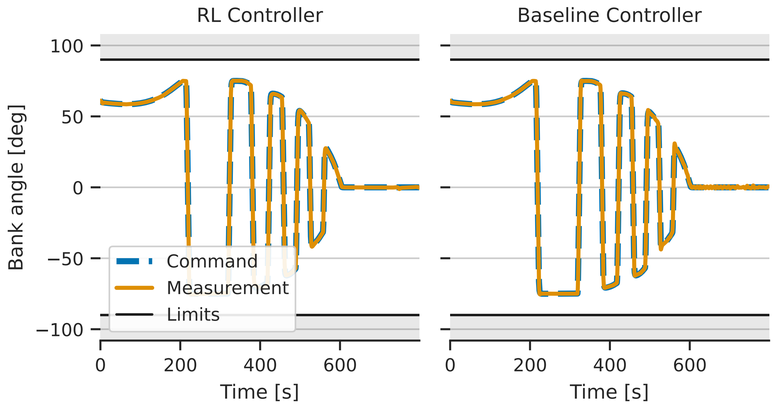
1. Test SotA deep RL for hypersonic re-entry → MR.Q excels, surpasses baseline controller
2. Test OOD generalization: hybrid vs. pure RL → hybrid control is usually better
3. Evaluate task scheduling vs. dynamics randomization → uniform random DR is sufficient

Simulation Environment



- 3-DOF spacecraft model along a pre-defined re-entry trajectory
- Controlled variables: angle of attack α, sideslip angle β, bank angle μ
- Actuators: aerodynamic flaps (pitch/roll) + RCS thrusters (yaw)
- Contextual MDP [1, 6] with mass, inertia tensor, and flap actuator bandwidth in context

Re-Entry Trajectory: Bank Angle



Task Formulation

Total reward:

$$r_{t'} = \text{attitude reward}\,(t) - \text{control cost}\,(t),$$

Transformation of the attitude cost to a positive range prevents early termination:

$$\text{attitude reward}\,(t) = \max(0.001, \exp(-w_c \cdot c(t))) \in [0.001, 1]\,.$$

Attitude cost:

$$c(t) = \left(\frac{\alpha_{\text{cmd}}(t) - \alpha(t)}{\alpha_{\text{range}}}\right)^2 + \left(\frac{\beta_{\text{cmd}}(t) - \beta(t)}{\beta_{\text{range}}}\right)^2 + \left(\frac{\mu_{\text{cmd}}(t) - \mu(t)}{\mu_{\text{range}}}\right)^2.$$

Method

RL Algorithms (no task-specific tuning): SAC [5] (max. entropy), TD3 [4] (deterministic policy gradient), TD7 [2] (state encoder), MR.Q [3] (model-based encoder)
Control Architectures: Baseline PID | Pure RL | Additive hybrid: $a = a_{\text{PID}}(o) + \pi(o)$ | GS hybrid: $a = a_{\text{PID}}(\pi(o))$
Evaluation: IQM + 95 % bootstrap CI, 10 seeds, 100 test contexts for dynamics randomization

Training under Nominal Conditions

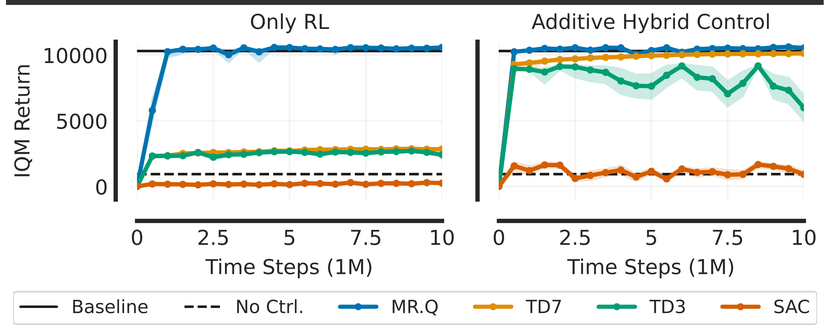


Figure: MR.Q surpasses the baseline controller. Learning curves with interquartile means (IQM) and 95% bootstrap confidence intervals for 10 training seeds.

Out-of-Distribution Evaluation

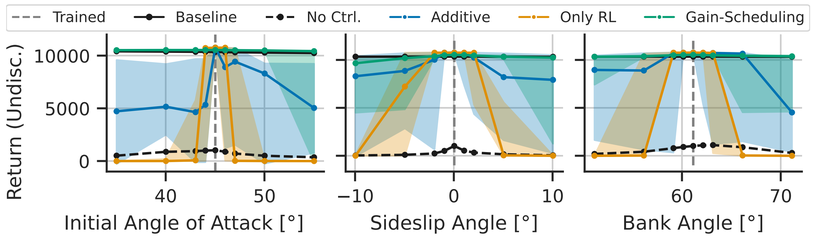


Figure: Performance with respect to changes of the initial attitude without training for it. Performance is measured at the initial attitude during training $+\epsilon \in [-10, -5, -2, -1, 0, 1, 2, 5, 10]$ degree per component.

Website: Paper and Code

Training under Domain Randomization: Learning Curves

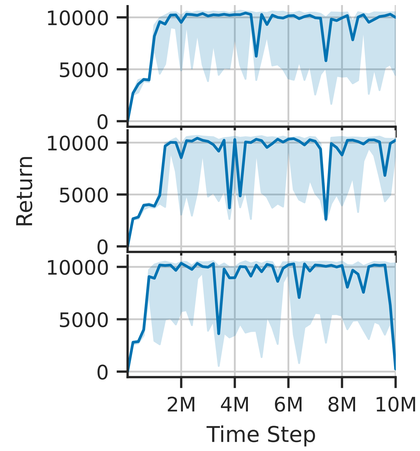


Figure: Learning curves of three training runs with DR and pure RL with MR.Q. Lines indicate median performance over 30 test contexts and shaded areas show the full range of returns.

Training under Domain Randomization: Error Statistics

Table: Evaluation of best controllers with domain-specific metrics. Cell background color indicates best values or values better than the baseline.

Training	Seed	Success	$\lvert e_\alpha \rvert$ [deg] Median	90-Perc.	95-Perc.	98-Perc.	$\lvert e_\beta \rvert$ [deg] Median	90-Perc.	95-Perc.	98-Perc.	$\lvert e_\mu \rvert$ [deg] Median	90-Perc.	95-Perc.	98-Perc.
Baseline	-	79%	0.05	0.46	0.67	2.96	0.02	0.20	0.36	0.57	0.06	0.72	1.43	2.35
Add. Hyb.	0	93%	0.05	0.23	0.34	0.59	0.07	0.25	0.41	0.65	0.17	0.42	0.59	0.87
Add. Hyb.	1	93%	0.07	0.37	0.56	0.83	0.07	0.30	0.53	0.85	0.15	0.60	0.84	1.11
Add. Hyb.	2	94%	0.07	0.33	0.47	0.71	0.11	0.39	0.60	0.84	0.17	0.60	0.84	1.19
Only RL	0	100%	0.12	0.47	0.63	0.86	0.09	0.45	0.88	1.87	0.23	1.14	1.63	2.19
Only RL	1	99%	0.10	0.45	0.62	0.84	0.10	0.53	1.06	2.03	0.34	1.31	1.82	2.57
Only RL	2	98%	0.10	0.39	0.55	0.80	0.09	0.49	0.95	1.76	0.25	0.69	0.98	1.49

Key Insights

- MR.Q generalizes to every variation seen during training (e.g., exploring actions)
- It does not generalize out-of-distribution (OOD)
- OOD generalization requires hybrid control: policy + domain-specific controller

References

[1] C. Benjamins, T. Eimer, F. Schubert, A. Mohan, S. Döhler, A. Biedenkapp, B. Rosenhahn, F. Hutter, and M. Lindauer. Contextualize me – the case for context in reinforcement learning. In *EWRL*, 2023.
[2] S. Fujimoto, W.-D. Chang, E. Smith, S. S. Gu, D. Precup, and D. Meger. For sale: State-action representation learning for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 61573–61624. Curran Associates, Inc., 2023.
[3] S. Fujimoto, P. D'Oro, A. Zhang, Y. Tian, and M. Rabbat. Towards general-purpose model-free reinforcement learning. In *International Conference on Learning Representations*, 2025.
[4] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR, 10–15 Jul 2018.
[5] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.
[6] A. Hallak, D. D. Castro, and S. Mannor. Contextual Markov decision processes. *CoRR*, abs/1502.02259, 2015.