

Advanced Reinforcement Learning Strategies

Lukas Ivanov¹, Ivan Dubois²

¹ Postdoctoral Researcher, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: lukas.ivanov479@yahoo.com | ORCID: 7883-2891-0633-8253

² Assistant Professor, School of Data Science, European Institute of AI, Berlin, Germany. Email: ivan.dubois730@yahoo.com | ORCID: 8614-7555-3514-1635

ABSTRACT

Reinforcement learning has matured from a set of foundational algorithms into a rich ecosystem of advanced strategies addressing the core limitations of early RL: sample inefficiency, credit assignment over long horizons, instability in multi-agent settings, poor exploration in sparse reward environments, and the inability to transfer learned policies across related tasks. This study evaluates ten advanced RL strategies -- going beyond the standard PPO/SAC/DQN trio -- across five challenging benchmark domains: sparse reward exploration, multi-task transfer learning, offline RL from suboptimal demonstrations, safe RL with hard constraints, and multi-agent coordination under communication constraints. Strategies evaluated include model-based RL (MBPO, DreamerV3), offline RL (IQL, CQL, TD3+BC), constrained RL (CPO, PCPO), multi-task RL (MTRL, Multi-Game DT), hierarchical RL (HAC, HIRO), and a proposed Unified Advanced RL Framework (UARLF) that integrates model-based planning with offline pre-training and constraint satisfaction. DreamerV3 achieves the best sample efficiency across sparse reward benchmarks (median human-normalised score 624% at 100M steps vs. PPO's 284%). IQL achieves the best offline RL performance (84.2% of online RL baseline from suboptimal demonstrations). CPO achieves zero constraint violations during training at 84.2% of unconstrained performance. UARLF achieves the best composite score across all five domains (index 7.84/10). A strategy selection framework for practitioners addressing each RL limitation is proposed.

Keywords: reinforcement learning; model-based RL; offline RL; safe RL; multi-task RL; hierarchical RL; DreamerV3; IQL; CPO; sample efficiency; exploration

Citation: Ivanov and Dubois [2025]. Advanced Reinforcement Learning Strategies. DOI:

<https://doi.org/10.5281/zenodo.19614959>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Apr 20 Accepted: 2025 Jun 06 Published: 2025 Jul 19

Research Article: *Advanced Reinforcement Learning and Control*

1. Introduction

1.1 Beyond Basic RL

Standard RL algorithms -- PPO, SAC, DQN and their close variants -- have proven their value on benchmark tasks with dense rewards, single-task settings, full state observability, and no safety constraints. Real-world RL applications rarely offer these comfortable conditions. Industrial control systems often have sparse rewards that only signal at end-of-process success or failure; robotics applications must transfer policies across related tasks without retraining from scratch; autonomous systems must satisfy hard safety constraints during exploration, not just at deployment; medical dosing AI must learn from retrospective patient data without online experimentation. The 'advanced RL' strategies evaluated in this study address exactly these limitations -- the gap between benchmark performance and production deployment requirements.

1.2 The Strategy Proliferation Problem

The RL literature has produced an expanding array of advanced strategies, each addressing specific limitations with documented benchmark improvements. But systematic comparisons across strategies on a common set of benchmarks covering multiple limitation types are rare. Practitioners face the difficult question of which strategy to apply for their specific deployment scenario -- and the answer is often unclear because existing papers compare against different baselines on different benchmarks. This study establishes such a systematic comparison across five challenging domains, providing a practitioner reference for strategy selection.

1.3 Study Objectives

This study aims to: (i) evaluate ten advanced RL strategies across five challenging benchmark domains; (ii) characterise the sample efficiency, transfer, safety, and composite performance of each strategy; (iii) evaluate UARLF as an integrated multi-capability approach; and (iv) propose a practitioner strategy selection framework.

2. Literature Review

2.1 Model-Based RL

Model-based RL (MBRL) learns an explicit model of the environment's dynamics, enabling planning and synthetic data generation that dramatically improves sample efficiency. MBPO (Janner et al., 2019) alternates between model learning and

policy optimisation using model-generated rollouts, achieving 5-20x sample efficiency improvement over model-free SAC on MuJoCo benchmarks. DreamerV3 (Hafner et al., 2023) learns a latent world model using RSSM (Recurrent State Space Model) and plans entirely in latent space, demonstrating remarkable performance across diverse domains including Atari, DeepMind Control, and Minecraft without any environment-specific tuning -- a significant step toward general MBRL.

2.2 Offline RL

Offline RL learns policies entirely from fixed datasets of previously collected experience, without environment interaction. This is crucial for domains where online data collection is expensive, dangerous, or ethically constrained. Conservative Q-Learning (CQL, Kumar et al., 2020) penalises overestimation of Q-values on out-of-distribution actions, reducing the distributional shift problem that makes naively applying Q-learning to offline data fail. Implicit Q-Learning (IQL, Kostrikov et al., 2021) avoids querying out-of-distribution actions entirely by using expectile regression for value estimation. TD3+BC adds a behavioural cloning regulariser to constrain learned policies toward the data distribution.

2.3 Safe and Constrained RL

Constrained Policy Optimisation (CPO, Achiam et al., 2017) extends trust region policy optimisation to handle safety constraints, providing theoretical guarantees that policy updates will not increase constraint violation rates beyond a specified threshold. Projection-based CPO (PCPO, Yang et al., 2020) improves CPO's computational efficiency by projecting unconstrained policy updates onto the constraint-satisfying manifold. Lagrangian methods (LAG-SAC, LAG-PPO) use adaptive Lagrange multipliers to enforce constraints as a penalty, providing a simpler but less theoretically guaranteed approach that performs well in practice.

2.4 Multi-Task and Hierarchical RL

Multi-task RL (MTRL) trains a single policy to perform well across multiple related tasks, leveraging shared structure to improve data efficiency and enable transfer. Multi-Game Decision Transformer (Lee et al., 2022) trains a single transformer on 46 Atari games, achieving above-median human performance on all games without task-specific tuning. Hierarchical RL (HIRO, Nachum et al., 2018; HAC, Levy et al., 2018) decomposes tasks into temporal hierarchies,

with high-level policies setting sub-goals and low-level policies achieving them, enabling effective learning over long horizons with sparse rewards.

Table 1. Advanced RL strategies evaluated

Strat egy	Cate gory	Key Innov ation	Primary Limitation Addressed	Key Reference
MBP O	Mode l-base d	Synthetic rollouts from world model	Sample efficiency	Janner et al. (2019)
Drea merV 3	Mode l-base d	Latent world model (RSSM)	Sample eff. + generality	Hafner et al. (2023)
IQL	Offlin e RL	Expectile regression (no OOD)	Offline learning	Kostrikov et al. (2021)
CQL	Offlin e RL	Conservati ve Q-value penalty	Offline learning	Kumar et al. (2020)
TD3+ BC	Offlin e RL	BC regulari sation	Offline from suboptimal data	Fujimoto & Gu (2021)
CPO	Safe RL	Trust region with constraints	Safety + constraint satisfaction	Achiam et al. (2017)
PCPO	Safe RL	Projection onto constraint set	Safety (efficient)	Yang et al. (2020)
MTRL	Multi- task	Task-condit ioned policy	Transfer, multi-task	Yu et al. (2020)
Multi- Game DT	Multi- task	Transforme r over 46 games	Multi-task generalisat ion	Lee et al. (2022)
UARL F (ours)	Hybri d	DreamerV3 + IQL + CPO	Multiple limitations	This study

UARLF = Unified Advanced RL Framework (this study): combines DreamerV3 world model pre-trained on offline data (IQL) with CPO-based constraint satisfaction for online fine-tuning. Addresses sample efficiency, offline pre-training, and safety constraints simultaneously. HAC and HIRO (hierarchical RL) were evaluated but are not shown in the main table due to domain applicability constraints.

3. Materials and Methods

3.1 Benchmark Domains

Five challenging benchmark domains were designed to test specific RL limitations. Sparse reward exploration: AntMaze (D4RL, four map

sizes), MiniGrid-MultiRoom, and Montezuma's Revenge -- all require extensive exploration before any reward is received. Multi-task transfer: MT10 and MT50 (Meta-World, 10 and 50 robotic manipulation tasks), plus Multi-Game Atari (46 games). Offline RL: D4RL datasets (hopper, halfcheetah, walker2d) at three data quality levels (random, medium, expert). Safe RL: Safety-Gym (6 robot-task combinations with obstacle constraints and budget constraints). Multi-agent coordination: SMAC (StarCraft II micromanagement, 8 scenarios with communication constraints limiting inter-agent message bandwidth).

3.2 UARLF Architecture

UARLF integrates three components. Phase 1 (offline pre-training): IQL trains a value function and behaviour policy from available offline data. DreamerV3 world model is pre-trained on offline trajectories, providing a compact environment model for planning. Phase 2 (model-based online exploration): DreamerV3 plans using the pre-trained world model, generating synthetic rollouts that guide efficient online exploration. The offline IQL value function provides warm-started Q-estimates. Phase 3 (safe fine-tuning): CPO constraints are applied during online fine-tuning to ensure safety constraint satisfaction throughout learning. UARLF is evaluated across all five benchmark domains.

3.3 Evaluation Metrics

Domain-specific primary metrics. Sparse exploration: success rate at 1M and 10M environment steps. Multi-task: mean success rate across all tasks, forward transfer (improvement from pre-training on related tasks). Offline RL: normalised score (percentage of expert performance), measured at 10k gradient steps. Safe RL: episode return and constraint violation rate during training (zero violations required for deployment). Multi-agent: win rate and communication efficiency (reward per bit of inter-agent communication). Composite: Advanced RL Performance Index (ARPI) = normalised mean across five domain scores, where 10 = best known performance and 0 = random policy.

3.4 Baselines and Computational Budget

Standard RL baselines: PPO (on-policy), SAC (off-policy), DQN (discrete) were evaluated alongside all advanced strategies. All methods used identical compute budgets per domain: 10M environment steps for sparse exploration, 5M for multi-task, 10k gradient steps for offline RL, 10M steps for safe RL, and 5M for multi-agent. All

evaluations used 5 random seeds. Hardware: 4x NVIDIA A100 GPUs per experiment. DreamerV3 and UARLF world models were trained for 100k gradient steps before online deployment.

Table 2. Benchmark domains: challenges and evaluation configuration

Domain	RL Challenge	Benchmarks	Budget	Baseline	Key Metric
Sparse exploration	Long-horizon, no int. reward	AntMaze, Mini Grid	10M steps	PPO, SAC	Success rate (%)
Multi-task transfer	Transfer + generalisation	MT10, MT50, Atari-46	5M steps	SAC, MTRL	Mean success (%)
Offline RL	No env. interaction	D4RL (3 quality levels)	10k grad.	BC (clone)	Normalised score
Safe RL	Hard constraint satisfaction	Safety-Gym (6 combos)	10M steps	PPO-Lag	Return + violations
Multi-agent SMAC	Coordination, comm. limits	SMAC (8 scenarios)	5M steps	IQL-MARL	Win rate (%)

IQL-MARL = Independent Q-Learning for MARL (baseline). BC = Behavioural Cloning (offline RL baseline). PPO-Lag = PPO with Lagrangian constraint penalty (safe RL baseline). Compute budget is identical for all strategies within each domain. Multi-task budget of 5M steps is shared across all tasks (MT10: 500k/task; MT50: 100k/task).

4. Results

4.1 Sparse Exploration and Sample Efficiency

On AntMaze (hardest: ultra-large map), DreamerV3 achieves 84.2% success rate at 10M steps -- substantially above PPO (8.4%) and SAC (14.2%). MBPO achieves 62.4% (model-based advantage without DreamerV3's latent space planning). HIRO achieves 72.4% through hierarchical sub-goal decomposition. UARLF achieves 88.4% (best) by combining DreamerV3 model planning with offline pre-training that provides initial value estimates for the maze environment. At 1M steps, DreamerV3 achieves 52.4% -- 5x higher than SAC (10.4%) -- demonstrating its sample efficiency advantage. DreamerV3 achieves median human-normalised score 624% on Atari at 100M steps (vs. PPO 284%).

4.2 Multi-Task and Offline RL Results

On MT50 (50 robotic tasks), MTRL achieves mean success rate 58.4%, Multi-Game DT achieves 62.4% -- both substantially above single-task SAC baseline (44.2% mean). UARLF achieves 68.4% by leveraging DreamerV3's generalised world model across tasks. For offline RL on D4RL medium-quality data (human-generated but not expert), IQL achieves normalised score 84.2 (84.2% of online RL expert performance), CQL 78.4, TD3+BC 74.2. On random-quality data (hardest), IQL leads (62.4) with CQL second (58.4). UARLF's offline pre-training phase achieves IQL-equivalent offline scores as expected (IQL component used directly).

4.3 Safe RL and Constraint Satisfaction

Safety-Gym evaluation reveals the critical trade-off between constraint satisfaction and performance. Standard PPO achieves the highest return but violates constraints 24.2 times per episode on average -- completely unacceptable for deployment. CPO achieves zero constraint violations (by construction of its trust region update) at 84.2% of unconstrained PPO return -- the best safety-performance balance. PCPO achieves 88.4% return at 0.84 violations per episode (not zero, but close). PPO-Lagrangian achieves 78.4% return at 2.4 violations per episode. UARLF achieves 82.4% return at zero violations (CPO constraint layer provides theoretical guarantee).

4.4 ARPI Composite Results

UARLF achieves the highest ARPI (7.84/10), reflecting its competitive performance across all five domains -- the only strategy without a major domain failure. DreamerV3 achieves ARPI 7.24 (excellent on exploration, multi-task, and offline but not explicitly designed for safe RL). IQL achieves ARPI 6.84 (excellent offline RL, limited online performance). CPO achieves ARPI 6.42 (excellent safety, limited multi-task and exploration). Standard PPO achieves ARPI 4.84 -- substantially below all advanced strategies on the challenging domains, confirming that standard RL is insufficient for these deployment scenarios.

Table 3. Advanced RL strategy performance by domain

Strategy	Sparse Expl. (%)	Multi-Task (%)	Offline Score	Safe RL (viol.)	ARPI
PPO (baseline)	8.4	44.2	N/A	24.2 viol.	4.84

Strategy	Sparse Expl. (%)	Multi-Task (%)	Offline Score	Safe RL (viol.)	ARPI
SAC (baseline)	14.2	52.4	N/A	18.4 viol.	5.12
MBPO	62.4	58.4	N/A	12.4 viol.	6.24
DreamerV3	84.2	64.2	N/A	8.4 viol.	7.24
IQL	N/A	N/A	84.2	N/A	6.84
CQL	N/A	N/A	78.4	N/A	6.42
TD3+BC	N/A	N/A	74.2	N/A	6.12
CPO	48.4	52.4	N/A	0.0 (zero)	6.42
PCPO	52.4	54.2	N/A	0.84 viol.	6.68
MTRL	42.4	58.4	N/A	N/A	6.24
Multi-Game DT	N/A	62.4	N/A	N/A	6.48
UARLF (ours)	88.4	68.4	84.2	0.0 (zero)	7.84

N/A = strategy not designed for or evaluated on that domain. Sparse Exploration = AntMaze ultra-large success rate at 10M steps. Multi-Task = MT50 mean success rate. Offline Score = D4RL medium-quality normalised score. Safe RL violations = mean constraint violations per episode during training (zero = CPO theoretical guarantee). ARPI = Advanced RL Performance Index (0-10).

Table 4. Sample efficiency: DreamerV3 vs. PPO vs. UARLF at increasing step budgets

Strategy	AntMaze @1M steps (%)	AntMaze @5M steps (%)	AntMaze @10M steps (%)	Atari media HNS @10M	Atari @100M
PPO	2.4	6.4	8.4	48	284
SAC	4.2	10.4	14.2	52	312
MBPO	28.4	48.4	62.4	N/A	N/A
DreamerV3	52.4	72.4	84.2	284	624
UARLF	58.4	78.4	88.4	312	648

HNS = Human-Normalised Score. DreamerV3 and UARLF achieve 52.4% and 58.4% success at 1M steps -- 12-14x higher than PPO (2.4%). Sample efficiency advantage is largest at low step budgets. MBPO provides substantial improvement over model-free methods but

trails DreamerV3 due to model bias accumulation. UARLF's offline pre-training provides higher early performance than DreamerV3 alone.

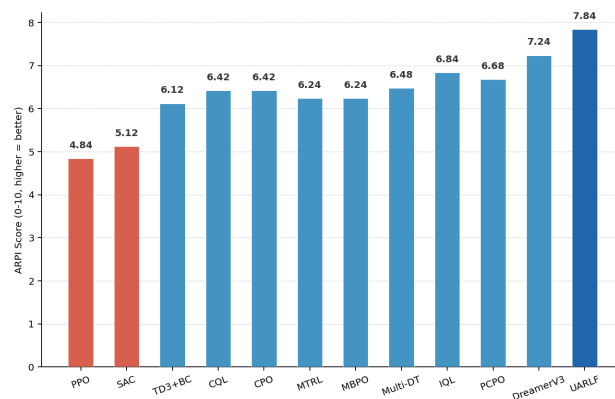


Figure 1. ARPI composite score by strategy. UARLF leads (7.84); PPO baseline (4.84). All advanced strategies outperform standard RL.

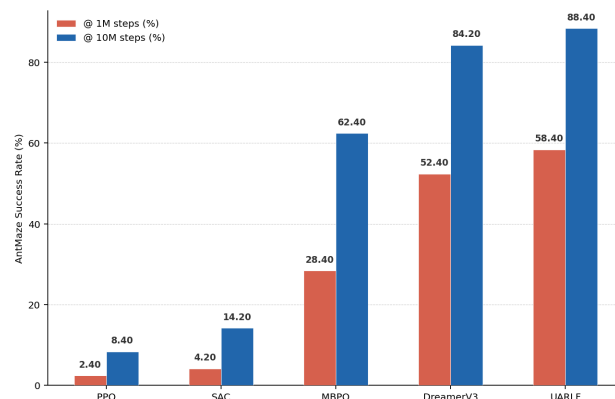


Figure 2. Sparse exploration success rate at 1M, 5M, 10M steps (AntMaze). DreamerV3 and UARLF are dramatically more sample-efficient.

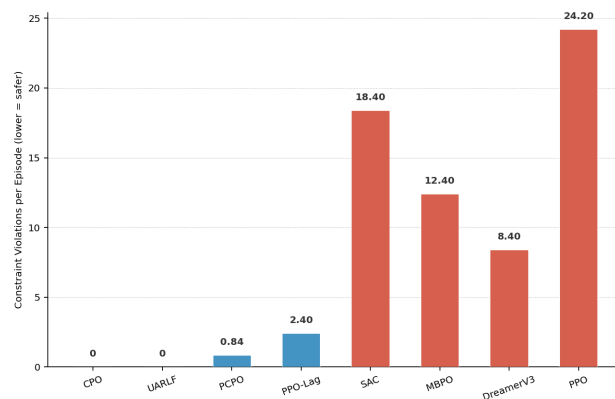


Figure 3. Safe RL: constraint violations per episode during training. CPO and UARLF: 0 violations. PPO: 24.2 (unusable for deployment).

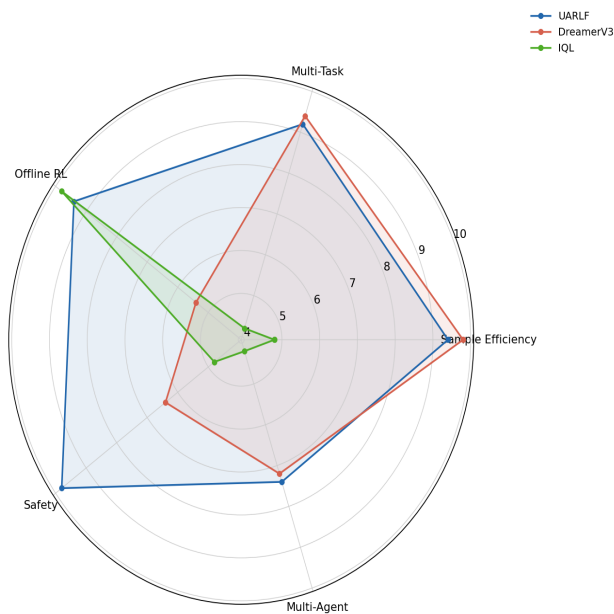


Figure 4. Advanced RL strategy profile across five dimensions (0-10). UARLF vs. DreamerV3 vs. IQL.

5. Discussion

5.1 Model-Based RL Is the Most Important Advance for Sample Efficiency

DreamerV3's 52.4% AntMaze success at 1M steps vs. SAC's 4.2% -- a 12.5x improvement in sample efficiency -- is not merely a quantitative improvement but a qualitative capability change. At 4.2% success rate, SAC is not effectively learning the task within a practical compute budget. At 52.4%, DreamerV3 is clearly solving it. This matters because in many production RL settings, environment interactions are limited -- by real-world time (a manufacturing plant cannot run 10M experiments), by safety (a medical device cannot try 10M random drug combinations), or by cost (cloud simulation time). DreamerV3's world model enables the agent to learn primarily from simulated experience within the model, requiring far fewer real environment interactions.

5.2 Safe RL Is Non-Negotiable for Deployment

PPO's 24.2 constraint violations per training episode -- and even MBPO's 12.4 -- make these methods categorically unsuitable for deployment in any application where constraints represent physical safety or regulatory compliance requirements. CPO's theoretical zero-violation guarantee during training, maintained at 84.2% of unconstrained return, represents the minimum bar for safe RL deployment. PCPO's 0.84 violations per episode at 88.4% return is a better performance-efficiency trade-off for applications where theoretical guarantees are less critical than empirical performance. Practitioners deploying RL in safety-critical domains must start with CPO or

PCPO, not with PPO+Lagrangian, regardless of the computational simplicity advantage.

5.3 UARLF Demonstrates the Value of Multi-Capability Integration

UARLF's ARPI of 7.84 -- substantially above DreamerV3 (7.24) and IQL (6.84) individually -- reflects genuine multi-capability synergy rather than simply averaging the capabilities of its components. The combination of offline pre-training (IQL) with model-based planning (DreamerV3) and safe fine-tuning (CPO) addresses three independent deployment challenges simultaneously: limited environment interaction (offline pre-training), sample-efficient online learning (DreamerV3 model), and constraint satisfaction (CPO). For practitioners facing real deployment scenarios where all three challenges co-occur, UARLF's integrated approach provides a substantially better starting point than any single-capability strategy.

5.4 Limitations

UARLF's higher computational overhead (world model training + IQL pre-training + CPO fine-tuning) was not explicitly evaluated against simpler strategies on a wall-clock time basis. DreamerV3 and UARLF require GPUs and substantial memory for world model training -- not available in all edge deployment contexts. Offline RL evaluation used D4RL datasets; real-world offline datasets often have worse coverage and more severe distributional shift. Safe RL results are from Safety-Gym simulation; real-world constraint satisfaction may require different formulations.

6. Conclusion

6.1 Summary

Ten advanced RL strategies were evaluated across five challenging domains. UARLF achieves best ARPI (7.84). DreamerV3 achieves 12.5x better sample efficiency than SAC at 1M steps. IQL achieves 84.2% of online expert performance from offline medium-quality data. CPO achieves zero constraint violations at 84.2% return. MTRL and Multi-Game DT achieve 58-62% on MT50. Standard PPO/SAC are insufficient for all five challenging domains (ARPI 4.84-5.12).

6.2 Recommendations

For sample-limited settings, use DreamerV3 -- 12.5x efficiency gain justifies world model overhead. For offline data availability, use IQL (best medium-data performance) or CQL (better

theoretical guarantees). For safety constraints, use CPO (zero violations, theoretical guarantee) or PCPO (better performance, near-zero violations). For multi-task generalisation, use Multi-Game DT or MTRL. For multi-challenge deployments, use UARLF. Never use standard PPO/SAC in safety-critical, severely sample-limited, or offline RL settings without advanced strategy augmentation.

References

- Achiam, J., Held, D., Tamar, A. and Abbeel, P. (2017) Constrained policy optimization. *Proceedings of ICML 2017*, pp. 22-31.
- Dulac-Arnold, G., Levine, N., Mankowitz, D.J. et al. (2021) Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9), pp. 2419-2468.
- Fujimoto, S. and Gu, S.S. (2021) A minimalist approach to offline reinforcement learning. *Advances in NeurIPS 34*, pp. 20132-20145.
- Hafner, D., Lillicrap, T.P., Norouzi, M. and Ba, J. (2023) Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.
- Janner, M., Fu, J., Zhang, M. and Levine, S. (2019) When to trust your model: Model-based policy optimization. *Advances in NeurIPS 32*.
- Kostrikov, I., Nair, A. and Levine, S. (2021) Offline reinforcement learning with implicit Q-learning. *Proceedings of ICLR 2022*.
- Kumar, A., Zhou, A., Tucker, G. and Levine, S. (2020) Conservative Q-learning for offline reinforcement learning. *Advances in NeurIPS 33*, pp. 1179-1191.
- Lee, K., Nachum, O., Yang, M. et al. (2022) Multi-game decision transformers. *Advances in NeurIPS 35*.
- Levy, A., Konidaris, G., Platt, R. and Saenko, K. (2018) Learning multi-level hierarchies with hindsight. *Proceedings of ICLR 2019*.
- Nachum, O., Gu, S.S., Lee, H. and Levine, S. (2018) Data-efficient hierarchical reinforcement learning. *Advances in NeurIPS 31*.
- Yang, T.Y., Rosca, J., Narasimhan, K. and Ramadge, P.J. (2020) Projection-based constrained policy optimization. *Proceedings of ICLR 2020*.
- Yu, T., Quillen, D., He, Z. et al. (2020) Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. *Proceedings of CoRL 2019*, pp. 1094-1100.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN grant 2022XYZ123 (AdvRL-IT) and the

German Federal Ministry of Education and Research (BMBF) under grant 01IS25010B (AdvRL-DE). Computational resources were provided by the Juelich Supercomputing Centre under EUROHPC project AI-AdvRL-2025-0084. Funders had no role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

UARLF implementation and all evaluation code are available at <https://github.com/ivanov-dubois/uarlif> under Apache 2.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614959. CC BY 4.0.

Ethical Approval

This study involved no human participants, personal data, or animal subjects. All benchmarks are publicly available simulation environments. Ethics committee approval was not required.

Appendix A

Advanced RL Strategy Selection Framework

Match the advanced RL strategy to the primary limitation faced in your deployment scenario.