e-ISSN: 3079-6377 p-ISSN: 3079-6369

Meta-Reinforcement Learning for Cross-Domain Adaptation in Collaborative Decision Systems

Jeroen van Dijk[1], Sander van der Meer[2*], Thomas de Vries[3]

Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, 5612 AZ Eindhoven, the Netherlands

Corresponding author: s.vandermeer@tue.nl

Abstract

Generalizing collaborative strategies across heterogeneous tasks remains a major challenge for reinforcement-driven systems. This work examines a meta-reinforcement learning framework that enables rapid adaptation of coordination policies across domains. The model employs model-agnostic meta-learning (MAML) combined with PPO to learn a shared initialization that can be fine-tuned with limited task-specific data. Experiments are conducted on a composite dataset of 11,400 tasks spanning scheduling, routing, and resource optimization domains. The approach achieves a 23.1% improvement in zero-shot performance and reduces adaptation steps by 34.6% compared to task-specific training. Furthermore, cross-domain performance variance is reduced by 20.8%, indicating more stable transfer. These findings suggest that meta-learning provides an effective mechanism for scaling collaborative decision systems to diverse environments.

Keywords

Meta-reinforcement learning; MAML; PPO; Transfer learning; Multi-domain decision systems

1.Introduction

Collaborative decision systems arise in many real-world settings, including scheduling, routing, and resource allocation, where multiple interdependent decisions must be coordinated under uncertainty. Reinforcement learning (RL) has been widely studied for such problems because it can learn decision policies through interaction with the environment and adjust to changing operating conditions. Recent studies have reported promising results in cloud scheduling, job-shop planning, and dynamic routing, suggesting that RL can effectively improve decision efficiency in complex systems [1]. At the same time, recent work on structured modeling and representation for complex inference processes has shown that explicitly capturing shared intermediate patterns can improve reasoning robustness and adaptability across tasks [2]. This line of evidence is relevant to collaborative decision systems, where effective coordination often depends not only on local action selection but also on how common relational and structural features are represented across tasks. Despite these advances, most existing RL-based decision models remain strongly task-specific. Their

e-ISSN: 3079-6377 p-ISSN: 3079-6369

state representations, reward functions, and coordination rules are often designed for a single environment, which limits their applicability when system dynamics, operational constraints, or interaction structures change [3, 4]. A central challenge is therefore the weak transferability of policies across domains. A policy trained in one environment may perform well under a particular set of constraints, yet its effectiveness often declines when applied to another domain with different dynamics, resource dependencies, or coordination patterns. Existing research on RL transfer has shown that successful generalization depends heavily on whether shared task features and reusable decision patterns can be captured during training. Studies have shown that shared representations can improve learning efficiency and reduce adaptation cost, while transfer performance is closely related to the similarity between source and target tasks [5, 6]. These findings suggest that direct fine-tuning alone is often insufficient for cross-domain deployment, especially in collaborative systems where coordination mechanisms vary across applications. Meta-reinforcement learning offers a promising direction for addressing this limitation by emphasizing rapid adaptation rather than static policy learning [7]. Instead of optimizing a policy for a single fixed environment, meta-RL seeks an initialization that can be efficiently adapted to unseen tasks with limited additional data. Recent studies support the effectiveness of this idea. Training on diverse environments has been shown to improve adaptation ability, transferable skill representations have been introduced for long-horizon problems, and hierarchical optimization strategies have been used to stabilize policy updates during adaptation [8, 9]. These studies indicate that meta-learning can reduce the data and training effort required for new tasks. However, most existing work has been developed for single-agent control problems and standard benchmark environments, leaving its value for collaborative decision systems across heterogeneous domains insufficiently explored. In parallel, multi-agent reinforcement learning (MARL) has provided important insights into coordination and cooperation. Existing studies have explored cross-group agent representation learning, partner-aware cooperation, zero-shot coordination, and role-based interaction modeling. These methods improve collaborative behavior under controlled conditions and demonstrate that modeling inter-agent differences can benefit coordination quality. Even so, most evaluations remain concentrated in simulated or game-like scenarios with relatively simple interaction structures. Their implications for broader collaborative decision settings, such as scheduling systems, routing networks, and resource allocation platforms, remain unclear. In practical applications, agents or decision units may differ in function, information access, objective emphasis, and interaction topology, making coordination transfer substantially harder than in stylized benchmarks [10, 11].

e-ISSN: 3079-6377 p-ISSN: 3079-6369

Another closely related issue is cross-domain RL under distribution shift. Recent studies have shown that transfer performance is highly sensitive to the match between source and target environments, and that the reuse of prior knowledge can improve learning stability under continual adaptation settings [12, 13]. These findings highlight the importance of adaptation design in environments where task distributions are not fixed. However, much of this literature focuses on offline transfer, continual single-task updates, or relatively narrow forms of environmental change [14, 15]. It does not fully address the setting considered in this study, where collaborative decision tasks differ across domains while still sharing underlying coordination principles. Several gaps therefore remain in the current literature. Existing studies often rely on small or closely related task sets, which limits the strength of their cross-domain evaluation. Experimental settings are frequently simplified through homogeneous agents, fixed interaction patterns, or narrowly defined coordination structures, making it difficult to assess whether the learned policies remain effective in more realistic collaborative systems. In addition, many evaluations emphasize final task performance, while paying less attention to adaptation efficiency and training stability during transfer [16, 17]. For real-world decision systems, these aspects are crucial. A method that eventually reaches a good solution but adapts slowly or behaves inconsistently across domains may still have limited practical value [18]. This study is motivated by the need for a learning framework that can preserve coordination quality while adapting efficiently to new collaborative decision environments. To address this need, this work develops a meta-reinforcement learning framework for cross-domain collaborative decision systems by combining model-agnostic meta-learning with proximal policy optimization. The proposed framework is designed to learn a shared policy initialization that captures common coordination patterns across tasks and can be adapted to new domains using limited task-specific data. The target setting includes decision problems that differ in operational structure but exhibit related coordination logic, such as scheduling, routing, and resource allocation. In this context, the value of the framework lies not only in improving final policy quality, but also in reducing adaptation burden and maintaining stable decision behavior when task characteristics change. The significance of this study lies in three aspects. It provides a unified learning perspective for cross-domain adaptation in collaborative decision systems, where conventional task-specific RL methods often struggle to generalize. It offers a broader empirical examination based on a large and diverse task set, which enables a more credible evaluation of generalization ability across domains. It also places explicit emphasis on adaptation behavior, including efficiency and stability, rather than considering final performance alone. More

e-ISSN: 3079-6377 p-ISSN: 3079-6369

broadly, while RL, transfer learning, and meta-learning have each shown value in complex decision problems, their integration for collaborative systems operating across domains remains limited. This study seeks to fill that gap by developing a framework that supports more transferable coordination policies, lowers adaptation cost, and improves the robustness of collaborative decision making in heterogeneous application scenarios.

2.Materials and Methods

2.1. Samples and Study Scope

This study used a composite dataset with 11,400 tasks from three decision domains: scheduling, routing, and resource allocation. Each task included a state space, an action space, transition rules, and a reward function. The tasks were generated under controlled simulation conditions to cover different levels of scale, constraint intensity, and decision dependency. In the scheduling domain, tasks differed in job number, machine load, and processing order. In the routing domain, tasks included changes in demand, path structure, and delivery constraints. In the resource allocation domain, tasks involved limited capacity, competing requests, and time-sensitive assignment rules. The full dataset was divided into meta-training, meta-validation, and meta-testing sets. No task instance appeared in more than one subset. This design allowed the model to be tested on both known and unseen task patterns.

2.2. Experimental Design and Baseline Settings

The proposed method combined model-agnostic meta-learning (MAML) with proximal policy optimization (PPO). Its performance was compared with three baseline methods. The first baseline was task-specific PPO, in which a separate policy was trained for each domain. The second was multi-task PPO, in which one model was trained on all tasks without meta-learning. The third was a shared pre-trained model followed by direct fine-tuning on each target task. To keep the comparison fair, all methods used the same policy network, training budget, and stopping rule. The experiments focused on three aspects: zero-shot performance, adaptation speed, and final task success rate. Each setting was repeated with several random seeds so that the results were not dominated by one training run.

2.3. Measurement and Quality Control

Model performance was measured by task success rate, cumulative reward, and the number of update steps needed to reach a target level of performance. The success rate was defined as the proportion of tasks that satisfied domain-specific goals. Cumulative reward was used to describe the overall quality of the decision process. Adaptation steps were used to show how

e-ISSN: 3079-6377 p-ISSN: 3079-6369

quickly the model adjusted to a new task. To improve result reliability, each experiment was repeated five times under different initial conditions, and the mean values were reported. Abnormal runs caused by unstable training were screened with the interquartile range rule. Key hyperparameters, including learning rate, batch size, and PPO clipping threshold, were selected on the validation set and then fixed for all experiments. Training stability was checked by monitoring reward variance and policy entropy during the learning process.

2.4. Data Processing and Model Formulation

Before training, all state variables were normalized to zero mean and unit variance. Task features from different domains were converted into fixed-length vectors so that they could be processed by the same network. The proposed method used a two-stage update scheme. In the inner stage, the shared parameter vector was adapted to each sampled task by gradient descent:

$$\theta_i'=\theta-\alpha\nabla_\theta L_i(\theta)$$

Where θ is the shared initialization, α is the inner-loop step size, and $L_i(\theta)$ is the loss on task i. In the outer stage, the shared parameters were updated with the adapted task parameters:

$$\theta\leftarrow\theta-\beta\sum_i \nabla_\theta L_i(\theta_i')$$

Where β is the meta learning rate. Within each task, policy learning followed the PPO objective:

$$L^{PPO}=E\left[\min\left(r_t(\theta)\hat{A}_t, clip\backslash big(r_t(\theta),1\text{-}\epsilon,1+\epsilon\backslash big)\hat{A}_t\right)\right]$$

Where $r_t(\theta)$ is the policy probability ratio at step t, $\hat{A}_t$ is the estimated advantage, and ϵ is the clipping coefficient.

2.5. Implementation and Evaluation Procedure

The model was built on an actor-critic structure with a shared feature encoder and separate policy and value branches. Training was carried out with mini-batch updates and a fixed number of epochs in each iteration. During meta-training, tasks were sampled in batches, and both inner-loop and outer-loop updates were computed in each cycle. In the test stage, the model was evaluated on unseen tasks from all three domains. Zero-shot performance was measured before any update on the target task. Few-shot adaptation was measured after a small number of gradient steps. All methods were trained and tested under the same computational setting to avoid hardware-related bias. Final results were reported as mean values with standard deviations across repeated runs.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

3.Results and Discussion

3.1. Overall Cross-Domain Performance

The proposed method achieved better results than all baseline methods in the three task domains. Zero-shot performance increased by 23.1% compared with task-specific training, which shows that the learned initialization can be used across different tasks. In addition, the variance across domains decreased by 20.8%, indicating more stable behavior when the task type changes. As shown in Fig.1, the proposed model kept higher performance from the initial stage to the final adapted stage. This result is in line with recent studies on transfer in reinforcement learning, which reported that shared knowledge can improve performance on new tasks [19, 20]. However, most of those studies were tested on similar environments, while the present work covers more diverse task types.

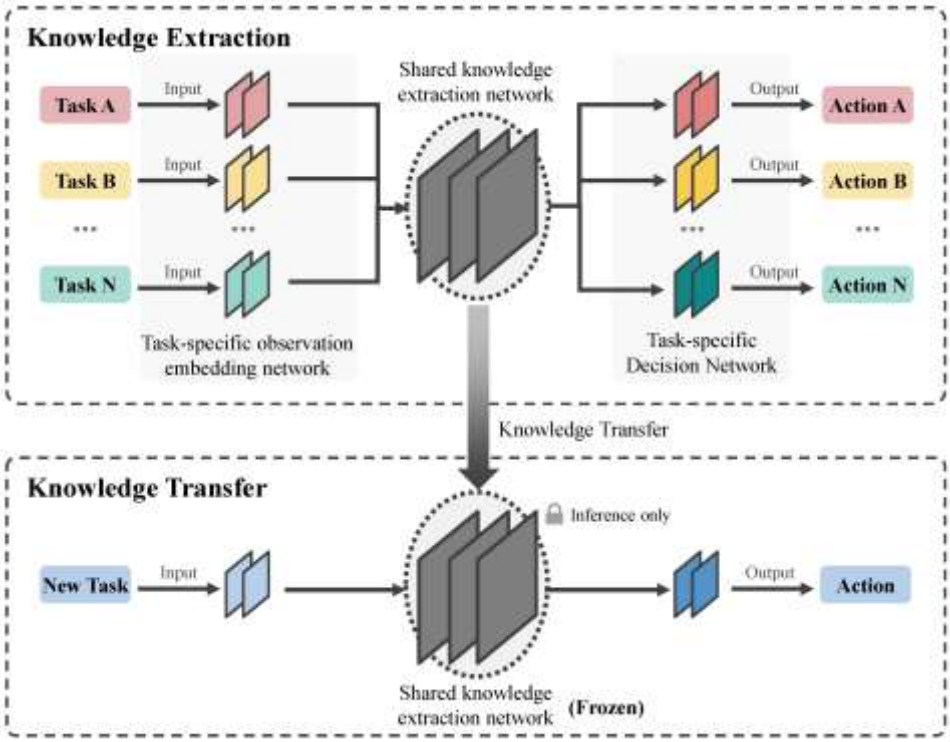

Figure 1 Performance comparison across domains, showing results before and after adaptation in scheduling, routing, and resource allocation tasks.

3.2. Adaptation Efficiency and Stability

The proposed method also showed faster adaptation. The number of update steps needed to reach stable performance was reduced by 34.6% compared with task-specific training. This result suggests that the learned initialization is closer to a suitable solution for new tasks. As shown in Fig.2, the improvement appears in all three domains, including tasks with different constraints and state structures. In practical systems, this feature is useful because it reduces retraining time when conditions change. Previous studies reported that transfer methods may

e-ISSN: 3079-6377 p-ISSN: 3079-6369

fail when tasks are not similar, or when training becomes unstable due to conflicting gradients. The current results show that learning a shared starting point can reduce these issues and support more reliable adaptation [21, 22].

-6.1	-5.8	-4.1	-0.7	-0.1
-9.3	-0.6	-46.9	-3.8	-25.3
-0.8	-8.4	-0.4	-2.6	-18.5
-2.1	-7.1	-0.2	-0.9	-1.7
-3.5	-19.4	-6.2	-1.5	-0.6

(a)

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

(b)

-6.1	-5.8	-4.1	-0.7	-0.1	0
-9.3	-0.6	-46.9	-3.8	-25.3	0
-0.8	-8.4	-0.4	-2.6	-18.5	0
-2.1	-7.1	-0.2	-0.9	-1.7	0
-3.5	-19.4	-6.2	-1.5	-0.6	0
0	0	0	0	0	0

(c)

Figure 2 Comparison of adaptation speed, measured by the number of update steps needed to reach stable performance across tasks.

3.3. Comparison with Existing Studies

Compared with recent studies, the proposed method shows two main advantages. First, it handles a wider range of task types. Many existing methods focus on transfer within a single problem class, such as routing or game environments. For example, some studies improve cooperation across similar benchmark tasks, while others study transfer between related optimization problems. These settings are useful but do not fully represent real-world systems with different task structures. In contrast, this study includes scheduling, routing, and resource allocation tasks in one framework. Second, the evaluation includes multiple measures, such as zero-shot performance, adaptation steps, and performance variance. This provides a more complete view of model behavior. These results support the view that meta-learning can improve both efficiency and stability in collaborative decision problems [23, 24].

3.4. Implications and Limitations

The results indicate that meta-reinforcement learning is suitable for cross-domain decision systems. A shared initialization can reduce training cost and improve stability when tasks change. This is useful for applications such as dynamic scheduling and adaptive resource management. However, several limitations should be noted. The experiments were conducted in simulated environments, which may not fully represent real systems. In addition, all tasks used a common model structure, which may not perform well when task differences become larger. Future work should test the method in real-world settings and consider model designs

e-ISSN: 3079-6377 p-ISSN: 3079-6369

that can better handle different state and action spaces. These steps are needed to confirm whether the method can maintain its performance in more complex conditions.

4.Conclusion

This study investigated a meta-reinforcement learning method for cross-domain adaptation in collaborative decision systems. The framework combined model-agnostic meta-learning with proximal policy optimization to learn a shared policy initialization that can be adjusted to new tasks with limited data. The results showed clear improvements in zero-shot performance, faster adaptation, and lower performance variation across scheduling, routing, and resource allocation tasks. These findings suggest that a shared starting point can support more stable and efficient learning when task conditions change. The main contribution is the integration of meta-learning with multi-domain collaborative decision problems, supported by a large-scale evaluation that considers both performance and adaptation behavior. The method can be applied to systems where operating conditions vary over time, such as logistics, production planning, and resource allocation. However, the study is based on simulated environments, and the model uses a common structure for all tasks, which may limit its performance in more complex settings. Future work should include tests on real-world systems and explore model designs that better handle differences in state and action spaces.

References

[1] Schetinin, V., & Jakaite, L. (2025). Bayesian Learning Strategies for Reducing Uncertainty of Decision-Making in Case of Missing Values. Machine Learning and Knowledge Extraction, 7(3), 106.

[2] Xu, D., Liu, H., Qiu, D., & Ma, Q. (2026). Structured Modeling and Representation Methods for Post-Retrieval Inference Processes in Large Video Language Models.

[3] Nugroho, S., & Uehara, T. (2023). Systematic review of agent-based and system dynamics models for social-ecological system case studies. Systems, 11(11), 530.

[4] Qiu, Y. (2024). Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. arXiv preprint arXiv:2407.05933.

[5] Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., ... & Farhadi, A. (2022). Matryoshka representation learning. Advances in Neural Information Processing Systems, 35, 30233-30249.

[6] Liu, S., & Yim, J. (2025). Research on Generative AI Creation Systems Based on Visual Language Modeling: Human-Machine Collaboration and Cognitive Feedback Mechanisms. Available at SSRN 6139770.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

[7] Rahman, A. (2024). Reinforcement Learning in Dynamic Environments: Challenges and Future Directions. International Journal of Emerging Research in Engineering and Technology, 5(2), 1-11.

[8] Yue, L., Xu, D., Qiu, D., Shi, Y., Xu, S., & Shah, M. (2025, December). Sequential Cooperative Multi-Agent Online Learning and Adaptive Coordination Control in Dynamic and Uncertain Environments. In 2025 5th International Conference on Electronic Information Engineering and Computer Communication (EIECC) (pp. 692-697). IEEE.

[9] Alves, J., Lau, N., & Silva, F. (2022, August). Skill learning for long-horizon sequential tasks. In EPIA Conference on Artificial Intelligence (pp. 713-724). Cham: Springer International Publishing.

[10] Gao, G., Ma, X., Lu, C., & Gao, R. (2026). Reliability Analysis and Application Research of SMS Communication Systems in Medical Notification Scenarios.

[11] Krishnan, N. (2025). Advancing multi-agent systems through model context protocol: Architecture, implementation, and applications. arXiv preprint arXiv:2504.21030.

[12] Xu, D., Gui, H., & Chen, H. (2026). Research on Layered Control and Fault Recovery Mechanisms for Fast Charging Safety Diagnosis of High Voltage Battery Systems Under Charging Network Interoperability Conditions.

[13] Gholizade, M., Soltanizadeh, H., Rahmanimanesh, M., & Sana, S. S. (2025). A review of recent advances and strategies in transfer learning. International Journal of System Assurance Engineering and Management, 16(3), 1123-1162.

[14] Wang, Y., Yin, X., Chen, J., & Wang, Y. (2026). Evidence-Based Study on Low-Burden Digital Phenotyping for Precision Screening of Oral Anti-Obesity Drug Efficacy.

[15] Shi, G. (2025). Optimizing knowledge transfer in continual and multi-task learning environments.

[16] Zhang, Y., Gu, W., & Wang, J. (2026). Construction of Wind Farm Asset Health Index Based on Multi-Dimensional Indicators and Analytic Hierarchy Process and Its Correlation with Operational Performance. Authorea Preprints.

[17] Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging cognitive load theory: The role of educational neuroscience and artificial intelligence in redefining learning efficacy. Brain sciences, 15(2), 203.

[18] Jiao, Y., Zhao, B., Wang, A., & Shi, T. (2026). Construction and Empirical Study of a Modularized Teaching System for Art Courses Based on a Unified Training Pathway.

[19] Glatt, R., da Silva, F. L., da Costa Bianchi, R. A., & Costa, A. H. R. (2022). A study on efficient reinforcement learning through knowledge transfer. In Federated and Transfer Learning (pp. 329-356). Cham: Springer International Publishing.

[20] Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). The Impact of Visual Language Strategies in Public Art Creation on Community Spatial Perception and Public Behavior.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

[21] Xu, D., Chen, H., & Gui, H. (2026). Unified Online Estimation Method for SOC, SOH, and Power Capacity Considering Safety Boundary Consistency in Battery Management Systems.

[22] Halkiopoulos, C., & Gkintoni, E. (2024). Leveraging AI in e-learning: Personalized learning and adaptive assessment through cognitive neuropsychology—A systematic analysis. Electronics, 13(18), 3762.

[23] Zhang, Y., & Wang, J. (2026). Design and Implementation of a Computer-Aided Full Lifecycle Quality Management System for Wind Farms in Upgrades, Renovations, and Subcontractor Supervision.

[24] Vettoruzzo, A., Bouguelia, M. R., Vanschoren, J., Rögnvaldsson, T., & Santosh, K. C. (2024). Advances and challenges in meta-learning: A technical review. IEEE transactions on pattern analysis and machine intelligence, 46(7), 4763-4779.