Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38108

Robot Learning Mechanisms and Frontier Paradigms under the Dual Challenges of Few-Shot Learning and Sparse Rewards

Ruiqing Feng

Faculty of Science and Engineering, University of Nottingham, Ningbo, China
ssyrf1@nottingham.edu.cn

Abstract. The amalgamation of Embodied AI with physical robot learning constitutes a pivotal route toward the realization of Artificial General Intelligence (AGI). Robot learning has important application value in real complex tasks. However, the interweaving of extremely high interactive sampling cost and sparse feedback in the real physical world constitutes a "double curse" that hinders robot autonomous skill learning. In the existing research on robot learning, there are a large number of reviews on Robot Learning under few-shot and sparse rewards. However, few scholars have systematically reviewed Robot Learning under the dual effects of few-shot and sparse rewards. This essay systematically discusses the technical trajectory and paradigm shift of solving this dual difficulty. Analyzes the failure mechanism of traditional reinforcement learning (RL) under few-shot and sparse rewards. Sorted out solutions based on three directions: Rapid Generalization (Meta RL, Few-Shot IL), Intrinsic Exploration (HER, Intrinsic Motivation), and Cognitive World Modeling (World Models, Offline RL, VLA embodied model). Specifically, Rapid Generalization focuses on addressing the few-shot learning problem, Intrinsic Exploration targets the sparse reward challenge, and Cognitive World Modeling aims to jointly resolve both issues through integration. In addition, this essay also constructs a systematic taxonomy comparison framework and looks forward to future research directions such as security exploration and embodied alignment.

Keywords: robot learning, few-shot, sparse reward, world model, embodied basic model

1. Introduction

As artificial intelligence moves from virtual space to physical entities, embodied intelligence systems require robots to have the ability to perform autonomous perception, decision-making, and control in complex and unstructured environments. The trial-and-error mechanism based on Deep Reinforcement Learning (DRL) has made remarkable achievements in chess games, virtual simulation, and other fields, but landing on real physical robots faces severe challenges.

In the real physical environment, robot learning faces two core pain points. Data scarcity (Few-shot) because of high sample cost and sparse environmental feedback (Sparse Reward). Few-shot and sparse reward not only exist independently, but also constitute a vicious circle of mutual

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38108

deterioration. Sparse reward signals cause agents to carry out massive random exploration to reach the reward state, but the high cost of physical sampling strictly limits the number of explorations of agents, which ultimately leads to the complete failure of policy optimization.

2. Failure mechanism of traditional reinforcement learning

Traditional reinforcement learning is a kind of reinforcement learning algorithm based on tabular methods and linear approximation, but it has many limitations. Firstly, the application of traditional reinforcement learning may lead to dimensional disasters. Once the status/action space is slightly larger, the table cannot be stored. Secondly, in traditional discrete action Q-Learning, the update of the target value depends on

$$\max_{a'} Q(s', a') \tag{1}$$

where max represents Maximum, a' represents Next Action, $Q(,)$ represents Q-Function, s' represents the next state.

The traditional Actor-Critic paradigm (such as DDPG) attempts to extend Q-Learning to the continuous action space of robots, using $\pi_\phi(s')$ to replace operation max,

$$y = r + \gamma Q_\theta(s', \pi_\phi(s')) \tag{2}$$

where y represents TD Target, r represents Reward, γ represents Discount Factor, s' represents Next State, ϕ represents Policy Parameters, θ represents Critic Parameters, $Q(,)$ represents Q-Function, $\pi_\phi(s')$ represents Next Action.

But it is proven that as long as there is an approximation error in Critical, the Actor will actively look for actions with an overestimated Critical [1]. Lastly, traditional reinforcement learning is difficult to deal with complex perceptual inputs, such as images, voice, and other original high-dimensional signals. In the recorded data from the comparison of Fetch Robot grasping tasks, the accuracy of the traditional model without randomization is extremely low in the real visual background, which directly proves that the simulation training strategy of the traditional simulation is not available on the real machine [2].

3. Paradigm classification

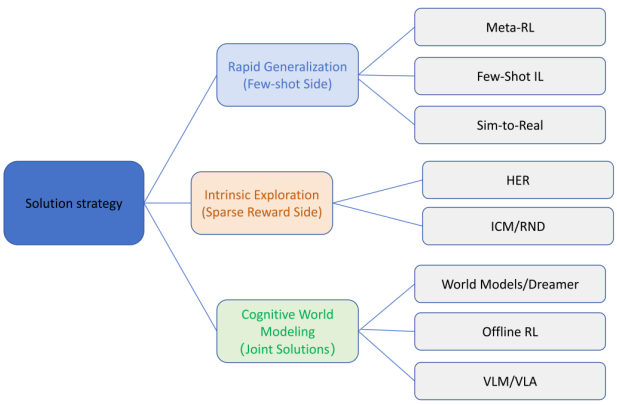


Figure 1. Forming three mainstream paradigms (photo/picture credit: original)

DOI: 10.54254/2755-2721/2026.38108

In order to break the vicious circle mentioned above, this essay has proposed solutions from different dimensions, forming three mainstream paradigms in Figure 1.

4. Rapid generalization (data efficiency side)

Rapid generalization enables the intelligent agent to rapidly adapt to new tasks with minimal new data through meta-learning or a small amount of demonstrations. Instead of blindly exploring in the target task without reward, it uses prior knowledge to narrow the strategy search space.

4.1. Rapid adaptation based on meta reinforcement learning (meta RL)

Meta-RL advocates "Learning to Learn". In the meta-training phase, agents accumulate experience on multiple associated task distributions $p(\mathcal{T})$ with intensive rewards or easy exploration. In the testing phase, only a few trial-and-error samples are needed to adjust the strategy for new tasks. However, the previous meta RL methods are all on-policy, and the sample utilization rate is still very low. PEARL makes meta RL off-policy and uses probabilistic context variables for task inference, which greatly improves the sample utilization rate when the robot faces unknown tasks [3].

4.2. Few-shot/one-shot imitation

When the reward is extremely sparse, 1 to 10 human expert demonstration trajectories are the most direct strategic constraints. One-Shot Imitation Learning uses attention and context encoders to enable robots to map actions to new object arrangements by watching a human demonstration video only once [4].

4.3. Sim-to-real

The sampling in the simulation environment is almost cost-free, and the reward is easy to shape. Randomize physical attributes such as friction, quality, and visual texture in the simulator through Domain Randomization and solve the exploration phase of extremely costly samples in the simulation [2]. Finally, only a small amount of real physical data is needed to fine-tune the policy.

5. Intrinsic exploration (sparse reward side)

If there is a lack of external prior, the environment reward or exploration mechanism must be changed so that agents can still obtain a gradient when there is a lack of external reward.

5.1. Hindsight experience replay (HER)

HER is a milestone method to solve the sparse reward target conditioning task [5]. Even if an attempt fails to reach the original target g, the agent must reach a certain actual state s_T . If it reset the goal of this failed attempt to $g' = s_T$, this track is a perfect success track for learning "how to reach s_T ". The transformation mechanism is as follows:

$$\mathcal{D}_{replay} \leftarrow \mathcal{D}_{replay} \cup \{(s_t, a_t, r_t, s_{t+1}, g)\} \cup \{(s_t, a_t, r'_t, s_{t+1}, g')\} \tag{3}$$

where $r'_t = \mathcal{R}(s_t, a_t, g')$, $\mathcal{D}_{replay}$ is Replay buffer/Experience replay memory, $(s_t, a_t, r_t, s_{t+1}, g)$ is Original transition tuple, $(s_t, a_t, r'_t, s_{t+1}, g')$ is Relabeled transition tuple

DOI: 10.54254/2755-2721/2026.38108

Through this mechanism, HER converts 100% of the original sparse failure samples into data containing positive reinforcement signals, completely solving the sparse reward problem in goal-oriented robot control (such as Reach, Push, Pick-and-Place). The subsequent Curriculum HER further improved the convergence speed in high-dimensional complex tasks by dynamically selecting virtual targets with progressive difficulty [6].

5.2. Intrinsic motivation & curiosity

When the environment has no goal definition at all, the Intrinsic Reward can be used to drive the robot to explore. Inspired by cognitive psychology, the Intrinsic Curiosity Module (ICM) regards the prediction error of the environmental dynamics model as an internal reward [7].

$$r_t^{int} = \frac{\eta}{2} \left\| \hat{\phi}\left(s_{t+1}\right) - \phi\left(s_{t+1}\right) \right\|_2^2 \tag{4}$$

Where $\hat{\phi}(s_{t+1})$ is the predicted state and $\phi(s_{t+1})$ is the real coding state. Agents tend to explore areas that surprise their own dynamic prediction models. In addition, Random Network Decomposition (RND) overcomes the bottleneck of exploration in complex high-dimensional environments by taking the prediction difference between a fixed random network and a trainable network as an uncertainty measurement [8].

6. Joint solutions: cognitive world modeling

In recent years, the most significant progress in the field of embodied intelligence is the joint solution of "few-shot" and "sparse rewards".

6.1. Latent imagination: virtual trial-and-error based on world models

Model-based RL attempts to learn the environmental transition dynamics model $p_\phi(s_{t+1}|s_t, a_t)$ which describes the distribution of next state s_{t+1} after executing action a_t in state s_t. DreamerV3 and other world models map high-dimensional visual observations to a low-dimensional latent space and establish a cyclic state space model (RSSM) in the latent space [9].

Representation Model: $\sim q_\theta(z_t|z_{t-1}, a_{t-1}, x_t)$ which infers the latent state z_t from the current latent state z_{t-1}, the previous action a_{t-1}, and the current observation x_t.

Transition Model: $\hat{z}_t \sim p_\phi(\hat{z}_t|z_{t-1}, a_{t-1})$ predicts the next latent state $\hat{z}_t$ in the latent space based on the previous latent state z_{t-1} and the action a_{t-1}.

In this case, robots do not need to perform massive trial and error in the real physical world, but instead generate millions of latent trajectories for imagination and strategy updates within the virtual brain (world model) [10].

In the field of physical robotics, DayDreamer migrated this architecture to a real quadruped robot (Unitree A1) and a robotic arm (UR5) [10]. Under the rigid constraints of the absence of offline pre-training and extremely sparse reward signals, Unitree A1 learned to flip its back to the ground and walk steadily while standing after only one hour of continuous online physical interaction in the real world. In contrast, conventional model-free RL requires several days of training and frequently incurs physical collisions with hardware. In the visual pick-and-place experiment of a robotic arm (UR5), DayDreamer relies on the world model to optimize the imagined trajectory in the latent space [10]. With only 8 hours of online interactive learning in the physical world, the grasping and

DOI: 10.54254/2755-2721/2026.38108

placing speed reaches 2.5 objects per minute (2.5 objects/min), which is close to the proficiency level of human remote operation (about 3 objects/min). In contrast, traditional model-free RL algorithms (such as PPO and Rainbow) are almost unable to learn effective strategies within 8 hours due to the lack of dense gradients (with an average capture count close to 0). This significantly validates the extremely high sample efficiency and feasibility of real machine deployment of the world model on sparse-reward physical systems.

6.2. Completely eliminate online exploration: offline RL

Offline RL attempts to directly learn the optimal strategy from the pre-collected non-interactive historical dataset $\mathcal{D}_{static}$, which contains a large number of failed and suboptimal attempts, completely avoiding the dangerous trial and error in sparse-reward physical environments.

Traditional Q-learning can experience a significant increase in Q values due to out-of-distribution (OOD) actions in Offline settings. Conservative Q-Learning (CQL) adds a penalty term to the objective function for the Q-value of OOD actions to solve the problem, while Implicit Q-Learning (IQL) constrains value estimation within the dataset support set through quantile regression, achieving efficient strategy extraction in extremely sparse or no-reward logs [11, 12].

In the representative offline benchmark D4RL AntMaze containing extremely sparse rewards, the traditional model-free offline SAC is affected by the explosion of OOD action Q-values, and the task success rate is almost 0% [12]. The action-constrained method BCQ has a normalized score of only 6.2% in long-distance and high-difficulty large cross tasks. CQL increased the task score to 14.9% by introducing explicit lower-bound constraints on unseen actions in the objective function. IQL completely avoided explicit evaluation of unseen actions and achieved a normalized score of 47.5% on the antmaze-arge-diverse-v0 task, a significant increase of 218.8% compared to CQL. This confirms the ability of IQL to extract optimal strategies in complex operations without the need for online hazard exploration.

6.3. Pre-training and transfer of embodied foundation large models (VLM / VLA)

With the development of Large Language Models (LLM) and Visual Language Models (VLM), embodied intelligence is ushering in a paradigm revolution. The vision-language-action (VLA) models such as RT-2, Octo, and OpenVLA directly represent multimodal data (text, video, action logs) on the Internet scale as token sequences [13, 14].

In 29 universal multitasking real physics evaluations covering multiple hardware platforms such as the WidowX robotic arm and Google Robot, OpenVLA with only 7B parameters achieved an absolute success rate jump of 16.5% compared to the representative large model RT-2-X with closed source and 55B parameters.

Internet-level common sense a priori endows robots with powerful semantic understanding and generalization capabilities. The VLA model avoids blind exploration when encountering novel objects or sparse instructions such as 'place the red fruit on the plate'. Its internal common sense prior can directly provide highly reasonable action predictions, and the real machine only needs a few example prompts or fine-tuning to successfully execute.

7. Comparison table of core method classification (taxonomy table)

To clearly delineate the performance boundaries of representative paradigms, Table 1 provides a comprehensive evaluation of mainstream approaches along adaptability and limitations.

DOI: 10.54254/2755-2721/2026.38108

Table 1. Taxonomy table

Core category	core technology	mechanism for dealing with few-shot	mechanism for dealing with sparse reward	Core limitation
Rapid Generalization	Meta-RL	Few rollouts during the testing phase to adapt	-	The meta-training stage relies on a huge amount of sample consumption.
	Few-Shot IL	Use few human demonstration trajectories.	-	The performance is limited by the level of the presenter.
	Sim-to-Real	Cost-free sampling	-	Simulation can never perfectly replicate reality.
Intrinsic Exploration	HER	Reuse failed trajectories	Transform the endpoint of failure into a virtual goal	Difficult to be applied to non-target tasks
	Intrinsic Motivation & Curiosity	-	Convert the state prediction error into a custom continuous intrinsic reward.	Sensitive to environmental noise, lacking task orientation
	World model	Majority of training completed in latent space	Millions of "virtual trial-and-error" processes in the latent space model	Model Bias
Joint Solutions	Offline RL	No need to resample in a real environment	Extract the value distribution from the historical success/failure logs	Limited by the coverage and quality of the static dataset
	Embodied foundational large model	Relying on massive pre-trained data	Pre-trained model comes with inherent priors	High computational delay and insufficient accuracy of the lower-level force control.

8. Open challenges & future directions

Although significant progress has been made in methods based on cognitive world modeling, there are still the following core challenges in pushing robot learning towards truly large-scale industrial and household implementation.

8.1. Safety-critical exploration

Under sparse rewards, intelligent agents are prone to producing intense, irrational high-frequency actions. How to strike a balance between exploring unknown intrinsic motivations and ensuring safety constraints is a prerequisite that real robot learning must address. In the future, it is necessary to combine Control Barrier Functions (CBF) in control theory with restricted RL algorithms.

8.2. Generalization of touch and force control in contact-rich tasks

At present, most strategies for solving small-sample learning, such as VLA large models, rely too much on visual perception. However, in complex assembly, screwing, and other contact-rich tasks, visual occlusion is prone to occur, and environmental rewards rely on tactile and six-dimensional force sensors. In recent years, in the benchmark testing of precision assembly and physical interaction, the pure visual strategy and the tactile fusion strategy have shown a discontinuous

DOI: 10.54254/2755-2721/2026.38108

performance gap. Lee et al. found in experiments on robotic arm insertion and long-range contact operations that, due to visual blind spots and sensor delays, the assembly success rate of pure visual behavior cloning models was less than 45%, and these issues frequently caused mechanical jamming and overload [15]. After integrating visual and multidimensional force feedback through a differentiable filter, the success rate of the task jumps to over 90%. This confirms the decisive role of tactile signals in suppressing dynamic distribution shifts. Therefore, how to introduce tactile signals into world models and offline RL frameworks is an important direction for improving the accuracy of low-level physical interactions.

8.3. Embodied RLHF

In the research on efficient preference learning for continuous robot control, Hejna and Sadigh pointed out that Traditional methods for training independent reward models require thousands of expensive annotations, and compound errors are prone to amplification [16]. The proposed inverse preference learning (IPL) directly optimizes the Q function implicitly from preference data, completely eliminating the need for explicit reward networks, achieving performance comparable to complex baselines with minimal annotation and better stability. In the field of real machine interaction correction, Cui et al. confirmed in real robotic arm operations that using large models to construct language alignment spaces can significantly improve the success rate of unseen complex tasks with only a small amount of online natural language coarse-grained correction by humans [17].

In general, large models, reinforcement learning based on human feedback (RLHF) successfully aligns language models with human intentions. In the future, how to use human high-level calibration (such as "action too light") as extremely sparse and high-quality guidance signals in a sparse-reward real environment, combined with RLHF for efficient fine-tuning of embodied large models, will become one of the most promising research hotspots.

9. Conclusion

The dual difficulty of "few-shot" and "sparse reward" in physical robot learning is essentially a mapping of exploration efficiency and information density scarcity. This essay provides a systematic summary of paradigms dedicated to either the few-shot or sparse-reward problem individually, and of unified approaches that simultaneously address both challenges, clearly expounding Rapid Generalization (Meta RL, Few-Shot IL), Intrinsic Exploration (HER, Intrinsic Motivation), and Cognitive World Modeling (World Models, Offline RL, VLA embodied model).

The future embodied intelligent system will no longer be isolated training of a single algorithm, but a comprehensive system based on massive pre-trained VLA, world models as an imagination space, and a minimal amount of real machine safety exploration aligned with humans. The maturity of this technical approach will significantly reduce the deployment cost of robots and promote intelligent robots to truly enter households and the high-end manufacturing sector.

References

[1] Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 1587-1596). PMLR.

[2] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 23-30). IEEE. https://doi.org/10.1109/IROS.2017.8202133

DOI: 10.54254/2755-2721/2026.38108

[3] Rakelly, K., Zhou, A., Finn, C., Levine, S., & Quilodrán, C. (2019). Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 5331-5340). PMLR.
[4] Duan, Y., Andrychowicz, M., Stadie, B., Ho, J., Schneider, J., Sutskever, I., Abbeel, P., & Zaremba, W. (2017). One-shot imitation learning. *Advances in Neural Information Processing Systems, 30*.
[5] Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., & Zaremba, W. (2017). Hindsight experience replay. *Advances in Neural Information Processing Systems, 30*.
[6] Fang, M., Zhou, C., Zhang, B., Yang, L., & Liu, W. (2019). Curriculum hindsight experience replay. *Advances in Neural Information Processing Systems, 32*.
[7] Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (pp. 16-17). https://doi.org/10.1109/CVPRW.2017.70
[8] Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). Exploration by random network distillation. *International Conference on Learning Representations*.
[9] Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). Mastering diverse domains through world models. *arXiv* preprint arXiv:2301.04104. https://doi.org/10.48550/arXiv.2301.04104
[10] Wu, P., Escontrela, A., Dan, D., Hafner, D., Goldberg, K., & Abbeel, P. (2023). Daydreamer: World models for physical robot learning. In *Proceedings of the Conference on Robot Learning* (pp. 2226-2240). PMLR.
[11] Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems, 33*, 1179-1191.
[12] Kostrikov, I., Nair, A., & Levine, S. (2021). Offline reinforcement learning with implicit Q-learning. *International Conference on Learning Representations*.
[13] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., ... & Zeng, A. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the Conference on Robot Learning*.
[14] Kim, M. J., Karlinsky, L., Levine, S., et al. (2024). OpenVLA: An open-source vision-language-action model for generalist robotic manipulation. *arXiv* preprint arXiv:2406.09246. https://doi.org/10.48550/arXiv.2406.09246
[15] Lee, M. A., Florence, P., Tremblay, J., Terry, N., & Bohg, J. (2022). Multimodal sensor fusion with differentiable filters for contact-rich robotic manipulation. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. https://doi.org/10.1109/ICRA46639.2022.9811988
[16] Hejna, J., & Sadigh, D. (2023). Inverse preference learning: Preference-based RL without a reward function. *Advances in Neural Information Processing Systems, 36*.
[17] Cui, Y., Karamcheti, S., Palleti, R., Shivakumar, N., Liang, P., & Sadigh, D. (2023). "No, to the right" – Online language corrections for robotic manipulation via shared autonomy. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2023)*. https://doi.org/10.1145/3568162.3576974