SCIENCE CHINA
Information Sciences

• REVIEW •

September 2026, Vol. 69, Iss. 9, 191101:1–191101:30
https://doi.org/10.1007/s11432-024-5010-x

Evolving embodied intelligence via reinforcement learning: challenges, barriers, and pathways

Tianying JI[1], Fuchun SUN[1*] & Lei LV[2]

[1]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*
[2]*Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 201210, China*

Received 25 December 2024/Revised 8 December 2025/Accepted 7 February 2026/Published online 1 September 2026

Abstract Embodied artificial intelligence—which involves agents perceiving, reasoning, and acting within physical environments—represents a new frontier towards artificial general intelligence. This study explores the evolving landscape of embodied intelligence, and presents a comprehensive survey on leveraging reinforcement learning (RL) to enhance embodied intelligence, systematically addressing the evolution from single-task performance in controlled settings to sophisticated cross-environment and cross-task capabilities. The study delves into several virtuoso capabilities, such as morphological intelligence (MI), human-like values and ethics, spontaneous reward generation, and reflective equilibrium. These capabilities are essential for developing adaptable, ethical, and intelligent embodied agents capable of navigating complex real-world scenarios. Furthermore, the study identifies and analyzes critical barriers in embodied RL, including sample efficiency, generalization and transferability, safety and robustness, real-time learning and adaptation, and interpretability and transparency. For each of these barriers, we offer an overview of current approaches and propose potential avenues for future research.

Keywords reinforcement learning, embodied artificial intelligence, embodied reinforcement learning, robotics, foundation models, safety and robustness, comprehensive survey

Citation Ji T Y, Sun F C, Lv L. Evolving embodied intelligence via reinforcement learning: challenges, barriers, and pathways. Sci China Inf Sci, 2026, 69(9): 191101, https://doi.org/10.1007/s11432-024-5010-x

1 Introduction

Embodied artificial intelligence (AI) represents the next frontier in the development of autonomous systems. The concept of embodied AI finds its roots in the idea that intelligence is not merely a product of abstract computation but is deeply intertwined with an agent's interaction with the physical world. This perspective draws inspiration from the philosophy of embodied cognition [1,2], which posits that cognition is shaped by the body's interactions with its environment. The term embodied AI refers to systems that not only perceive and reason about the world but also act within it, using physical sensors and actuators to perform tasks, learn from experiences, and adapt to changing conditions. As such, embodied AI is considered an essential step toward achieving more autonomous, generalizable, and adaptive AI systems [3].

At its core, embodied AI refers to intelligent systems that integrate sensory feedback, perception, and action within a physical body. These systems can learn from real-world interactions, adapt to dynamic environments, and perform tasks autonomously, ranging from simple actions like object manipulation to more complex decision-making in uncertain, real-world contexts [4]. Unlike traditional AI, which works in isolated environments (e.g., analyzing data or simulating problems), embodied AI has a physical or simulated presence that allows it to directly interact with its surroundings. This interaction is what makes it particularly suited for applications like robotics, autonomous driving, healthcare assistance [5], and even virtual environments like gaming or metaverse avatars.

The significance of embodied AI lies in its ability to bridge the gap between abstract reasoning and physical, real-world actions. By learning through interaction, embodied AI systems are capable of adapting to new situations, handling uncertainty, and performing tasks that require both high-level reasoning and sensory-motor integration. In fields such as autonomous robotics, embodied AI is crucial for enabling systems to operate in dynamic environments, interact with humans naturally, and carry out tasks without constant human supervision [6]. This adaptability and autonomy are essential for applications such as search and rescue, elderly care, precision surgery, and self-driving vehicles, where real-time decision-making and physical actions are crucial [7].

* Corresponding author (email: fcsun@tsinghua.edu.cn)

Table 1 Comprehensive performance on LIBERO benchmark. The table integrates data from VLA+RL, skill-based RL, and traditional baselines, highlighting the dominance of VLA fine-tuning. Success rates (%) are reported for spatial, object, goal, and long-horizon suites. Data sourced from [22, 27, 29].

Method	Paradigm	Spatial	Object	Goal	Long	Avg
VLA + RL fine-tuning						
SimpleVLA-RL [22]	VLA + Online RL (GRPO)	99.4	99.1	99.2	98.9	99.1
SRPO [23]	VLA + Self-Ref RL	98.8	100.0	99.4	98.6	∼99.2
RLinf-VLA [24]	VLA + RL	99.4	99.8	98.8	94.0	98.0
OpenVLA-OFT [25]	VLA + Optimized SFT	98.4	97.9	94.5	97.6	97.1
TGRPO [26]	VLA + RL	90.4	92.2	81.0	59.2	80.7
VLA/foundation baselines						
π_0	VLA (flow matching)	96.8	98.8	95.8	85.2	94.2
OpenVLA (Base)	VLA (SFT)	76.5	84.7	73.5	52.1	71.7
Diffusion policy	CNN/Transformer + BC	80.0	85.0	75.0	58.5	74.6
Skill-based&traditional						
EXTRACT [27]	Skill-based offline RL	85–88	88-90	88–90	50–52	–
SPiRL [28]	Continuous latent skills	65–68	60–62	70–72	30–32	–
BC (ResNet)	Imitation learning	<10	<15	<15	<5	–
SAC (Pixel)	Model-free RL	∼0	∼0	∼0	∼0	∼0

Reinforcement learning (RL) has achieved remarkable progress in solving complex decision-making problems, ranging from strategic games [8–10], often played in virtual, disembodied environments, to real-world robotic applications through sim-to-real deployment [11–13], and even in fields as specialized as magnetic control of tokamak plasmas [14], champion-level drone racing [15], and autonomous stratospheric balloon control [16]. These advancements underscore RL's potential in both synthetic and real-world contexts. Among its many successes, RL has been a key driver of innovations like ChatGPT [17], further demonstrating its versatility across diverse fields. RL is highly suitable for embodied AI as it facilitates sequential decision-making and supports long-term optimization, even with delayed rewards. It also enables agents to continuously adapt to new challenges, making it ideal for autonomous systems operating in complex and unpredictable environments.

Existing studies and our contributions. Existing surveys on embodied AI have predominantly concentrated on foundational aspects, such as the integration of perception, motor control, and basic interaction capabilities necessary for AI systems to operate within physical bodies [4, 18, 19]. These surveys have provided comprehensive overviews of underlying tasks—sensor integration, embodied interaction mechanisms, and control strategies, laying the groundwork for understanding how AI can function effectively in embodied contexts. However, there remains a notable gap in the literature concerning the higher-level capabilities required for embodied AI to approach the versatility and adaptability envisioned for artificial general intelligence (AGI).

This study seeks to address that gap by offering a comprehensive survey that extends beyond the basic functionalities of embodied AI to examine the **emerging levels of embodied intelligence** and the critical capabilities required for AGI. We introduce a novel framework that classifies embodied intelligence into multiple levels, ranging from single-task performance in controlled environments to cross-task and cross-environment adaptability. This hierarchical perspective provides a roadmap for the evolution of more sophisticated and versatile embodied AI systems.

Furthermore, while previous research has explored various RL techniques for embodied AI [20, 21], there is a lack of **a systematic survey that examines how RL can be leveraged to enhance embodied intelligence** across these evolving capabilities. We provide a detailed analysis on numerous RL techniques, highlighting their potential to address current challenges in embodied AI and to guide future research directions.

To provide a rigorous quantitative foundation, we analyze performance boundaries across three premier benchmarks: LIBERO, CALVIN, and RLBench.

Table 1 demonstrates the dominance of VLA fine-tuning in lifelong learning [22–29]. On the LIBERO benchmark, SimpleVLA-RL achieves near-perfect success rates, correcting the trajectory drift that causes standard pixel-based RL and skill-based priors to fail in long-horizon tasks.

For offline-to-online adaptation, Table 2 contrasts the performance collapse of uncalibrated algorithms (e.g., CQL) with the stability of calibrated methods. Notably, PA-RL improves OpenVLA's success rate by 75% on CALVIN, quantifying the efficacy of policy-agnostic fine-tuning in preventing catastrophic forgetting [30–35].

Regarding perception modality, Table 3 confirms the necessity of structured 3D priors. Standard RL fails completely on RGB-based RLBench tasks but exceeds 90% success on ManiSkill2 when provided with point cloud

Table 2 Performance analysis on CALVIN benchmark (ABCD → D setting). Hierarchical VLAs and generative models dominate long-horizon tasks (Avg. Len.), while traditional offline RL struggles with long-term dependencies. Data sourced from [34, 35]. The best results are in bold.

Method	Learning paradigm	Avg. Len.	1-instr SR (%)	5-instr SR (%)
Hierarchical/generative				
UniVLA [30]	Hierarchical VLA + Planning	**3.90–4.10**	**>95**	**∼75**
FLOWER [31]	Flow-matching policy	3.80	∼93	∼68
MDT [32]	Multimodal diffusion transformer	3.50	∼90	∼60
HULC [33]	Transformer implicit planning	2.80–3.00	∼85	∼25
Reinforcement learning				
Cal-QL [34]	Calibrated offline RL	1.50–2.00	50–60	<5
IQL	Implicit Q-learning	∼1.20	∼40	∼0
Baseline				
BC baseline	ResNet + LSTM	0.64	48.9	0.08

Table 3 Impact of perception modality on manipulation success. Comparison across RLBench (RGB/voxel) and ManiSkill2 (point cloud) highlights the necessity of structured priors for RL. Data sourced from [36, 37]. The best results are in bold.

Benchmark	Method	Input modality	Success rate (%)	Root cause analysis
RLBench (18 Tasks)	RVT [38]	Multi-view	**62.9**	Multi-view attention captures 3D geometry.
	PerAct [36]	3D Voxels	49.4	34× gain vs 2D baselines via spatial bias.
	GNFactor [39]	3D Voxel + NeRF	46	Neural fields aid in semantic understanding.
	PPO/SAC	2D RGB	∼0	Fails to infer geometry from sparse rewards.
RLBench (Stacking)	PPO-LSTM	State + Memory	**100**	Recurrent memory maintains belief state under occlusion.
	PPO-MLP	State only	25	Fails due to POMDP nature of occluded manipulation.
ManiSkill2 (PickCube)	PPO (DAPG)	Point cloud	**94**	Explicit geometry enables efficient state extraction.
	PegInsertion (PPO)	Point cloud	1	Pure RL struggles with high-precision contact dynamics.

inputs, underscoring that explicit geometry is a prerequisite for pure RL in contact-rich manipulation [36–39].

Beyond reviewing the current state of the art, this work identifies promising future avenues and breakthroughs that are critical for the next generation of embodied intelligence.

We explore **how RL can empower virtuoso capabilities**, such as morphological intelligence, human-like ethics and values, reflective equilibrium, and spontaneous reward generation—key factors in the development of adaptable, ethical, and intelligent embodied agents. Through this study, we aim to provide a valuable resource for researchers and practitioners who seek to advance the boundaries of embodied AI through the strategic application of reinforcement learning.

2 Leveraging reinforcement learning for enhancing embodied intelligence

This section explores how RL can be applied to enhance embodied intelligence across four distinct levels of capability. These levels reflect the increasing complexity of tasks and environments that embodied systems must handle. The initial level addresses single-task capability in a controlled environment, where RL optimizes performance in specific, well-defined tasks. As complexity increases, enabling single-task capability across multiple environments allows the system to adapt to different settings while performing the same task. Meanwhile, a multi-task capability in a single environment requires the system to handle multiple tasks simultaneously within the same environment, demanding advanced decision-making and coordination. At the highest level, cross-environment and cross-task capability combine the previous challenges, enabling the agent to transfer knowledge across diverse tasks and environments. We discuss how RL can be applied to tackle the unique challenges at each level, driving advancements in robotics and intelligent systems. Through these discussions, we emphasize the value of RL in enhancing embodied systems' adaptability, autonomy, and effectiveness in real-world, dynamic environments. We provide an overview in Figure 1.

2.1 Single-task capability in a controlled environment

Reinforcement learning has made substantial strides in addressing a range of complex decision-making tasks, from strategic games like chess [9], Go [8] and computer games [10, 40]—typically played in virtual, disembodied environments—to real-world tasks such as locomotion, manipulation, and navigation [11–13, 41, 42]. The success of RL in these domains largely depends on access to unbiased, interactive environments, which may be high-fidelity simulators or the real-world systems themselves, and on the ability to conduct millions of unrestricted trial-and-error interactions [43, 44].

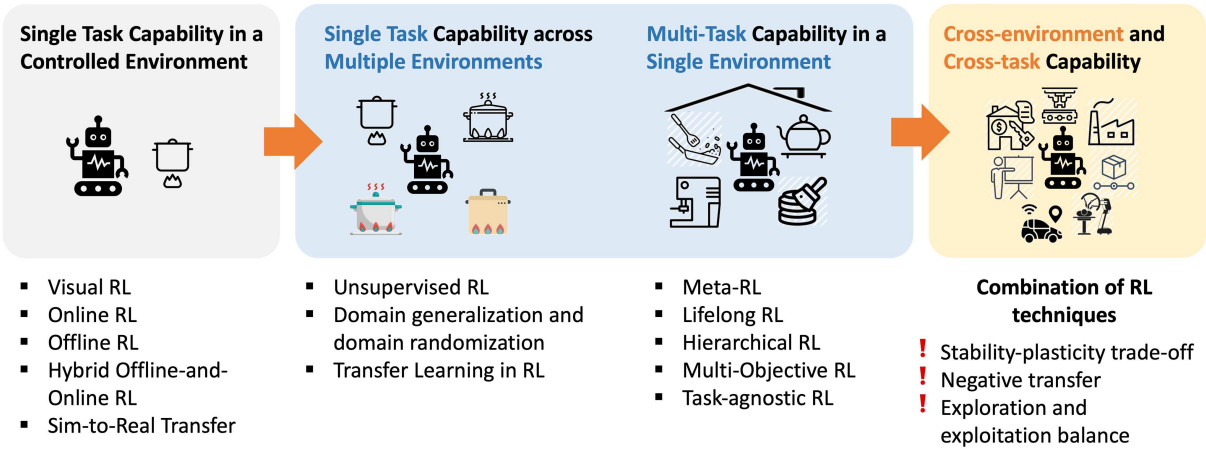

Figure 1 Evolving landscape of embodied intelligence and promising RL techniques. From left to right, the figure illustrates the progression from single-task capability in a controlled environment, to single-task capability across multiple environments, multi-task capability within a single environment, and finally, cross-environment and cross-task capability. Each level represents an increasing complexity of tasks and environments, demonstrating how RL can enhance the adaptability, versatility, and intelligence of embodied systems.

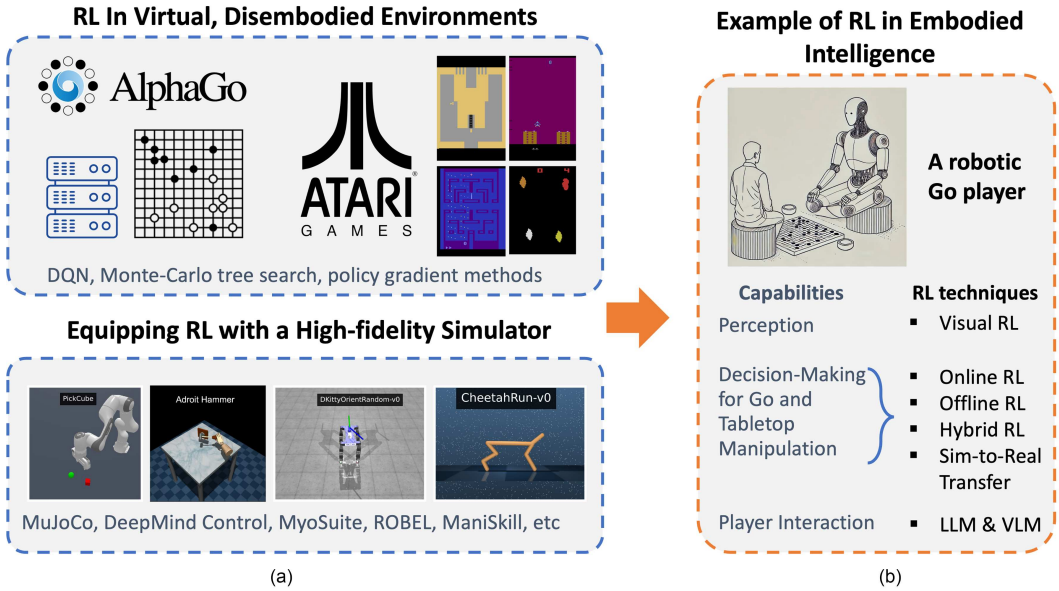

Figure 2 Illustration of a Go player evolving from disembodied to embodied. (a) RL has achieved success in disembodied environments, such as virtual games (e.g., chess, Go), relying on unbiased, interactive simulations or real-world systems. (b) Embodied RL is exemplified by a robotic Go player, which integrates perception (for sensing the game state), RL-driven decision-making for gameplay, manipulation of pieces, and player interaction via language models, allowing seamless engagement with the physical environment.

Conversely, in the realm of embodied intelligence, RL must extend beyond abstract decision-making to control a physical robot capable of interacting directly with its environment. For instance, imagine a robotic chess player who not only strategizes but also physically engages with the board and interacts with an opponent. Take, for example, a robotic Go player—not just an AI that plays the game virtually but a robot that physically interacts with the board and engages with its opponent. This setup demands several core capabilities. (1) Perception: equipped with sensors, the robot must observe the chessboard, accurately identifying piece positions to understand the game state. (2) Decision-making for Go: using RL-based strategies, the robot evaluates potential moves and adapts its gameplay based on the opponent's actions. (3) Decision-making for tabletop manipulation: the robot must pick up and place chess pieces with precision, relying on RL or control algorithms to handle objects smoothly. (4) Player interaction: through language model capabilities, such as a large language model (LLM) or a vision language model (VLM) powered by RL (e.g., ChatGPT [17], GPT-4V [45]), the robot interacts conversationally with the player, maintaining relevant and engaging exchanges. We illustrate the robotic Go player and the potential engaged RL techniques in Figure 2.

2.1.1 *RL methods for pure decision-making tasks in simulations and offline datasets*

For solving pure decision-making within simulations or offline datasets, RL methods are generally adequate, allowing agents to effectively learn optimal strategies and control policies through repeated interaction with a simulated environment or pre-collected data.

Online RL with a high-fidelity simulator. When with access to a high-fidelity simulator, **online RL** paradigm is especially advantageous as it allows agents to conduct extensive trial-and-error cycles, iteratively refining their decision-making processes unburdened by the physical or safety limitations of real-world settings [37, 46–48]. Online RL has achieved notable breakthroughs in virtual domains like strategic games (e.g., Go) and video game environments, enabling agents to attain superhuman skills through prolonged self-play and policy optimization over many simulated episodes [8–10, 40]. Further, the high-fidelity simulator enables agents to learn through rich, simulated interactions that mimic real-world physics, allowing them to master complex behaviors, such as controlling robotic arms, navigating dynamic environments, and manipulating objects [41, 49–52].

Offline RL with pre-collected data. In scenarios where real-time interaction is impractical, unsafe, or too costly, and where constructing high-fidelity simulators may be prohibitively expensive or even unfeasible, **offline RL** is an appealing alternative. Offline RL learns effective policies exclusively from previously collected data, without requiring any further interaction with the environment.

While most off-policy online RL algorithms can technically be applied in offline settings by filling the replay buffer with static data, challenges arise in improving the policy beyond the baseline set by the behavior policy that generated the dataset. Specifically, policy optimization often involves querying the Q-function for values of actions generated by the new policy, which are frequently out-of-distribution and thus absent in the dataset. This leads to extrapolation errors, where the Q-function may inaccurately estimate the value of unseen actions, destabilizing learning [53, 54].

Offline RL addresses the distribution shift issue primarily by adding pessimism to the learning objective—for instance, by constraining the policy to stay close to the behavior policy [54–56] or through conservative value regularization [57–59]. Recently, emerging approaches have also employed in-sample learning techniques [60, 61], which avoid querying the value function for any unseen actions. Consequently, offline RL methods are generally well-suited for deriving decision-making capabilities with pre-collected datasets.

Hybrid offline-and-online RL (hybrid RL). Hybrid RL combines the benefits of offline and online RL, addressing the key limitations of each [43, 62–64]. Offline RL trains agents using pre-collected datasets, which is ideal when interaction with the environment is costly or unsafe. Despite this, it often faces a distributional shift, where the learned policy encounters states not represented in the dataset. Online RL, on the other hand, enables real-time exploration and adaptability but is sample-inefficient, typically requiring extensive environment interactions.

The hybrid RL addresses these challenges by using offline data to establish an initial policy, reducing the need for broad exploration, and then refining this policy through selective online interactions. By integrating online updates, the agent can address distributional shifts dynamically, correcting any policy errors that arise from unfamiliar states. This iterative adjustment enables the agent to build on the offline data foundation while fine-tuning its policy based on real-world feedback, resulting in a more robust, adaptable, and sample-efficient learning process.

2.1.2 *Vision-based RL for integrating perception into decision-making*

When integrating perception into decision-making, **visual RL** becomes a critical approach for managing complex continuous control tasks that depend on high-dimensional pixel data. Visual RL leverages advanced online RL algorithms as its backbone and equips agents to interpret raw visual inputs, enabling perception, object recognition, and navigation in dynamic environments [65, 66]. For instance, in high-dimensional control tasks like dexterous hand manipulation (Adroit suite [67]) or locomotion (DeepMind Control Suite [68]), visual RL helps agents acquire sophisticated motor skills by directly learning representations and policies from visual data.

With advanced techniques like data augmentation [69, 70], self-supervised representation learning [71–73], and temporal difference regularization [74], visual RL addresses challenges such as partial observability, complex kinematics, and high degrees of freedom. Furthermore, **model-based visual RL** enhances this learning process by constructing predictive models of the environment, known as "world model" [75–78]. These models allow agents to simulate potential outcomes from pixel observations before acting, which reduces sample complexity by enabling learning through imagined scenarios rather than direct interactions with the environment.

2.1.3 *Sim-to-real transfer in RL for real-world deployment*

Sim-to-real transfer in RL aims to bridge the gap between training in simulations and deploying policies in real-world environments [13, 79, 80]. The primary challenge is the reality gap: simulated environments often simplify or fail to capture real-world dynamics, leading to discrepancies in performance when transferring policies. Popular techniques to mitigate these gaps include the following. (1) Domain randomization: randomize simulator parameters (e.g., visual features, dynamics) to train policies that generalize better to unseen real-world variations [81, 82]. For example, adding variability in textures, lighting, or friction coefficients makes the learned policies robust to diverse real-world conditions. (2) Domain adaptation: this method aligns the feature space of the simulated and real domains, often using adversarial learning. By learning invariant features, agents can generalize learned skills despite domain-specific discrepancies [83, 84]. (3) Meta-learning and knowledge distillation: meta-learning [85] prepares the agent to adapt quickly to real-world scenarios by learning over a distribution of simulated tasks. Knowledge distillation involves transferring a robust, pre-trained "teacher" policy to a more adaptable "student" policy optimized for real-world conditions [86]. In short, sim-to-real transfer enables effective RL policy deployment in real-world robotics and complex control applications by addressing sensory and dynamic mismatches through these methods.

2.2 Single-task capability across multiple environments

Enabling agents to perform the same task effectively across diverse and dynamic settings is a hallmark of advanced embodied AI. In contrast to single-environment scenarios, where tasks are learned under fixed and controlled conditions, multi-environment setups introduce significant variability in dynamics, sensory inputs, and external factors [87, 88]. These variations necessitate seamless adaptability while ensuring consistent performance. Such adaptability is essential for real-world deployment, where environments are inherently unpredictable and diverse. For example, a single manipulation task can be executed seamlessly across various robotic arms, quadruped robots adapt their locomotion to both urban streets and natural terrains, and autonomous vehicles handle road conditions ranging from icy highways to sandy trails.

Key RL techniques address these challenges through pre-training policies that generalize across environments, adaptive fine-tuning, and environment-specific modeling. Methods often include intrinsic motivation, unsupervised RL, domain generalization, and transfer learning. Below, we detail the primary avenues for achieving single-task capability in multiple environments.

2.2.1 *Unsupervised RL*

Unsupervised RL focuses on enabling agents to learn robust and generalizable behaviors across diverse environments without relying on explicit reward signals. This paradigm leverages intrinsic objectives to guide exploration and policy development, preparing agents to perform single tasks effectively across varied settings. Recent advancements in unsupervised RL can be grouped into several key approaches, each addressing specific challenges in multi-environment adaptability and task generalization.

Intrinsic motivation for exploration. Curiosity and diversity-based intrinsic rewards [40] are central to many unsupervised RL methods. Agents are encouraged to explore less-visited or unpredictable states, maximizing their knowledge of the environment. For example, curiosity-driven frameworks reward agents for reducing predictive errors in state transitions, enabling effective exploration of unknown areas. Similarly, diversity-driven approaches, such as skill discovery, optimize for policies that maximize behavioral diversity, allowing agents to acquire a wide range of reusable skills applicable across environments. These methods ensure broad state coverage and prepare agents for downstream task fine-tuning with minimal additional training.

Multi-environment policy optimization. A widely used approach involves updating the agent's policy by leveraging shared experiences from multiple environments. In this framework, the agent encounters a range of tasks (environments) and learns to optimize a policy that performs effectively across all of them. Policy gradient methods are commonly applied, where the gradient of the total reward, accumulated from different environments, is used to adjust the policy parameters [89]. Additionally, optimizing for expected performance across all environments during training improves adaptability and ensures stability, which is crucial in situations where agents must function under diverse conditions [90], such as in robotics tasks with varied workspace setups or autonomous navigation in changing weather conditions.

State history and memory augmentation. In complex environments where single-step state observations are insufficient, non-Markovian policies incorporate memory to handle temporal dependencies [91]. By considering

past state histories, agents can develop a deeper understanding of long-term dependencies, enabling more effective decision-making across environments with partial observability or dynamic changes.

Meta-learning for adaptation. Meta-learning techniques in unsupervised RL aim to equip agents with the ability to adapt quickly to new environments by leveraging shared knowledge across tasks. These methods learn meta-policies that can be generalized across multiple environments [92], requiring minimal fine-tuning for new conditions. Such approaches are particularly beneficial for tasks like robotic manipulation, where physical parameters (e.g., object weight or surface friction) can vary significantly between environments.

2.2.2 *Domain generalization and domain randomization*

Domain randomization (DR). DR trains agents in simulations with varied parameters, such as textures, lighting, dynamics, and object properties, fostering robustness against unseen changes during deployment. In robotics, DR can randomize factors like object mass, friction, and visual conditions, enabling agents to better handle real-world variability.

A major benefit of DR is its role in sim-to-real transfer, where agents trained in simulation operate effectively in real-world settings. Randomizing parameters helps align the training environment with real-world conditions, narrowing the reality gap. For instance, Tobin et al. showed that DR allowed a robotic arm trained in simulations to perform precise object manipulation in real life [93], while Hwangbo et al. demonstrated improved legged robot control through dynamics randomization [12]. Despite its advantages, excessive randomization can create unrealistic scenarios, hindering learning. Additionally, randomization must sufficiently represent the target domain to ensure reliable transfer.

Domain generalization (DG). DG enables agents to perform in unseen environments without prior exposure during training by learning invariant features or policies effective across domains with shared task characteristics. Unlike DR, which diversifies training conditions, DG focuses on isolating task-relevant information while ignoring domain-specific variations.

Techniques like adversarial training align feature spaces to ensure policies are not domain-dependent. For instance, domain-adversarial neural networks (DANN) use adversarial learning to align feature distributions across domains, facilitating skill transfer [94]. Similarly, DARLA disentangles representations to enable zero-shot policy transfer to unseen environments [95]. DG is particularly valuable when simulating diverse training environments, which is costly or impractical, but its success hinges on effectively identifying domain-invariant patterns in highly variable conditions [96].

2.2.3 *Transfer learning in RL*

Transfer learning in RL focuses on enabling agents to leverage knowledge acquired in one environment or task to improve learning in a different, often related, environment or task. Transfer learning aims to generalize learned behaviors and efficiently adapt to new scenarios by transferring skills, policies, or representations.

Policy transfer. Policy transfer in reinforcement learning enables agents to leverage pre-trained policies from a source task to accelerate learning in a related target task. The simplest form, direct policy reuse, applies the pre-trained policy to the target task without modifications and is effective when tasks share similar dynamics or objectives. When task discrepancies arise, policy fine-tuning adjusts the pre-trained policy to better suit the target environment, retaining useful knowledge while adapting to new requirements [80, 97, 98].

For tasks involving hierarchical decision-making, hierarchical policy transfer reuses sub-policies for low-level behaviors, such as locomotion, while adapting high-level strategies like navigation to task-specific constraints [99]. When multiple source tasks are available, multi-source policy transfer integrates knowledge from diverse sources, often using meta-learning frameworks like model-agnostic meta-learning (MAML) to identify and adapt relevant policies efficiently [100].

An alternative approach, soft policy transfer, combines the pre-trained policy with exploratory strategies in the target task, balancing exploitation of prior knowledge with the discovery of new behaviors. This probabilistic integration mitigates overfitting to the source task and encourages robust exploration [81, 101]. Despite challenges such as negative transfer, advancements in hierarchical methods, meta-learning, and feature alignment continue to enhance the effectiveness of policy transfer across diverse RL tasks [102].

Representation transfer. Representation transfer reuses learned features or latent representations from a source task to accelerate learning in related target tasks, improving sample efficiency and reducing training complexity [103, 104]. Pre-trained neural networks often reuse early layers for general features, and fine-tuning later layers for adaptation [105]. Feature space alignment, such as adversarial learning, minimizes domain discrepancies, ensuring compatibility across tasks [106].

In hierarchical RL, disentangled representations enable agents to adapt shared dynamics across tasks while minimizing retraining [107]. Representation transfer also supports multi-task learning by leveraging shared features to generalize across environments [108]. Meta-learning approaches like gradient-based optimization further refine representations for fast adaptation [109]. By mitigating challenges like negative transfer with selective fine-tuning, representation transfer enhances adaptability in robotics and navigation systems.

Value function transfer. Value function transfer reuses value functions—estimations of expected cumulative rewards—from a source task to accelerate learning in a related target task, improving sample efficiency and reducing training time. This technique aligns state and action spaces across tasks to facilitate effective transfer [110]. Multi-agent settings benefit from N-step returns to transfer value functions between single-agent and multi-agent tasks, enhancing coordination [111]. Universal value function approximators (UVFAs) generalize value functions across multiple goals or tasks, enabling agents to adapt efficiently to diverse environments [112]. Challenges such as negative transfer, where mismatched value functions hinder learning, are addressed through reward shaping and regularization. Value function transfer remains a vital tool for leveraging prior knowledge to improve RL performance across related tasks.

Reward function transfer. Reward function transfer in RL reuses reward structures from a source task to accelerate learning in a related target task. Maintaining consistent objectives across tasks with similar goals but differing environments enables agents to adapt efficiently without redefining rewards. For example, a reward function designed for autonomous driving in one city can be applied in another with different road layouts, allowing the agent to focus on learning environmental dynamics [113]. This method is particularly effective for tasks with shared high-level goals but varying dynamics or state spaces. Techniques like generalized reward shaping enhance alignment between source and target tasks, ensuring meaningful and efficient transfer [114]. Reward function transfer also supports multi-task RL by unifying reward structures across tasks while maintaining flexibility for environment-specific nuances [115]. Even so, careful design is crucial to avoid negative transfer, where mismatched rewards can misguide learning.

2.3 Multi-task capability in a single environment

In embodied AI, the ability to handle multiple tasks within a single environment is a critical milestone, complementing the challenge of mastering individual tasks across diverse settings.

The practical significance of this capability becomes evident in scenarios such as warehouse robots navigating through aisles, sorting objects, and stocking shelves—all within the same workspace. Similarly, smart home assistants must seamlessly manage tasks like climate control, cleaning, and security, while gaming agents adapt to diverse mechanics such as exploration and combat. In industrial settings, factory robots are required to transition effortlessly between assembly, inspection, and error correction. These examples highlight the necessity of multi-task learning, where agents must address diverse objectives either concurrently or sequentially without retraining.

Multi-task capability hinges on balancing flexibility to handle diverse tasks, generalization to transfer skills, and resource efficiency to optimize performance. Addressing these challenges is vital for robust embodied intelligence in real-world environments. Reinforcement learning approaches like meta-reinforcement learning (meta-RL) for rapid adaptation, Lifelong RL for skill retention, hierarchical RL (HRL) for task decomposition, multi-objective RL (MORL) for balancing competing priorities, and task-agnostic RL for generalization collectively advance multi-task capabilities.

2.3.1 *Meta-RL*

Meta-RL is a specialized subfield of RL that focuses on enabling agents to adapt rapidly to new tasks by leveraging prior experience. Unlike traditional RL, where agents learn a single policy for a specific task, meta-RL aims to train agents across a distribution of tasks so they can quickly generalize and adapt to unseen tasks with minimal data. This adaptability is essential in dynamic environments where tasks may vary significantly, requiring agents to respond effectively without extensive retraining.

At the core of meta-RL is the ability to exploit shared patterns across tasks to prepare agents for rapid adaptation. For instance, MAML has been a foundational approach, training agents to find an optimal initialization of parameters that allows them to adapt to new tasks with just a few gradient updates [100]. Similarly, RL^2 uses recurrent neural networks (RNNs) to encode task information from past experiences, enabling agents to adapt dynamically without explicit task labels [3]. Probabilistic embeddings provide another perspective by learning task-specific context variables that inform policy adaptations, enhancing efficiency and scalability [116].

Despite its promise, meta-RL faces several challenges. One of the key issues is **task distribution shift**, where the training and test tasks differ significantly, leading to performance degradation when encountering out-of-distribution

tasks. Sample efficiency is another major hurdle, as meta-RL requires sufficient training data from diverse tasks to generalize effectively [117]. Moreover, the exploration-exploitation trade-off becomes more complex in meta-RL, as agents must explore new tasks sufficiently to understand their dynamics while exploiting prior knowledge to perform optimally. Computational complexity is also a concern, particularly for methods involving large neural networks, which require significant resources to train on diverse task distributions.

Recent advancements in meta-RL have sought to address several challenges in task inference and adaptation. Temporal convolutions have been introduced to efficiently capture task-specific information from sequences of interactions, thus improving task inference and adaptation [118]. In another development, methods for structured exploration strategies have been proposed to optimize agents' exploration behaviors across tasks, thereby reducing the need for task-specific exploration [119]. Furthermore, benchmarks like Meta-World have been developed to evaluate meta-RL methods in robotic manipulation, providing a critical tool for assessing progress in the field [92]. Together, these advancements underscore the rapid progress in meta-RL and its potential to revolutionize applications in robotics, autonomous systems, and beyond.

2.3.2 *Lifelong reinforcement learning (lifelong RL)*

Lifelong RL enables agents to continuously acquire, retain, and adapt knowledge across sequential tasks in dynamic, evolving environments. By emphasizing cumulative learning, it allows agents to leverage prior knowledge to accelerate learning and improve performance on new tasks while preserving previously acquired skills. This ability is essential for real-world applications where agents are required to adapt to non-stationary task distributions over time [120,121]. Key approaches in lifelong RL are commonly grouped into regularization-based methods, replay-based methods, and architectural strategies.

Regularization-based methods tackle catastrophic forgetting by restricting updates to parameters crucial for earlier tasks. For example, elastic weight consolidation penalizes large changes to important parameters, preserving prior knowledge [120], while synaptic intelligence dynamically tracks and protects critical weights throughout learning [122].

Replay-based methods prevent forgetting by periodically revisiting previous tasks. This is achieved by storing past experiences in a buffer or generating synthetic data from learned models, allowing agents to reinforce earlier knowledge while training on new tasks [121, 123].

Architectural strategies, such as modular and compositional designs, adapt dynamically by allocating new resources for new tasks. Techniques like Progressive Neural Networks [97] and PathNet [124, 125] preserve older knowledge in separate pathways while enabling transfer through lateral connections, balancing the stability of past knowledge with the plasticity needed for new learning.

Lifelong RL faces several key challenges. **Catastrophic forgetting** is a primary concern, where acquiring new tasks can overwrite prior knowledge, leading to degraded performance on earlier tasks. While methods such as regularization and experience replay partially address this issue, they often incur significant memory or computational overhead [120,121]. Another challenge is **negative transfer**, where knowledge from unrelated tasks interferes with learning new ones. This issue typically arises from misjudged task similarity and requires improved measures of task relationships and adaptive mechanisms for knowledge sharing [126]. Additionally, **scalability** becomes increasingly problematic as task sequences grow in complexity and diversity, demanding efficient management of memory and computational resources to ensure sustained learning [127].

2.3.3 *HRL*

HRL introduces a framework to tackle complex tasks by decomposing them into hierarchical structures, enabling agents to learn and execute subtasks in a structured manner. This decomposition aligns well with the demands of multi-task reinforcement learning, where agents must perform a variety of tasks efficiently within shared environments. By organizing learning into layers of abstraction, HRL facilitates skill reuse, improves learning efficiency, and enhances scalability in solving complex, long-horizon problems [128–131].

Central to HRL is the principle of task decomposition, where high-level policies govern decision-making over abstract goals, and low-level policies handle specific actions to achieve these goals. This hierarchical structure allows agents to solve tasks by combining reusable sub-policies, significantly reducing the complexity of policy search. Approaches such as the options framework have been instrumental in formalizing this decomposition, providing a structured way to define temporal abstractions [132]. The feudal reinforcement learning paradigm further refines this idea by introducing hierarchical managers that issue subgoals to subordinate agents, facilitating task specialization and inter-level communication [133]. Advances in subtask learning focus on enabling policies to autonomously discover, refine, and reuse subtasks. Methods such as Option-Critic [134, 135] and H-DQN [136]

highlight how agents can learn subtasks and hierarchical policies simultaneously. Similarly, approaches in **skill discovery** use unsupervised techniques or intrinsic rewards to identify reusable behaviors, enabling efficient skill transfer across tasks [137, 138].

HRL faces significant challenges. Designing effective hierarchies is non-trivial, as rigid or poorly structured hierarchies can hinder learning. Approaches such as dynamic hierarchies and adaptive subtask granularity address this by allowing policies to refine their structure based on environmental feedback [139]. Scalability is another issue, as hierarchical policies introduce computational overhead, particularly in maintaining consistency between layers. Additionally, exploration-exploitation balancing becomes more complex in HRL, especially in sparse reward environments [133]. Further, transferring learned hierarchies across tasks remains a challenge, demanding robust methods for handling task variability [136].

2.3.4 *MORL*

MORL addresses decision-making challenges involving multiple, often conflicting objectives. Instead of optimizing a single goal, it enables agents to balance trade-offs, such as speed versus safety or cost versus efficiency. This capability is particularly critical in multi-task environments, where agents must consider trade-offs to achieve optimal outcomes across diverse objectives [140, 141].

MORL enables agents to learn policies that account for the relative importance of multiple objectives by modeling the environment to provide a vector of rewards for each objective. To address the challenge of balancing competing objectives, key methods include scalarization techniques, Pareto-based approaches, and policy search strategies. Scalarization combines multiple objectives into a single scalar value, typically using weighted sums that require careful tuning to reflect trade-offs, with advanced methods like the Tchebycheff approach systematically balancing objectives to efficiently learn Pareto-optimal policies [142, 143]. Pareto-based methods aim to identify the Pareto front, where no objective can be improved without compromising another. Frameworks like non-stationary policy gradients enhance this process by capturing the full Pareto front, enabling flexible multi-objective solutions [141,144]. Policy search strategies directly optimize policies for multi-objective tasks, often using gradient-based techniques to explore complex landscapes. Recent innovations, such as the latent-conditioned policy gradient, improve scalability by approximating the entire Pareto set within a single neural network [145, 146]. Together, these methods provide a comprehensive toolkit for advancing MORL in complex, multi-objective environments.

MORL faces several challenges. Balancing computational efficiency with the complexity of optimizing multiple objectives remains a significant issue, particularly in high-dimensional spaces. Accurately estimating Pareto fronts or defining effective scalarization functions can also be difficult. Furthermore, conflicting objectives can destabilize learning, as improvements in one dimension may negatively impact others. Overcoming these challenges requires the development of scalable algorithms, robust scalarization techniques, and conflict resolution mechanisms [140, 141].

2.3.5 *Task-agnostic reinforcement learning (TARL)*

TARL is a paradigm in reinforcement learning where agents explore environments without predefined reward functions, focusing on acquiring generalizable skills applicable across multiple tasks. Through intrinsic motivation, such as curiosity-driven exploration, agents discover novel states and actions, building a deep understanding of environment dynamics [147]. TARL emphasizes unsupervised task space learning, where agents autonomously identify potential goals and refine behaviors that transfer across scenarios [148]. This allows agents to practice skills beneficial for future tasks without relying on external guidance.

Despite its advantages, TARL faces notable challenges. Designing intrinsic reward signals that effectively drive exploration without encouraging redundant or suboptimal behaviors is a significant hurdle. Ensuring that acquired skills are diverse yet transferable across tasks also requires careful optimization of exploration strategies. Moreover, integrating knowledge gained from task-agnostic exploration into task-specific learning phases remains an open problem. Addressing these challenges is essential for fully realizing TARL's potential [149].

2.4 Cross-environment and cross-task capability

As embodied AI systems continue to advance, their ability to operate across diverse tasks and environments becomes increasingly crucial. Previous sections explored methods for single-task learning in controlled settings, multi-task learning within a single environment, and adaptation to multiple environments. Now, to address the complexities of real-world applications, agents must integrate these capabilities to handle a wide range of tasks across different environments, while maintaining flexibility, adaptability, and generalization.

For instance, consider a household robot that must adapt its behavior to different rooms, such as the kitchen, living room, and garage. Each room presents unique challenges—different spatial layouts, objects, and types of interactions—while each task, whether it is cleaning, cooking, or organizing, demands specific actions and priorities. To handle this complexity, agents must generalize knowledge from one task or environment and successfully apply it in others. The goal is to maintain consistent performance across contexts while adapting to new challenges, without losing the ability to execute previously learned tasks.

Achieving this versatility requires a combination of several reinforcement learning techniques. **Transfer learning**, **multi-task RL**, and **domain generalization** are essential for enabling agents to adapt across varying tasks and environments. Techniques like **sim-to-real transfer**, **unsupervised RL**, and **domain randomization** help generalize policies across different environments. Multi-task RL methods, such as **meta-RL**, **lifelong RL**, and **hierarchical RL**, enable agents to acquire reusable skills and adapt to new tasks without starting from scratch. Additionally, **task-agnostic RL**, driven by intrinsic motivations, supports the development of versatile behaviors that can be applied across a wide range of tasks, even without explicit supervision. To tackle the complexities of multi-task and multi-environment learning, agents must effectively manage the **stability-plasticity trade-off**, ensuring they retain learned knowledge from prior tasks while efficiently acquiring new skills. They also need mechanisms to dynamically prioritize objectives, particularly when handling conflicting tasks within a single environment. Techniques such as **compositional RL**, which modularizes policies for flexible recombination, and **hierarchical transfer**, which extends subtask learning across tasks and environments, offer promising solutions. Moreover, combining **multi-objective RL** with **meta-learning** can help balance competing priorities and facilitate quick adaptation to new tasks and contexts. **Task-agnostic exploration** and **lifelong learning** frameworks further enhance agents' ability to generalize and integrate diverse skills across multiple scenarios.

Nevertheless, integrating these techniques presents several challenges. One key issue is managing the balance between task-specific knowledge and environment-specific contexts, which can lead to problems like **negative transfer**, where knowledge from one task or environment hinders learning in another. Scaling learning frameworks to accommodate many tasks and environments also requires robust memory management and adaptable architectures. Additionally, balancing **exploration and exploitation** becomes more difficult in multi-task settings, where agents must effectively prioritize tasks while ensuring learning remains effective across all objectives.

In short, while integrating these techniques provides the foundation for enabling true cross-scenario and cross-task capabilities, overcoming the challenges of knowledge transfer, scalability, and task prioritization is essential for achieving adaptable, intelligent embodied AI systems capable of operating in complex real-world environments.

2.5 Embodied foundation models

The progression of embodied intelligence has culminated in the emergence of "Embodied Foundation Models". While earlier sections discussed isolated capabilities, the field is now converging towards unified vision-language-action (VLA) architectures. However, a critical consensus is emerging: while imitation learning (IL) provides a strong initialization, reinforcement learning post-training is essential for achieving robustness and surpassing the limitations of demonstration data. We analyze this paradigm shift through three interconnected frontiers.

2.5.1 *From specialized agents to generalist VLA models*

The quest for cross-environment capability (Subsection 2.4) has led to the development of generalist VLA models. A prime example is π_0 [150], which marks a departure from traditional diffusion policies by utilizing flow matching to model continuous action distributions. This architecture enables zero-shot transfer across diverse robot morphologies—from quadrupeds to dexterous hands—validating the scaling laws of physical intelligence.

Simultaneously, deployment efficiency remains a barrier. Projects like **OpenVLA** [151] address this by adapting quantized language backbones (e.g., Llama) for robotic control. These models demonstrate that generalist reasoning and high-frequency manipulation (10–20 Hz) can coexist on edge devices, bridging the gap between massive foundation models and real-world constraints.

2.5.2 *RL post-training for VLAs*

While VLA models trained via behavioral cloning act as powerful "sponges" for diverse data, they suffer from compounding errors and are fundamentally upper-bounded by the quality of human demonstrations. To break this ceiling, the field is pivoting towards RL post-training to refine these pre-trained policies.

Reinforced fine-tuning (RFT). Recent approaches integrate RL directly into the fine-tuning process of VLAs to optimize task success rather than just perplexity. For instance, ConRFT [152] introduces a consistency policy to

regularize the RL update, preventing the catastrophic forgetting of language capabilities. Similarly, RL-100 [153] demonstrates that large-scale, sample-efficient RL on real robots can significantly outperform pure imitation learning baselines, particularly in contact-rich manipulation tasks where precision is paramount.

Value-aware architectures and flow-RL. Equipping VLAs with value estimation capabilities is another rising trend. Methods like VLA-Critic [154] incorporate critic networks to guide the generation process, allowing the model to "self-correct" based on long-term utility. Furthermore, applying RL to flow-based policies (like π_0) presents unique challenges compared to discrete action spaces. Emerging studies [155] are exploring gradient-based guidance to enable end-to-end RL updates for flow-matching backbones, aiming to combine the expressivity of generative models with the optimality of reinforcement learning.

Optimization challenges in flow-based policies. A distinct challenge arises when applying RL to emerging *flow-based architectures* (such as π_0). Unlike discrete or Gaussian policies, flow-matching models generate actions via continuous vector fields (ODEs), rendering standard policy gradient methods (e.g., PPO) computationally intractable or unstable due to the difficulty in likelihood estimation. Addressing this optimization gap is critical for the next generation of foundation models. Recent theoretical frameworks, such as the work by Lv et al. [156], provide potential solutions to this problem, offering rigorous mechanisms to align flow-matching dynamics with reinforcement learning objectives for effective fine-tuning.

2.5.3 *Generative world models*

Supporting these RL post-training paradigms requires massive interaction data, which is costly to collect in the real world. Generative interactive environments (Genie) [157] pioneers the use of generative world models to hallucinate interactive physics from internet video. By serving as a learned simulator, it allows agents to practice and acquire RL signals in a latent dreamscape, effectively solving the data scarcity bottleneck for post-training and enabling "sim-to-real 2.0".

2.6 Real-world reinforcement learning

Direct training on physical hardware circumvents the reality gap but necessitates overcoming bottlenecks in sample efficiency, autonomy, and reward specification. Recent advancements address these through high-efficiency optimization architectures, autonomous reset mechanisms, and foundation model integration as illustrated in Table 4.

2.6.1 *Sample-efficient policy optimization*

Achieving industrial-grade reliability within limited physical interaction time requires algorithmic innovations that maximize data utility and control stability.

• **RL-100** [153] establishes a zero-failure standard by integrating consistency distillation into a hierarchical pipeline. This compresses diffusion inference steps, enabling high-frequency control (>50 Hz) essential for dynamic tasks.

• **SERL** [158] provides a robust infrastructure for contact-rich manipulation. By combining off-policy SAC with regularized learning from demonstrations, it efficiently anchors value estimates to human priors, enabling rapid policy convergence.

• **CrossQ** [159] introduces a minimalist architecture that removes target networks in favor of batch normalization. This allows for aggressive update-to-data (UTD) ratios, enabling agents to learn complex locomotion behaviors in approximately 8 min.

2.6.2 *Autonomous reset-free learning*

Scalability in real-world learning is fundamentally constrained by the need for manual interventions. To enable continuous unsupervised training, systems such as MEDAL++ [160] implement forward-backward policy cycles. By learning to reset the environment as part of the task, agents can practice skills like cloth manipulation continuously for hours without human oversight.

2.6.3 *Foundation model fine-tuning*

Pre-trained VLA models often lack the precision required for physical actuation. Post-training frameworks bridge this gap by leveraging foundation models for rewards and regularization.

• **Reward generation:** Frameworks like PA-RL [161] and RoboFuME [162] utilize vision-language models (VLMs) to generate dense, semantic reward signals, enabling the fine-tuning of policies where ground-truth state estimation is unavailable.

Table 4 Comparison of real-world reinforcement learning methods. Key mechanisms focus on overcoming efficiency, autonomy, and forgetting constraints.

Method	Key mechanism	Data source	Primary constraint
RL-100 [153]	Consistency distillation	Real demos + Online	Reset mechanism
SERL [158]	SAC + RLPD	Real demos + Online	Expert data cost
CrossQ [159]	Batch normalization	Real online only	Safe exploration
MEDAL++ [160]	Forward-backward cycles	Real online only	Task reversibility
PA-RL [161]	VLM rewards	Pre-trained + Online	Inference latency
ConRFT [152]	Consistency regularization	Pre-trained + Online	Catastrophic forgetting

• **Consistency regularization:** ConRFT [152] addresses the catastrophic forgetting inherent in fine-tuning large models. It introduces a consistency policy to regularize updates, ensuring the agent retains generalist language capabilities while adapting to specific physical dynamics.

3 Virtuoso capabilities in embodied intelligence and future directions

To advance embodied intelligence, the focus shifts beyond merely enabling agents to perform predefined tasks. For intelligent systems to thrive in dynamic and complex environments, they need to exhibit higher-order cognitive abilities that reflect human-like flexibility, adaptability, and decision-making. These abilities, referred to as "virtuoso capabilities", extend far beyond basic task execution. They enable agents to autonomously generate new tasks, align with human values, and engage in reflective, ethical reasoning.

Imagine a robot that can not only complete a variety of chores but also respond to unexpected situations, make moral decisions, and adapt its behavior to new contexts. Instead of blindly following instructions or resorting to trial and error, such a robot could recognize when a human request is unreasonable or unsafe, flagging the task's irrationality and offering suggestions for a safer or more efficient course of action. Whether adjusting its movement in a cluttered environment or revising its plan after a change in human intent, these virtuoso capabilities would allow the robot to function effectively across diverse environments. This section explores four critical areas of virtuoso capabilities for embodied intelligence: **morphological intelligence**, **human-like values and ethics**, **reflective equilibrium**, and **spontaneous reward generation**. By exploring these areas, we emphasize their scientific importance, the challenges they introduce, and how they contribute to the development of more adaptive, ethical, and intelligent systems. Each capability is illustrated through practical scenarios that demonstrate its potential impact and application in real-world settings.

3.1 Morphological intelligence (MI)

MI enables robots to dynamically adjust their physical form to meet the demands of different tasks and environments. Inspired by biological systems, where animals like octopuses or caterpillars modify their bodies to navigate or manipulate objects, MI allows robots to optimize their morphology in real-time to solve complex problems. This flexibility extends the capabilities of robots far beyond traditional rigid designs, enabling them to operate autonomously in dynamic, unstructured environments, such as search-and-rescue missions, space exploration, and deep-sea robotics. Please refer to Figure 3 for a brief overview.

3.1.1 *RL for morphological intelligence*

The integration of RL with morphological intelligence has been pivotal in allowing robots to autonomously discover and refine their morphological adaptations. RL provides a flexible framework for robots to explore various configurations and learn how to modify their physical form in response to task-specific goals and environmental conditions. Through continuous interaction with the environment and feedback from task outcomes, robots can optimize their morphology over time.

In soft robotics, RL could be used to enable robots with flexible actuators to autonomously modify their morphology for delicate tasks, such as picking up fragile objects or navigating through tight spaces. With RL, robots could explore various body configurations, learning to optimize their shape in real time to maximize task success. Instead of manually programming specific body behaviors, RL could allow robots to discover efficient strategies by trial and error, adjusting their form based on environmental feedback and task-specific goals [163].

Similarly, in modular robotics, RL could guide robots composed of reconfigurable units to adapt their morphology on the fly. These robots could learn to self-organize their modules depending on the task at hand—whether

for locomotion, object manipulation, or multi-agent coordination. RL could empower robots to autonomously reconfigure themselves without human intervention, optimizing their body for a wide range of applications by receiving rewards based on successful task completion [20].

Furthermore, biologically-inspired control systems, when combined with RL, could enhance robots' ability to autonomously adjust their morphology in response to environmental stimuli, similar to how animals adapt their bodies. RL can facilitate continuous learning, allowing robots to refine their body structures and movements based on sensory feedback, thus improving adaptability in complex or unpredictable environments [163].

Additionally, subequivariant RL provides a promising avenue for optimizing morphology in complex 3D environments. In this approach, robots could learn to adjust their form while respecting physical symmetries, ensuring that their adaptations are both stable and efficient. In environments involving multiple interacting entities–such as robots working alongside humans or other machines—RL could enable robots to modify their morphology in response to these dynamics, optimizing collaboration and interaction [164, 165].

3.1.2 *Challenges and future directions*

Despite its promise, MI faces several challenges that must be overcome for practical implementation in robotics.

Design and fabrication of robots capable of real-time morphological adaptation remains a major hurdle. These robots require flexible, durable materials and actuators that allow for the robot to change shape while maintaining stability and performance. Developing such actuators, materials, and sensors that enable real-time shape modifications is complex. Here, RL can play a vital role. RL allows robots to autonomously learn which body configurations are most effective through trial and error, without requiring explicit human guidance. By interacting with their environment, robots can improve their morphological design iteratively, adjusting their body structure based on feedback, and ultimately optimizing for robustness and functionality over time. RL's ability to guide autonomous exploration of design space helps in overcoming the lack of manual expertise required to develop adaptive robots [163].

Real-time control is another significant challenge. Robots need to process large amounts of sensory data quickly and autonomously decide on the best morphological changes for task success. This involves fast decision-making and dynamic adaptation to changing environmental conditions. RL helps in this regard by enabling online learning: robots can continuously refine their control policies through experience. By receiving rewards based on task success, robots can learn which morphological adaptations yield the best performance in real-time. This continuous feedback loop enables robots to improve their decision-making over time, ensuring that adaptations are both efficient and stable under varying conditions [164, 165]. RL's ability to adapt continuously based on real-time feedback from sensors makes it particularly suited for controlling dynamic, adaptive systems.

Task generalization remains a significant issue. A robot that learns to adapt its morphology to one specific task or environment may struggle to apply that knowledge to new, unforeseen tasks. This is particularly true when robots encounter environments with different constraints or interact with new entities. Meta-learning and transfer learning—subfields of RL—offer promising solutions. Meta-learning enables robots to learn how to learn, allowing them to adapt their morphological strategies to new tasks more efficiently. In this way, robots can transfer learned behaviors from one task or environment to another without the need for retraining from scratch. Similarly, transfer learning allows robots to apply knowledge gained in one environment to similar, yet distinct, environments, reducing the need for extensive retraining [166, 167]. This allows MI to scale across a wider range of tasks and environments, making robots more versatile.

Looking ahead, **multi-modal sensing** and **adaptive learning algorithms** will be crucial for improving MI. Multi-modal sensors will provide richer, more detailed data from the robot's environment, enabling more precise morphological adaptations. RL-based systems can incorporate these sensor inputs to optimize morphological changes based on a more comprehensive understanding of the environment. Moreover, collaborative MI, where robots adjust their morphology in response to interactions with other agents (humans or robots), presents an exciting avenue for research. RL techniques that integrate collaborative learning could enable robots to optimize their morphology based on team dynamics, improving collaborative task performance in multi-agent environments. Robots could learn to adapt not only to their own environment but also to the presence and actions of other agents in real-time.

Finally, **evolutionary approaches** to MI, where robots evolve their morphology through competition or environmental adaptation, offer new pathways for optimizing both morphology and behavior. These methods, combined with RL, could lead to autonomous robotic evolution, allowing robots to adapt their physical form over time to meet new challenges without human intervention. RL could guide this evolutionary process, enabling robots to evolve more efficient and task-specific morphologies autonomously, based on long-term environmental feedback and

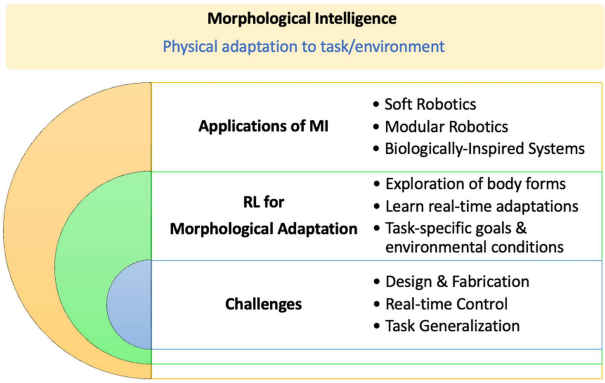

Figure 3 Morphological intelligence (MI). This diagram illustrates the applications and challenges of MI, highlighting the role of RL in enabling robots to adapt their morphology autonomously for task-specific and environmental demands.

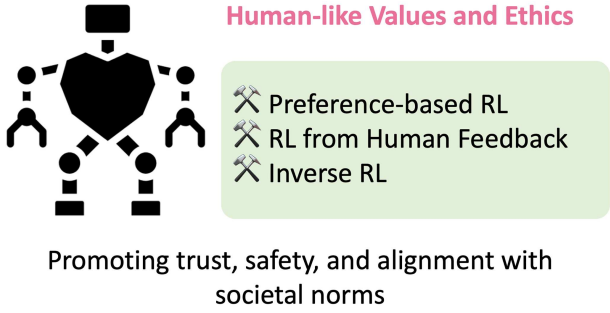

Figure 4 Human-like values and ethics. Preference-based RL, RLHF, and IRL offer promising approaches to align robots' behaviors with human preferences, ethical guidelines, and societal norms. These methods allow robots to adapt and learn ethical decision-making, addressing challenges such as value misalignment and evolving human expectations.

competition.

3.2 Human-like values and ethics

Equipping embodied AI systems with human-like values and ethics is essential for ensuring that robots interact with humans and the environment in a socially responsible and context-aware manner. As robots become more integrated into everyday life, especially in collaborative environments, their decisions will have direct consequences on human well-being, safety, and societal norms. For instance, autonomous vehicles must make split-second decisions that can have significant ethical implications, such as how to prioritize the safety of passengers versus pedestrians. Similarly, robots assisting in healthcare or elder care must understand human preferences and respect social and ethical guidelines, including privacy, autonomy, and fairness. Therefore, embedding human-like values into robots is crucial for promoting trust, safety, and alignment with societal norms.

3.2.1 *RL for instilling human-like values and ethics*

RL offers promising approaches to instilling ethical decision-making in robots, as shown in Figure 4.

One powerful method is **preference-based RL**, where agents learn to make decisions by aligning their actions with human preferences or ethical guidelines. In this framework, human feedback—often in the form of preferences or ratings on different outcomes—guides the agent's learning process, allowing it to optimize for behaviors that are consistent with human values. This approach can be applied in various domains, from healthcare, where robots must learn patient preferences, to autonomous vehicles, where they must learn safe and ethically acceptable driving behaviors [168, 169].

Another method that is gaining traction is **reinforcement learning from human feedback (RLHF)**. RLHF enables robots to learn ethical behavior by directly incorporating human-provided feedback during the training process. Instead of relying solely on predefined reward functions, RLHF allows robots to adjust their actions based on more complex, human-derived signals. This is especially useful for teaching robots values that are hard to formalize in traditional reward functions, such as empathy, fairness, or respect for human rights. In RLHF, the agent receives feedback in the form of human preferences or corrective signals, which it then uses to refine its policy [168, 170, 171]. This approach has been successfully used to train AI models to align with human values, such as avoiding harmful behaviors and prioritizing human safety.

Additionally, **inverse reinforcement learning (IRL)** can be used to extract human-like values by observing and modeling human behavior. By observing the decisions humans make in different scenarios, robots can infer the underlying value system guiding those decisions and apply it to their own decision-making processes. This method is especially useful when it is difficult to explicitly define reward functions that encapsulate human values. IRL helps robots understand the ethical framework of human actions, which can then be encoded into their decision-making algorithms [172].

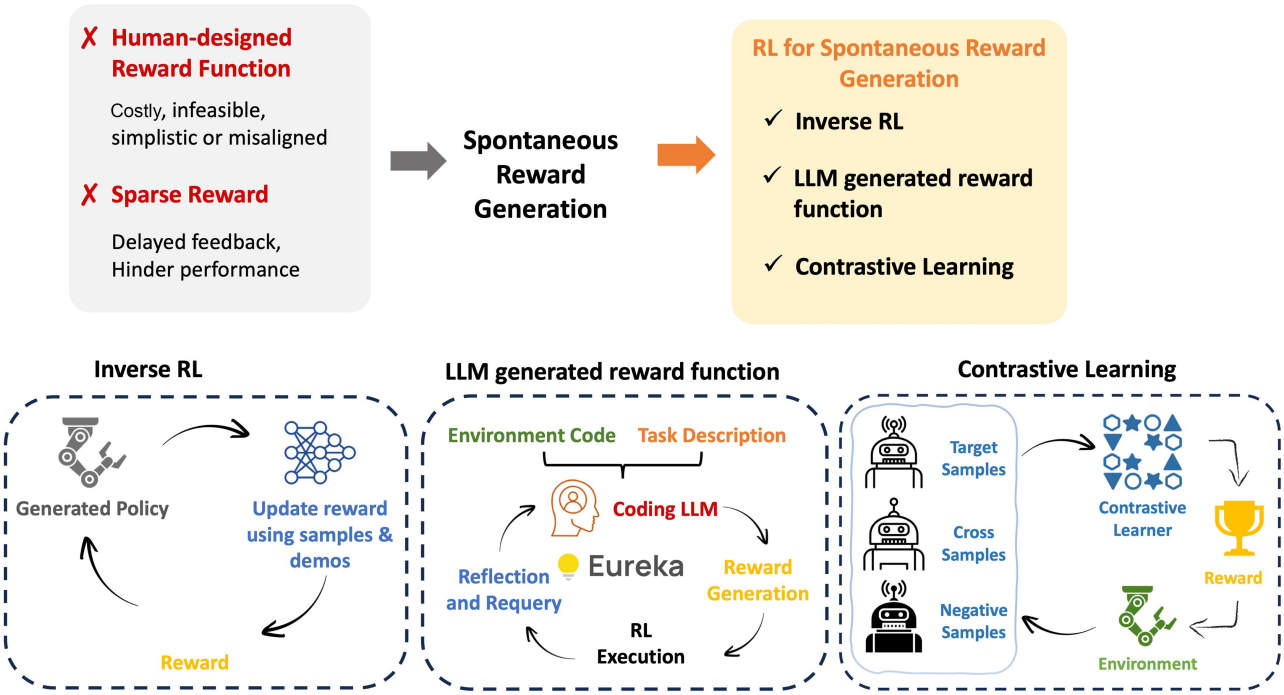

Figure 5 Spontaneous reward generation. Spontaneous reward generation aims to overcome the limitations of manually designed and sparse rewards by automatically producing informative reward signals for robot learning. As illustrated, representative approaches include inverse reinforcement learning, which updates rewards from samples and demonstrations; LLM-generated reward functions, which use task descriptions and environment code to synthesize and refine rewards, and contrastive learning, which derives rewards by comparing target, cross, and negative samples.

3.2.2 *Challenges and future directions*

Despite the progress in these areas, challenges remain. One of the primary difficulties is value misalignment, where the robot's learned values might diverge from the true intentions of humans. This can occur due to biased feedback, incomplete value representations, or the robot's misinterpretation of human preferences. Additionally, ensuring that robots can adapt to evolving human values over time is another challenge, as societal norms and ethical guidelines change. Developing RL-based techniques that can continuously learn and adapt to shifting ethical standards without violating previously learned principles is an ongoing research area.

Furthermore, MORL offers a robust framework for embedding ethical standards into agent behavior. Ethical decision-making often involves trade-offs between conflicting goals—such as maximizing operational efficiency versus adhering to safety or privacy norms. Instead of treating ethics merely as a binary constraint, MORL allows agents to treat ethical alignment as a distinct objective alongside task performance. By optimizing for the Pareto front of these competing objectives, embodied agents can dynamically adjust their policies to maintain high ethical standards without disproportionately sacrificing utility, adapting their "moral compass" based on the context's specific requirements.

To address these challenges, future directions include improving human-in-the-loop methods, where human feedback plays a more active role in continuous learning, and developing multi-agent reinforcement learning systems that allow robots to collaboratively learn ethical guidelines in shared environments. Such systems could involve multiple agents, both human and robotic, providing feedback to ensure that ethical behavior is consistently aligned with human expectations across various contexts.

3.3 Spontaneous reward generation

The reward function is a key factor that shapes an agent's learning process and action selection as shown in Figure 5. It directly influences the agent's ability to align its actions with task objectives and maximize overall performance [173]. The reward signal provides feedback to the agent, indicating how well its actions match the desired outcome, thus encouraging behavior that leads to goal achievement. In spite of this, designing effective reward functions is a complex task, particularly in dynamic and real-world environments where goals may be ambiguous, multi-objective, or subject to varying constraints [174].

Most RL systems rely on **human-designed reward functions**, where researchers or engineers explicitly define the rewards and punishments for various actions. This approach is common in many applications, such as robotics and video games, where designers craft detailed reward structures based on known objectives. Another common scenario is **sparse-reward settings**, where agents only receive feedback after achieving significant milestones, such as reaching a goal or completing a complex task. While human-designed and sparse rewards are widely used, they come with several limitations.

Even so, human-designed reward functions can be too simplistic or misaligned with broader, more complex human values, making it difficult to capture the full spectrum of behavior expected from an agent in dynamic, real-world settings. For example, in autonomous driving, designing a reward function that captures every nuance of safe driving behavior, while avoiding unintended consequences (such as overly cautious driving), can be an intricate task. Sparse rewards, on the other hand, present the challenge of credit assignment—the difficulty in determining which actions led to the reward, especially when feedback is delayed. In environments with sparse or delayed rewards, an agent may struggle to connect its actions with the consequences they caused, hindering effective learning [175, 176].

3.3.1 *RL for spontaneous reward generation*

Spontaneous reward generation aims to address the limitations of human-designed and sparse reward functions by enabling agents to generate their own reward signals based on environmental interactions and intrinsic motivations. Several approaches have been explored to facilitate this kind of reward generation.

IRL. IRL revolves around deducing reward functions that rationalize expert demonstrations, providing a robust framework for imitation learning. Reward design is pivotal in IRL as it defines the task in a succinct and transferable way, enabling policies to generalize across different scenarios. By emphasizing the reward function over direct policy modeling, IRL achieves structured policy search, reducing the policy space to those optimal under a learned reward function. This abstraction allows IRL methods to leverage the reward function for policy transfer across domains, making it instrumental in tasks such as robotics and human-computer interaction. Recent advances, such as those that utilize expert visitation distributions, aim to alleviate inefficiencies by resetting the learner to states frequently visited by the expert. These strategies, exemplified by methods like Moment Matching by Dynamic Programming (MMDP), streamline the policy optimization process, reducing the computational complexity of reward inference. By focusing on effective reward modeling, IRL not only facilitates robust policy learning but also enhances the interpretability of decision-making processes, critical in applications like autonomous systems and predictive modeling [177].

LLM generated reward function. LLMs play a transformative role in RL by automating and optimizing the critical task of reward design. The development of frameworks such as EUREKA [178] demonstrates that LLMs can generate high-performing reward functions through their code generation and in-context learning capabilities. These models excel in producing executable reward functions without requiring extensive task-specific prompting or templates, enabling generality across diverse RL environments. By iteratively refining rewards via evolutionary search and reflective feedback mechanisms, LLMs achieve human-level or even superhuman performance in designing rewards that effectively align agent behaviors with desired objectives. This approach eliminates reliance on trial-and-error reward engineering, reduces the likelihood of unintended agent behaviors, and supports robust learning even in complex environments like high-dimensional robot control tasks. These innovations highlight the synergy between LLMs and RL, particularly in addressing the long-standing challenge of reward design, paving the way for more efficient and scalable RL solutions.

Contrastive learning. Contrastive value learning (CVL) represents a novel approach in reinforcement learning, contrasting traditional model-based and model-free methods. It introduces an implicit, multi-step model of environmental dynamics, eschewing the need for reward functions or the high-dimensional prediction tasks typical of conventional models. Unlike classical temporal difference (TD) learning, CVL directly estimates the value of actions through an implicit representation of discounted state occupancy measures, trained via contrastive learning. This method avoids complex rollouts and high-dimensional observations, scaling effectively to high-dimensional tasks. By leveraging this forward-looking estimation of state dynamics, CVL aligns state occupancy measures with reward signals, yielding a computationally efficient yet robust estimation of Q-values. This innovative approach demonstrates superior performance in offline reinforcement learning scenarios, particularly in complex, continuous control benchmarks, as highlighted in empirical evaluations [179]. Contrastive learning in RL is emerging as a powerful paradigm, especially in contexts where explicit reward signals are unavailable or challenging to define. By framing the RL problem through contrastive learning, state-action pairs can be encoded into representations where the similarity between current and future states naturally serves as an implicit reward signal. This mechanism aligns well with goal-conditioned RL tasks, where proximity in the representation space correlates with task success. Con-

trastive learning methods leverage positive examples (e.g., states from the same trajectory) and negative examples (randomly sampled states) to shape the representation space, effectively serving as a surrogate for traditional reward functions. This framework eliminates the dependency on manually designed rewards, enabling the RL algorithm to focus on the intrinsic structure of trajectories. Empirical results have demonstrated that contrastive RL methods achieve robust performance, particularly in high-dimensional tasks such as image-based RL scenarios, by optimizing for state transitions that inherently encode goal-oriented behaviors [180].

3.3.2 *Challenges and future directions*

Despite the promising advancements in spontaneous reward generation, several challenges remain, and future research will need to address these obstacles.

Cognitive science as intuition. A promising future direction is to draw insights from cognitive science to understand how humans generate intrinsic rewards and apply these processes in artificial agents. Psychological theories suggest that human learning is driven by curiosity, exploration, and the desire for competence, rather than just external rewards. This could be integrated into RL systems, where agents are motivated to explore and learn autonomously, generating rewards based on curiosity or novelty [181]. Incorporating cognitive models of emotional and social rewards could further align RL systems with human-like decision-making processes.

Ensuring safety and ethical alignment. One of the critical challenges in spontaneous reward generation is ensuring that the agent's autonomously generated rewards align with human values and ethical standards [182–184]. There is a risk that agents may generate rewards that encourage unintended, unsafe, or unethical behaviors. Future research must focus on ensuring that these dynamic reward functions do not inadvertently promote harmful actions. Combining human feedback, ethical guidelines, and safety constraints with spontaneous reward generation systems will be crucial for building trustworthy and ethical autonomous agents.

3.4 Reflective equilibrium

3.4.1 *Concept*

Reflective equilibrium, a hallmark of advanced intellectual and ethical reasoning, refers to achieving coherence among beliefs by iteratively adjusting overarching principles and particular judgments. Rooted in existential philosophy, this concept addresses profound ethical challenges that humans frequently face, such as balancing individual autonomy with societal welfare or reconciling immediate actions with long-term consequences [185, 186]. In human decision-making, reflective equilibrium often involves re-evaluating goals, priorities, or ethical frameworks when faced with contradictions or failures, as shown in Figure 6. Conversely, robotic systems typically lack this capacity, as most are designed to follow predefined instructions rigidly. Consequently, these systems struggle to adapt or resolve conflicts in complex, contradictory real-world scenarios [187].

3.4.2 *Future vision of reflective robot*

Robots with reflective capabilities can reassess and adjust their goals in response to unforeseen challenges, addressing two key areas: deeming impossibilities and deeming immorality, as shown in Figure 6.

Deeming impossibilities. Reflective capabilities in robots allow them to adapt and revise their goals in response to unforeseen challenges. In industrial applications, for instance, a robotic system may face situations where its predefined goal becomes unachievable due to material defects or equipment malfunctions. Rather than rigidly continuing, a reflective robot could pause, reassess the situation, and either recalibrate its approach or seek human intervention [188]. This dynamic adaptability addresses what would otherwise be deemed an "impossibility"—overcoming obstacles that prevent the task's completion, thus ensuring operational continuity.

Deeming immorality. For instance, in autonomous systems, reflective reasoning enables ethical decision-making in morally complex scenarios. For instance, an autonomous vehicle might face an unavoidable accident where it must choose between two potential victims. A reflective system would weigh ethical considerations, such as minimizing harm, and make a decision in line with societal moral standards [189–191]. Beyond immediate choices, such a system could learn from these experiences, refining its ethical principles over time. This capacity for moral reflection allows robots to avoid actions that would be deemed immoral, ensuring their decisions align with human values and societal expectations.

3.4.3 *Challenges and future directions*

Despite its significance, the integration of reflective equilibrium into artificial systems remains a nascent area of research. Most current efforts are limited to preliminary explorations, such as basic re-evaluation of goals or tasks,

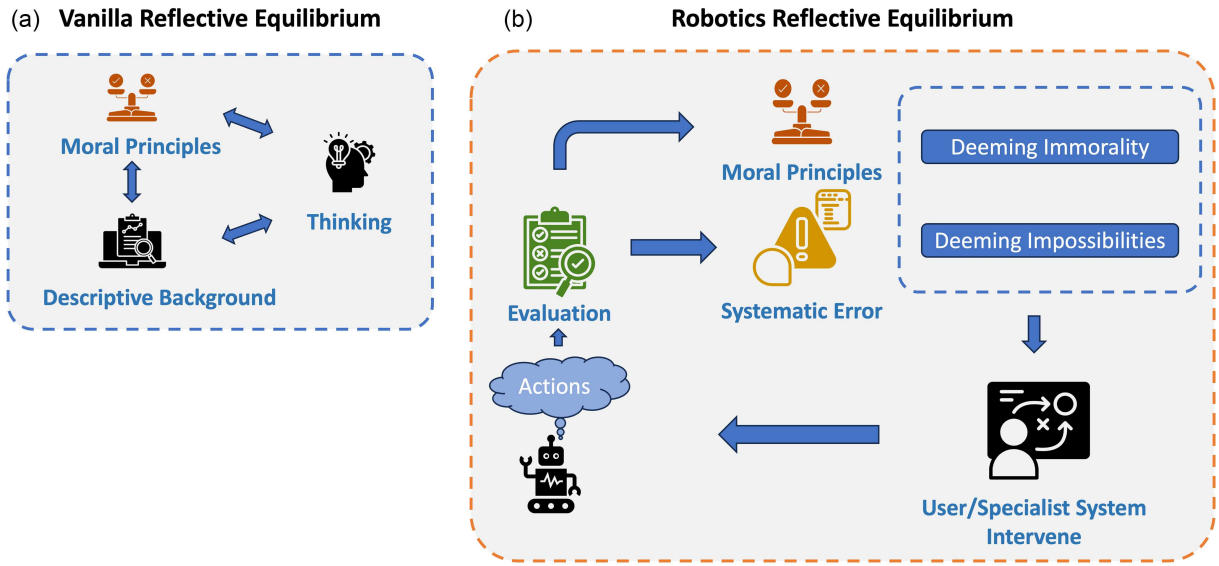

Figure 6 Illustration of reflective equilibrium. (a) Reflective equilibrium is a well-established theory in philosophy, generally referring to a mutual feedback loop between moral principles, reasoning, and the descriptive background. (b) In robotics, reflective equilibrium is expected to be achieved through a loop involving actions and evaluations. This loop determines whether an action violates systematic settings, restrictions, or certain moral principles. If a violation occurs, a feedback loop—either induced by the user or a specialized system—intervenes to help the robot achieve reflective equilibrium.

which lack the depth and adaptability required for true reflective reasoning [192–194]. These approaches typically address surface-level inconsistencies but fail to achieve the dynamic coherence and ethical adaptability necessary for real-world applications.

The challenges are particularly pronounced in embodied AI agents, which interact with the physical world in ways that magnify the ethical and practical stakes of their decisions. Unlike traditional reinforcement learning systems designed for single-task, single-scenario environments, embodied AI agents must navigate diverse, unpredictable situations. The absence of reflective capabilities in these systems results in a critical gap in their ability to handle complex ethical dilemmas, adapt to evolving environments, or make value-aligned decisions.

Significant advancements are needed to transition from traditional reinforcement learning to embodied reinforcement learning capable of cross-task and cross-environment adaptability. Future research must prioritize developing mechanisms that enable agents to engage in iterative self-assessment, dynamically adjust their goals, and integrate ethical considerations into their decision-making processes. This will be essential for bridging the gap between rigid task execution and the nuanced, reflective reasoning required for advanced intelligence.

4 Barriers and pathways in embodied reinforcement learning

In the exploration of leveraging RL for enhancing embodied intelligence, several challenges have been identified across different levels of capability and within the development of virtuoso abilities. These challenges highlight the complexities and limitations currently faced in the field, as well as the opportunities for future research and innovation.

4.1 Sample efficiency

Sample efficiency refers to the ability of RL algorithms to learn effective policies with minimal interaction with the environment. This concept is crucial for real-world applications in embodied intelligence, where data collection is often expensive, time-consuming, and risky. Improving sample efficiency is essential for deploying RL in complex environments such as robotics, autonomous vehicles, and healthcare, where excessive trial-and-error learning may lead to hardware degradation, safety concerns, or substantial computational overhead.

4.1.1 *Challenges in improving sample efficiency*

The primary challenge in improving sample efficiency arises from the large number of interactions typically required for RL algorithms to converge to an optimal policy. This issue is exacerbated in real-world environments, where,

(1) high cost of interaction: real-world data collection involves operational costs, physical wear, and safety risks, which make extensive exploration impractical; (2) slow learning in high-dimensional spaces: complex environments, such as those involving vision or high-dimensional control, require extensive data to explore effectively; (3) safety and ethical constraints in exploration: in critical domains like healthcare or autonomous driving, exploration must adhere to safety standards, limiting the agent's ability to perform risky actions during training [195–197].

4.1.2 *Methods for improving sample efficiency*

Several methods have been proposed to reduce the amount of data required for RL agents to learn effectively.

Model-based RL (MBRL). MBRL methods, which directly learn a transition model from significantly fewer samples and then derive the optimal policy from this model, have emerged as an attractive alternative for small-data and more practical scenarios. These MBRL algorithms can generally be grouped into several categories to highlight the range of uses of predictive models [198]: Dyna-style algorithms [199–202], policy search with back-propagation through time [51, 203–205] and shooting algorithms [206, 207].

Recently, there has been growing interest in developing large-scale world models. A world model refers to a system that can simulate or predict the behavior of the real world. These models are crucial for training agents that can operate effectively in the real world, as they allow for extensive experimentation within a virtual environment, thus avoiding the risks and costs associated with real-world interaction [208, 209].

Experience replay and off-policy learning. Storing and reusing past experiences allows agents to learn from data generated in earlier stages of training, thus improving sample efficiency. Experience replay has evolved from basic vanilla replay [10] to more sophisticated priority-based techniques, including single proxy [210], learnable priority scores [211], similarity diversity [212, 213], and goal-based strategies [214]. Recent advances in scaling the replay ratio—i.e., the number of parameter updates per environment interaction—have further enhanced sample efficiency by increasing the replay ratio to higher values [215]. Off-policy learning builds on experience replay. However, distribution shifts can introduce bias in the value function, leading to over- or under-estimation. To mitigate these biases, various techniques have been proposed. Double Q-learning addresses overestimation bias [216, 217], in-sample learning within experience replay helps reduce underestimation bias by bootstrapping Q-values [41], and methods like DICE compensate for distribution shifts [218].

Bootstrap RL with demonstrations. Bootstrap RL with demonstrations significantly improves sample efficiency by integrating expert-provided demonstrations with RL. Initially, the agent learns from a set of high-quality demonstrations, which provides a strong starting point and reduces the need for extensive exploration. This allows the agent to bypass much of the costly trial-and-error phase that would typically be required. After the agent has learned from the demonstrations, it continues refining its policy through bootstrapping, improving on the demonstrated behavior through further interaction with the environment. Such an approach accelerates the learning process by leveraging expert knowledge, enabling the agent to build on this prior data while still benefiting from its own exploration [219–221].

4.2 Generalization and transferability

As discussed in Section 2, generalization and transferability constitute essential capabilities for embodied intelligence systems, particularly when agents are required to operate in dynamic and unpredictable environments. Generalization allows agents to adapt learned behaviors to new contexts, while transferability enables the efficient transfer of knowledge across distinct tasks.

While techniques such as domain randomization and meta-learning have historically promoted these capabilities, the focus of contemporary research shifts towards large-scale systems where the curse of dimensionality complicates policy adaptation. Recent advances in scalable reinforcement learning aim to exploit the inherent structure of these systems [222–224]. However, ensuring robustness in such high-dimensional spaces requires moving beyond heuristic methods to rigorous theoretical frameworks that quantify generalization bounds.

4.2.1 *Theoretical bounds via distributionally robust optimization*

To address the brittleness of policies under distributional shifts, the field adopts distributionally robust optimization as a fundamental framework. Rather than optimizing the expected return under a single training distribution, this approach maximizes the worst-case performance within a specified ambiguity set. Contemporary research leverages the geometry of Wasserstein ambiguity sets to define a probability ball of radius ϵ around the empirical distribution defined as

$$\max_{\pi} \min_{P|W_p(P,P_{train})\leqslant\epsilon} J_R(\pi, P) \quad \text{s.t.} \quad \max_{P|W_p(P,P_{train})\leqslant\epsilon} J_C(\pi, P) \leqslant d. \tag{1}$$

A pivotal theoretical insight involves the duality that tractably converts this infinite-dimensional optimization into a finite regularization problem. Theoretical analysis demonstrates that the radius ϵ effectively serves as a Lipschitz regularization coefficient on the value function and thereby provides a mathematically guaranteed upper bound on constraint violations during transfer to perturbed environments [225]. This formulation transforms the generalization gap from an empirical unknown into a computable quantity controlled by the radius parameter and provides the theoretical grounding necessary for scaling embodied agents to unmodeled real-world scenarios.

4.2.2 *Quantification of optimality gaps*

The enforcement of strict generalization and safety constraints imposes a quantitative reduction in performance formally known as the price of safety. Empirical evaluations on constrained benchmarks reveal a convex Pareto frontier where the marginal cost of safety increases sharply as the violation threshold approaches zero. Specifically, enforcing a strict constraint violation rate of less than 0.1% typically incurs a 15% to 25% reduction in asymptotic reward compared to unconstrained baselines [226]. This optimality gap quantitatively reflects the restriction of the explorable state space as the policy is forced to maintain a conservative buffer distance from hazardous boundaries. Furthermore, this tension manifests in sample complexity where safe agents require approximately 30% more interaction steps to converge as they must navigate the delicate saddle points between reward maximization and constraint satisfaction.

4.3 Safety and robustness

Safety and robustness in embodied RL are important due to the significant consequences that can arise from harmful actions in real-world environments, especially in safety-critical domains such as autonomous vehicles, robotics, and healthcare. The deployment of RL in these areas requires algorithms that not only maximize rewards but also ensure that agents can safely interact with environments, avoiding unsafe states or actions, and maintaining stability at the same time.

In practice, general RL algorithms often fall short in ensuring safety and robustness. Traditional RL approaches allow agents to explore any behavior that leads to performance improvement, which leads to potential safety issues. Moreover, RL agents frequently rely on excessive amounts of data and struggle to generalize beyond their initial training process, leading to unsafe behavior in real-world environments when encountering out-of-distribution (OOD) scenarios or actions. These challenges arise because agents often overfit to limited training conditions, fail to model uncertainties, and lack the ability to adapt to new or unexpected situations.

To mitigate these issues, current techniques focus on several key areas. Constrained RL incorporates safety constraints [227,228] into the learning process to ensure adherence to predefined safety criteria, while safe exploration enforces hard constraints to prevent hazardous actions, such as collisions and system crashes. Additionally, concepts in control theory, like Lyapunov stability verification [195–197], are being integrated with RL during training to enhance safety and stability.

Building on constrained and stability-aware RL, current progress shifts from soft-penalty safety handling toward explicit feasibility modeling and certifiable guarantees. FISOR [229] formulates safe offline RL through feasibility-guided diffusion, decoupling safety satisfaction from reward optimization and demonstrating strong zero-violation safety performance on DSRL-style benchmarks. CDT [230] further enables post-training adaptation to different safety budgets via zero-shot conditioning.

For formal safety assurance, verification of neural control barrier functions is becoming scalable through linear bound propagation [231], making certification more practical for larger neural controllers and higher-dimensional systems. In parallel, Hamilton-Jacobi reachability with formally verified neural value functions [232] provides a scalable route to safety filtering with explicit formal guarantees under nonlinear dynamics.

To improve safety under sim-to-real mismatch, SPiDR [233] introduces pessimistic domain randomization to explicitly account for dynamics uncertainty during training, yielding stronger transfer-time safety robustness. Complementarily, VLM-based frameworks such as AHA [234] enable semantic failure detection and reasoning from visual context, extending safety assessment beyond pure geometric collision criteria.

4.3.1 *Rigorous stability guarantees via Lyapunov analysis*

Building upon the integration of control theory, the field adopts rigorous stability criteria to provide theoretical certification. A policy π is theoretically certified as safe if it induces a Lyapunov function $L(s)$ that serves as a scalar measure of system energy and satisfies a negative drift condition defined as

$$\mathbb{E}[L(s_{t+1}) - L(s_t) \mid s_t, a_t] \leqslant -\alpha L(s_t). \tag{2}$$

Algorithms such as the Lyapunov-based actor-critic [235] enforce this drift as a hard constraint, which ensures that the state trajectory remains within a stable region of attraction almost surely. This approach contrasts sharply with Lagrangian relaxation methods, which convert constraints into soft penalties and frequently exhibit oscillatory violation patterns during the primal-dual optimization process. Theoretical analysis confirms that Lyapunov-based constraints provide a stability guarantee that is asymptotically distinct from and superior to heuristic penalty methods.

4.3.2 *Forward invariance through control barrier functions*

Complementing Lyapunov stability, control barrier functions ensure the forward invariance of the safe set $\mathcal{C}$ defined as $\{s \mid h(s) \geqslant 0\}$. Functioning as a differentiable safety filter, these functions project the proposed action u_{RL} onto a safe subspace defined by the inequality $\dot{h}(s,u) \geqslant -\gamma(h(s))$. Quantitative evaluations on high-dimensional control benchmarks demonstrate that this hybrid approach achieves near-zero violations during the entire training phase [236]. This property is critical for real-world deployment where the exploratory violations typical of standard reinforcement learning are physically unacceptable.

4.3.3 *Quantification of the safety-adaptability trade-off*

Despite significant advancements, a fundamental tension exists between strict safety constraints and the exploration required for robustness. This trade-off is quantitatively defined as the price of safety. Empirical evaluations on constrained benchmarks reveal a non-linear relationship where enforcing a strict constraint violation rate of less than 0.1% typically incurs a 15% to 25% reduction in asymptotic reward compared to unconstrained baselines [226]. This optimality gap quantitatively reflects the restriction of the explorable state space as the policy is forced to maintain a conservative buffer distance from hazardous boundaries. Consequently, future research focuses on balancing robustness and adaptability by utilizing techniques such as adaptive constraint handling to dynamically adjust safety boundaries based on real-time environmental uncertainty.

4.4 Real-time learning and adaptation

Embodied agents often need to adapt to dynamic environments in real time, which poses a significant challenge for RL methods. RL typically relies on extensive offline training, where agents learn optimal policies through large amounts of pre-collected data or simulations. By contrast, in real-world applications, environments can change rapidly and unpredictably. Agents must be capable of adjusting their policies on the fly to handle these changes, ensuring optimal performance under different scenarios.

For real-time learning and adaptation, continual learning approaches [237, 238] further enhance real-time adaptation by addressing the issue of catastrophic forgetting, where agents forget previously learned tasks when learning new ones. Continual learning ensures that agents can retain and build upon their existing knowledge while acquiring new skills. Additionally, incorporating human feedback or guidance [237, 239] into the learning process can significantly improve the efficiency and accuracy of real-time adaptation. Human-in-the-loop systems allow experts to provide corrections, demonstrations, or advice, helping agents to learn more effectively and avoid suboptimal behaviors. Moreover, meta-learning [92, 109, 119] enables agents to "learn how to learn", allowing them to quickly adapt to new tasks with minimal data.

Despite these methodological advances, achieving robust real-time adaptation remains difficult in practice. Agents must adapt under limited computational budgets, maintain stability under continual task shifts, and improve online performance despite sparse or delayed feedback. Moreover, real-world deployment introduces additional variability, such as fluctuating sampling rates and system latency, which further complicate adaptation dynamics. These intertwined challenges motivate recent efforts to make real-time adaptation both computationally feasible and algorithmically stable.

Recent studies operationalize real-time adaptation by improving deployment efficiency and adaptation stability. LoRA-based PEFT combined with low-bit quantization makes VLA policies feasible to run on consumer-grade GPUs, lowering the barrier for on-device finetuning [240]. For continual learning under non-stationarity, DMPEL incrementally grows a low-rank expert library with a lightweight router and uses coefficient replay to reduce catastrophic forgetting while enabling forward transfer [241]. To mitigate reward sparsity, RFTF trains a value model to provide temporally informed dense feedback aligned with human instructions, enabling fast adaptation on CALVIN "D" within a few interaction episodes [242]. Finally, TAWM conditions dynamics prediction on the variable time-step size Δt, improving robustness under irregular sampling rates [243]. Overall, these methods sug-

gest complementary ingredients for robust real-time adaptation: efficient on-device finetuning, continual knowledge retention, instruction-aligned dense feedback, and time-aware dynamics modeling.

4.5 Interpretability and transparency

Interpretability and transparency are critical in the development and deployment of embodied intelligence systems, especially in high-stakes applications such as robotics, healthcare, and autonomous vehicles. These systems must not only perform tasks effectively but also operate in a manner that is understandable, predictable, and aligned with human values and safety standards. When agents make decisions in complex, dynamic environments, understanding the "why" behind their actions becomes essential to ensure that these systems are trustworthy, safe, and capable of operating under uncertain conditions, and could also enhance learning efficiency.

General RL often struggles with these aspects due to its "black-box" nature. While RL excels in learning policies to optimize long-term rewards, it typically lacks clear mechanisms for explaining why specific actions are chosen. This opacity arises from its reliance on correlation-based learning, which does not inherently reveal causal relationships between actions, states, and outcomes. Consequently, the lack of interpretability in conventional RL methods limits their adoption in critical real-world scenarios where accountability and transparency are non-negotiable.

A promising solution is causal inference, which seeks to identify cause-and-effect relationships rather than mere correlations. By leveraging causal reasoning, we can better understand how actions influence outcomes and provide interpretable insights into an agent's decision-making process. Causal models enable the representation of environment dynamics in a structured and human-comprehensible way, which aligns well with the need for transparency in embodied intelligence.

The emerging field of **causal RL** builds upon this idea, integrating causal inference into reinforcement learning frameworks to address both interpretability and sample efficiency. For example, frameworks like ACE [50] incorporate causality-aware entropy regularization to enhance learning while maintaining transparency. Similarly, causal discovery techniques [244, 245] identify underlying causal structures that guide agent behavior, improving alignment with real-world dynamics.

Causal RL not only enables agents to explain their decisions but also facilitates counterfactual reasoning, allowing agents to predict the outcomes of hypothetical actions (e.g., "What would happen if I took action A instead of B?"). This capability, rooted in causal principles [246, 247], is critical for building robust and adaptable systems. Recent advancements in the field, as surveyed by Zeng et al. [248], highlight the potential of integrating causal inference to address challenges in high-dimensional, complex environments, making RL systems more interpretable and efficient.

By leveraging causal reasoning, embodied intelligence systems can achieve a deeper understanding of the environments they operate in, enabling safe, reliable, and transparent decision-making. This integration is not only a step toward enhancing interpretability but also a pathway to improving the overall robustness and applicability of RL in real-world scenarios.

5 Conclusion

This comprehensive survey has elucidated the pivotal role of RL in enhancing embodied intelligence, delineating a hierarchical framework that spans from single-task capabilities in restricted environments to the sophisticated cross-task and cross-environment adaptability essential for achieving AGI. The exploration of virtuoso capabilities marks a significant leap towards more sophisticated embodied AI systems. Morphological intelligence empowers robots to dynamically adjust their physical forms in response to varying task demands, while the incorporation of human-like values and ethics ensures that these agents operate in socially responsible and context-aware manners. Additionally, spontaneous reward generation and reflective equilibrium introduce higher-order cognitive abilities, enabling agents to autonomously generate reward signals and engage in ethical reasoning, respectively. These advancements are crucial for developing intelligent systems that not only perform tasks effectively but also align with human values and societal norms.

Despite the potential, the path to fully realizing embodied RL is fraught with challenges. Enhancing sample efficiency remains a critical concern, particularly for real-world applications where data collection is costly and risky. Addressing issues of generalization and transferability is essential for enabling agents to adapt to unforeseen contexts and diverse environments. Ensuring safety and robustness is paramount, especially in safety-critical domains, necessitating the integration of safety constraints and robust policy designs. Moreover, real-time learning and adaptation demand the development of online and continual learning algorithms that can dynamically adjust to

evolving environments. Finally, improving interpretability and transparency through causal inference and other techniques is vital for building trustworthy and accountable embodied AI systems.

Acknowledgements This work was supported by Nanjing Major Science and Technology Special Project (Grant No. 202405017), Joint Funds of the National Key Research and Development Program of China (Grant No. 2024YFB4711102), and National Natural Science Foundation of China (Grant No. U22A2057).

References

1 Wilson M. Six views of embodied cognition. Psychonomic Bull Rev, 2002, 9: 625–636
2 Wilson A D, Golonka S. Embodied cognition is not what you think it is. Front Psychol, 2013, 4: 58
3 Duan Y, Schulman J, Chen X, et al. RL^2: fast reinforcement learning via slow reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 5368–5376
4 Liu Y, Chen W, Bai Y, et al. Aligning cyber space with physical world: a comprehensive survey on embodied AI. 2024. ArXiv:2407.06886
5 Fiske A, Henningsen P, Buyx A. The Implications of Embodied Artificial Intelligence in Mental Healthcare for Digital Wellbeing. Philosophical Studies Series. Cham: Springer, 2020. 207–219
6 Zhou Y, Huang L, Bu Q, et al. Embodied understanding of driving scenarios. In: Proceedings of European Conference on Computer Vision, 2025. 129–148
7 Huang K, Guo D, Zhang X, et al. CompEtevo: towards morphological evolution from competition. 2024. ArXiv:2405.18300
8 Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search. Nature, 2016, 529: 484–489
9 Silver D, Hubert T, Schrittwieser J, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. 2017. ArXiv:1712.01815
10 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
11 Agarwal A, Kumar A, Malik J, et al. Legged locomotion in challenging terrains using egocentric vision. In: Proceedings of Conference on Robot Learning, 2023
12 Hwangbo J, Lee J, Dosovitskiy A, et al. Learning agile and dynamic motor skills for legged robots. Sci Robot, 2019, 4: eaau5872
13 Tan J, Zhang T, Coumans E, et al. Sim-to-real: learning agile locomotion for quadruped robots. 2018. ArXiv:1804.10332
14 Degrave J, Felici F, Buchli J, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature, 2022, 602: 414–419
15 Kaufmann E, Bauersfeld L, Loquercio A, et al. Champion-level drone racing using deep reinforcement learning. Nature, 2023, 620: 982–987
16 Bellemare M G, Candido S, Castro P S, et al. Autonomous navigation of stratospheric balloons using reinforcement learning. Nature, 2020, 588: 77–82
17 Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. 2023. ArXiv:2303.08774
18 Duan J, Yu S, Tan H L, et al. A survey of embodied AI: from simulators to research tasks. IEEE Trans Emerg Top Comput Intell, 2022, 6: 230–244
19 Ma Y, Song Z, Zhuang Y, et al. A survey on vision-language-action models for embodied AI. 2024. ArXiv:2405.14093
20 Gupta A, Savarese S, Ganguli S, et al. Embodied intelligence via learning and evolution. Nat Commun, 2021, 12: 5721
21 Deitke M, Han W, Herrasti A, et al. RoboTHOR: an open simulation-to-real embodied AI platform. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 3164–3174
22 Li H, Zuo Y, Yu J, et al. SimpleVLA-RL: scaling VLA training via reinforcement learning. 2025. ArXiv:2509.09674
23 Fei S, Wang S, Ji L, et al. SRPO: self-referential policy optimization for vision-language-action models. 2025. ArXiv:2511.15605
24 Zang H, Wei M, Xu S, et al. Rlinf-VLA: a unified and efficient framework for reinforcement learning of vision-language-action models. 2025. ArXiv:2510.06710
25 Kim M J, Finn C, Liang P. Fine-tuning vision-language-action models: optimizing speed and success. 2025. ArXiv:2502.19645
26 Chen Z, Niu R, Kong H, et al. TGRPO: fine-tuning vision-language-action model via trajectory-wise group relative policy optimization, 2025. ArXiv:2506.08440
27 Zhang J, Heo M, Liu Z X, et al. EXTRACT: efficient policy learning by extracting transferable robot skills from offline data. 2024. ArXiv:2406.17768
28 Pertsch K, Lee Y, Lim J J. Accelerating reinforcement learning with learned skill priors. 2020. ArXiv:2010.11944
29 Liu B, Zhu Y F, Gao C K, et al. Libero: benchmarking knowledge transfer for lifelong robot learning. 2023. ArXiv:2306.03310
30 Wang Y, Li X, Wang W, et al. Unified vision-language-action model. 2025. ArXiv:2506.19850
31 Reuss M, Zhou H, Rühle M, et al. FLOWER: democratizing generalist robot policies with efficient vision-language-action flow policies. 2025. ArXiv:2509.04996
32 Reuss M, Yanğmurlu Ö E, Wenzel F, et al. Multimodal diffusion transformer: learning versatile behavior from multimodal goals. 2024. ArXiv:2407.05996
33 Mees O, Hermann L, Burgard W. What matters in language conditioned robotic imitation learning over unstructured data. IEEE Robot Autom Lett, 2022, 7: 11205–11212
34 Nakamoto M, Zhai Y X, Singh A, et al. Cal-QL: calibrated offline RL pre-training for efficient online fine-tuning. 2023. ArXiv:2303.05479
35 Mees O, Hermann L, Rosete E, et al. Calvin: a benchmark for language-conditioned policy learning. 2021. ArXiv:2112.03227
36 Shridhar M, Manuelli L, Fox D, et al. Perceiver-actor: a multi-task transformer for robotic manipulation. In: Proceedings of CoRL, 2022
37 Gu J, Xiang F, Li X, et al. Maniskill2: a unified benchmark for generalizable manipulation skills. 2023. ArXiv:2302.04659

38 Goyal A, Xu J, Guo Y, et al. RVT: robotic view transformer for 3D object manipulation. 2023. ArXiv:2306.14896
39 Ze Y, Yan G, Wu Y H, et al. GNFactor: multi-task real robot learning with generalizable neural feature fields. 2024. ArXiv:2308.16891
40 Pathak D, Agrawal P, Efros A A, et al. Curiosity-driven exploration by self-supervised prediction. 2017. ArXiv:1705.05363
41 Ji T, Luo Y, Sun F, et al. Seizing serendipity: exploiting the value of past success in off-policy actor-critic. In: Proceedings of Forty-first International Conference on Machine Learning, 2024
42 Sun F, Huang W, Luo Y, et al. Robot cognitive learning by considering physical properties. Engineering, 2024, 47: 168–179
43 Niu H, Qiu Y, Li M, et al. When to trust your simulator: dynamics-aware hybrid offline-and-online reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 36599–36612
44 Niu H, Ji T, Liu B, et al. H2o+: an improved framework for hybrid offline-and-online RL with dynamics gaps. 2023. ArXiv:2309.12716
45 Yang Z, Li L, Lin K, et al. The dawn of LMMs: preliminary explorations with GPT-4V (ision). 2023. ArXiv:2309.17421
46 Todorov E, Erez T, Tassa Y. Mujoco: a physics engine for model-based control. In: Proceedings of International Conference on Intelligent Robots and Systems, 2012. 5026–5033
47 Brockman G, Cheung V, Pettersson L, et al. Openai gym. 2016. ArXiv:1606.01540
48 Ahn M, Zhu H, Hartikainen K, et al. Robel: robotics benchmarks for learning with low-cost robots. In: Proceedings of Conference on Robot Learning, 2020
49 Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of International Conference on Machine Learning, 2018
50 Ji T, Liang Y, Zeng Y, et al. Ace: off-policy actor-critic with causality-aware entropy regularization. In: Proceedings of International Conference on Machine Learning, 2024
51 Hansen N, Wang X, Su H. Temporal difference learning for model predictive control. In: Proceedings of International Conference on Machine Learning, 2022
52 Luo Y, Ji T, Sun F, et al. Offline-boosted actor-critic: adaptively blending optimal historical behaviors in deep off-policy RL. In: Proceedings of Forty-first International Conference on Machine Learning, 2024
53 Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: Proceedings of International Conference on Machine Learning, 2019
54 Kumar A, Fu J, Soh M, et al. Stabilizing off-policy q-learning via bootstrapping error reduction. In: Proceedings of Advances in Neural Information Processing Systems, 2019
55 Fujimoto S, Gu S S. A minimalist approach to offline reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 20132–20145
56 Nair A, Gupta A, Dalal M, et al. Awac: accelerating online reinforcement learning with offline datasets. 2020. ArXiv:2006.09359
57 Kumar A, Zhou A, Tucker G, et al. Conservative q-learning for offline reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020
58 Lyu J, Ma X, Li X, et al. Mildly conservative q-learning for offline reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2022
59 Kostrikov I, Fergus R, Tompson J, et al. Offline reinforcement learning with fisher divergence critic regularization. In: Proceedings of International Conference on Machine Learning, 2021. 5774–5783
60 Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit q-learning. In: Proceedings of International Conference on Learning Representations, 2021
61 Xu H, Jiang L, Li J, et al. Offline RL with no ood actions: in-sample learning via implicit value regularization. In: Proceedings of International Conference on Learning Representations, 2023
62 Song Y, Zhou Y, Sekhari A, et al. Hybrid RL: using both offline and online data can make RL efficient. In: Proceedings of the Eleventh International Conference on Learning Representations, 2022
63 Ball P J, Smith L, Kostrikov I, et al. Efficient online reinforcement learning with offline data. In: Proceedings of International Conference on Machine Learning, 2023. 1577–1594
64 Xie T, Jiang N, Wang H, et al. Policy finetuning: bridging sample-efficient offline and online reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 27395–27407
65 Nair A V, Pong V, Dalal M, et al. Visual reinforcement learning with imagined goals. In: Proceedings of Advances in Neural Information Processing Systems, 2018
66 Xu G, Zheng R, Liang Y, et al. Drm: mastering visual reinforcement learning through dormant ratio minimization. In: Proceedings of the Twelfth International Conference on Learning Representations, 2023
67 Rajeswaran A, Kumar V, Gupta A, et al. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. 2017. ArXiv:1709.10087
68 Tunyasuvunakool S, Muldal A, Doron Y, et al. dm_control: software and tasks for continuous control. Software Impacts, 2020, 6: 100022
69 Yarats D, Fergus R, Lazaric A, et al. Mastering visual continuous control: improved data-augmented reinforcement learning. 2021. ArXiv:2107.09645
70 Laskin M, Lee K, Stooke A, et al. Reinforcement learning with augmented data. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 19884–19895
71 Laskin M, Srinivas A, Abbeel P. Curl: contrastive unsupervised representations for reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2020. 5639–5650
72 Schwarzer M, Anand A, Goel R, et al. Data-efficient reinforcement learning with self-predictive representations. In: Proceedings of International Conference on Learning Representations, 2021
73 Zheng R, Wang X, Sun Y, et al. Taco: temporal latent action-driven contrastive loss for visual reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2024
74 Cetin E, Ball P J, Roberts S, et al. Stabilizing off-policy deep reinforcement learning from pixels. In: Proceedings of International Conference on Machine Learning, 2022. 2784–2810
75 Hansen N, Lin Y, Su H, et al. MoDem: accelerating visual model-based reinforcement learning with demonstrations. In: Proceedings of the Eleventh International Conference on Learning Representations, 2022

76 Zhang M, Vikram S, Smith L, et al. Solar: deep structured representations for model-based reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2019. 7444–7453
77 Seo Y, Hafner D, Liu H, et al. Masked world models for visual control. In: Proceedings of Conference on Robot Learning, 2023. 1332–1344
78 Hafner D, Lillicrap T, Ba J, et al. Dream to control: learning behaviors by latent imagination. In: Proceedings of International Conference on Learning Representations, 2020
79 Sandha S S, Garcia L, Balaji B, et al. Sim2real transfer for deep reinforcement learning with stochastic state transition delays. In: Proceedings of Conference on Robot Learning, 2021. 1066–1083
80 Zhao W, Queralta J P, Westerlund T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In: Proceedings of 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020. 737–744
81 Peng X B, Andrychowicz M, Zaremba W, et al. Sim-to-real transfer of robotic control with dynamics randomization. In: Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018. 3803–3810
82 Andrychowicz O A M, Baker B, Chociej M, et al. Learning dexterous in-hand manipulation. Int J Robotics Res, 2020, 39: 3–20
83 Eysenbach B, Chaudhari S, Asawa S, et al. Off-dynamics reinforcement learning: training for transfer with domain classifiers. In: Proceedings of International Conference on Learning Representations, 2021
84 Liu J, Zhang H, Wang D. Dara: dynamics-aware reward augmentation in offline reinforcement learning. 2022. ArXiv:2203.06662
85 Wang J X, Kurth-Nelson Z, Tirumala D, et al. Learning to reinforcement learn. 2016. ArXiv:1611.05763
86 Rusu A A, Colmenarejo S G, Gulcehre C, et al. Policy distillation. 2015. ArXiv:1511.06295
87 Mutti M, Mancassola M, Restelli M. Unsupervised reinforcement learning in multiple environments. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022. 7850–7858
88 Parisi S, Dean V, Pathak D, et al. Interesting object, curious agent: learning task-agnostic exploration. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 20516–20530
89 Rabault J, Kuhnle A. Accelerating deep reinforcement learning strategies of flow control through a multi-environment approach. Phys Fluids, 2019, 31: 094105
90 Nafi N M, Poggi-Corradini G, Hsu W. Policy optimization with augmented value targets for generalization in reinforcement learning. In: Proceedings of 2023 International Joint Conference on Neural Networks (IJCNN), 2023. 1–8
91 Maldini P, Mutti M, De S R, et al. Non-markovian policies for unsupervised reinforcement learning. In: Proceedings of First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML, 2022
92 Yu T, Quillen D, He Z, et al. Meta-world: a benchmark and evaluation for multi-task and meta reinforcement learning. In: Proceedings of Conference on Robot Learning, 2020. 1094–1100
93 Tobin J, Fong R, Ray A, et al. Domain randomization for transferring deep neural networks from simulation to the real world. In: Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017. 23–30
94 Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks. J Mach Learn Res, 2016, 17: 1–35
95 Higgins I, Pal A, Rusu A, et al. Darla: improving zero-shot transfer in reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2017. 1480–1490
96 James S, Wohlhart P, Kalakrishnan M, et al. Sim-to-real via sim-to-sim: data-efficient robotic grasping via randomized-to-canonical adaptation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 12627–12637
97 Rusu A A, Rabinowitz N C, Desjardins G, et al. Progressive neural networks. 2016. ArXiv:1606.04671
98 Parisi G I, Kemker R, Part J L, et al. Continual lifelong learning with neural networks: a review. Neural Netws, 2019, 113: 54–71
99 Barreto A, Dabney W, Munos R, et al. Successor features for transfer in reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 30
100 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of International Conference on Machine Learning, 2017. 1126–1135
101 Taylor M E, Stone P. Transfer learning for reinforcement learning domains: a survey. J Mach Learn Res, 2009, 10: 1633–1685
102 Zhang Y, Zohren S, Roberts S. A deep reinforcement learning framework for financial portfolio management. In: Proceedings of Deep Reinforcement Learning Workshop, NeurIPS, 2018
103 Zhang A, McAllister R T, Calandra R, et al. Learning invariant representations for reinforcement learning without reconstruction. In: Proceedings of International Conference on Learning Representations, 2017
104 Sharma A, Gupta A, Levine S. Dynamics-aware unsupervised discovery of skills. In: Proceedings of International Conference on Learning Representations (ICLR), 2019
105 Lu K, Grover A, Abbeel P, et al. Frozen pretrained transformers as universal computation engines. In: Proceedings of AAAI Conference on Artificial Intelligence, 2022, 36: 7628–7636
106 Zhang Z, Zhao R, Wang T. Bridging feature and policy spaces for transfer in reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2022
107 Petangoda J C, Pascual-Diaz S, Adam V, et al. Disentangled skill embeddings for reinforcement learning. 2019. ArXiv:1906.09223
108 Teh Y W, Bapst V, Czarnecki W, et al. Distral: robust multitask reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2017
109 Hospedales T M, Antoniou A, Micaelli P, et al. Meta-learning in neural networks: a survey. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 5149–5169
110 Ferns N, Panangaden P, Precup D. Value-function-based transfer for reinforcement learning using structure mapping. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI), 2006. 720–725
111 Liu Y, Hu Y, Gao Y, et al. Value function transfer for deep multi-agent reinforcement learning based on n-step returns. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 2019. 465–471
112 Schaul T, Horgan D, Gregor K, et al. Universal value function approximators. In: Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015. 1312–1320
113 Li S, Barto A, Yu S. Generalized reward shaping for reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020
114 Devin C, Gupta A, Darrell T, et al. Learning modular neural network policies for multi-task and multi-robot transfer. In: Proceedings

of International Conference on Robotics and Automation (ICRA), 2017
115 Rajeswaran A, Mordatch I, Kumar V. Meta-reinforcement learning for reward function transfer. In: Proceedings of Advances in Neural Information Processing Systems, 2020
116 Rakelly K, Zhou A, Quillen D, et al. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In: Proceedings of the 36th International Conference on Machine Learning (ICML), 2019. 5331–5340
117 Zhang H, Zheng B, Guo A, et al. Scrutinize what we ignore: reining task representation shift in context-based offline meta reinforcement learning. 2024. ArXiv:2405.12001
118 Zintgraf L, Shiarli K, Kurin V, et al. Fast context adaptation via meta-learning. In: Proceedings of the 36th International Conference on Machine Learning (ICML), 2019. 7693–7702
119 Gupta A, Mendonca R, Liu Y, et al. Meta-reinforcement learning of structured exploration strategies. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 5302–5311
120 Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci USA, 2017, 114: 3521–3526
121 Rolnick D, Ahuja A, Schwarz J, et al. Experience replay for continual learning. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 348–358
122 Zenke F, Poole B, Ganguli S. Continual learning through synaptic intelligence. In: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, 3987–3995
123 Shin H, Lee J K, Kim J, et al. Continual learning with deep generative replay. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 2994–3003
124 Fernando C, Banarse D, Blundell C, et al. PathNet: evolution channels gradient descent in super neural networks. 2017. ArXiv:1701.08734
125 Wei Z, Chen H, Nan L, et al. PathNet: path-selective point cloud denoising. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 4426–4442
126 Tessler C, Givony S, Zahavy T, et al. A deep hierarchical approach to lifelong learning in minecraft. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2017. 31
127 Khetarpal K, Riemer M, Rish I, et al. Towards continual reinforcement learning: a review and perspectives. J Artif Intell Res, 2022, 75: 1401–1476
128 Barto A G, Mahadevan S. Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Syst, 2003, 13: 341–379
129 Dayan P, Hinton G E. Feudal reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 1993. 5: 271–278
130 Luo Y, Ji T, Sun F, et al. Goal-conditioned hierarchical reinforcement learning with high-level model approximation. IEEE Trans Neural Netw Learn Syst, 2025, 36: 2705–2719
131 Luo Y, Sun F, Ji T, et al. Bidirectional-reachable hierarchical reinforcement learning with mutually responsive policies. In: Proceedings of Reinforcement Learning Conference, 2024
132 Sutton R S, Precup D, Singh S. Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. In: Proceedings of Artificial Intelligence, 1999. 181–211
133 Vezhnevets A S, Osindero S, Schaul T, et al. Feudal networks for hierarchical reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017
134 Bacon P L, Harb J, Precup D. The option-critic architecture. In: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017. 2207–2216
135 Tiwari S, Thomas P S. Natural option critic. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 5175–5182
136 Kulkarni T D, Narasimhan K, Saeedi A, et al. Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 3675–3683
137 Eysenbach B, Gupta A, Ibarz J, et al. Diversity is all you need: learning skills without a reward function. 2018. ArXiv:1802.06070
138 Florensa C, Duan Y, Abbeel P. Stochastic neural networks for hierarchical reinforcement learning. In: Proceedings of International Conference on Learning Representations (ICLR), 2017
139 Nachum O, Gu S, Lee H, et al. Data-efficient hierarchical reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 3303–3313
140 Vamplew P, Dazeley R, Berry A, et al. Empirical evaluation methods for multiobjective reinforcement learning algorithms. Mach Learn, 2011, 84: 51–80
141 Roijers D M, Vamplew P, Whiteson S, et al. A survey of multi-objective sequential decision-making. J Artif Intell Res, 2013, 48: 67–113
142 Van M K, Nowé A. Multi-objective deep reinforcement learning. In: Proceedings of International Conference on Neural Information Processing, 2014. 144–151
143 Qiu S, Zhang D, Yang R, et al. Traversing pareto optimal policies: provably efficient multi-objective reinforcement learning. 2024. ArXiv:2407.17466
144 Chen Z, Zhou F, Trimponias G, et al. Multi-objective neural architecture search via non-stationary policy gradient. 2020. ArXiv:2001.08437
145 Parisi S, Pirotta M, Smacchia N, et al. Policy gradient approaches for multi-objective sequential decision making. In: Proceedings of 2014 International Joint Conference on Neural Networks (IJCNN), 2014. 2323–2330
146 Kanazawa T, Gupta C. Latent-conditioned policy gradient for multi-objective deep reinforcement learning. 2023. ArXiv:2303.08909
147 Zhang X, Ma Y, Singla A. Task-agnostic exploration in reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020
148 Oudeyer P, Finn C, Bramley N. Task-agnostic reinforcement learning workshop. In: Proceedings of ICLR 2019 Workshops: LLD, 2019
149 Caccia M, Mueller J, Kim T, et al. Task-agnostic continual reinforcement learning: gaining insights and overcoming challenges. 2022. ArXiv:2205.14495
150 Team P I. π_0: a flow-matching foundation model for physical intelligence. 2024. ArXiv:2410.24164
151 Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: an open-source vision-language-action model. 2024. ArXiv:2406.09246
152 Chen Y, Tian S, Liu S, et al. Conrft: a reinforced fine-tuning method for VLA models via consistency policy. 2025. ArXiv:2502.05450
153 Lei K, Li H Y, Yu D J, et al. RL-100: performant robotic manipulation with real-world reinforcement learning. 2025. ArXiv:2510.14830

154 Zhai S, Zhang Q, Zhang T Y, et al. A vision-language-action-critic model for robotic real-world reinforcement learning. 2025. ArXiv:2509.15937
155 Chen K, Liu Z H, Zhang T H, et al. π_{RL}: online RL fine-tuning for flow-based vision-language-action models. 2025. ArXiv:2510.25889
156 Lv L, Li Y, Luo Y, et al. Flow-based policy for online reinforcement learning. 2025. ArXiv:2506.12811
157 Bruce J, Dennis M, Edwards A, et al. Genie: generative interactive environments. In: Proceedings of International Conference on Machine Learning (ICML), 2024
158 Luo J, Hu Z Y, Xu C, et al. SERL: a software suite for efficient robotic reinforcement learning. 2024. ArXiv:2401.16013
159 Bhatt A, Palenicek D, Belousov B, et al. CrossQ: batch normalization in deep reinforcement learning for greater sample efficiency. 2024. ArXiv:1902.05605
160 Sharma A, Ahmed A M, Ahmad R, et al. Self-improving robots: end-to-end autonomous visuomotor reinforcement learning. In: Proceedings of the 7th Conference on Robot Learning, 2023. 3292–3308
161 Mark M S, Gao T, Sampaio G G, et al. Policy agnostic RL: offline RL and online RL fine-tuning of any class and backbone. 2024. ArXiv:2412.06685
162 Yang J, Mark M S, Vu B, et al. Robot fine-tuning made easy: pre-training rewards and policies for autonomous real-world reinforcement learning. In: 2024 IEEE International Conference on Robotics and Automation, 2024. 4804–4811
163 Liu H, Guo D, Sun F, et al. Morphology-based embodied intelligence: historical retrospect and research progress. Acta Autom Sin, 2023, 49: 1131–1154
164 Chen R, Han J, Sun F, et al. Subequivariant graph reinforcement learning in 3D environments. In: Proceedings of International Conference on Machine Learning, 2023. 4545–4565
165 Chen R, Wang L, Du Y, et al. Subequivariant reinforcement learning in 3D multi-entity physical environments. 2024. ArXiv:2407.12505
166 Pathak D, Lu C, Darrell T, et al. Learning to control self-assembling morphologies: a study of generalization via modularity. In: Proceedings of Advances in Neural Information Processing Systems, 2019
167 Furuta H, Iwasawa Y, Matsuo Y, et al. A system for morphology-task generalization via unified representation and behavior distillation. In: Proceedings of the Eleventh International Conference on Learning Representations, 2022
168 Christiano P F, Leike J, Brown T B, et al. Deep reinforcement learning from human preferences. In: Proceedings of Advances in Neural Information Processing Systems, 2017
169 Wirth C, Akrour R, Neumann G, et al. A survey of preference-based reinforcement learning methods. J Mach Learn Res, 2017, 18: 1–46
170 Ziebart B D, Maas A, Bagnell J, et al. Maximum entropy inverse reinforcement learning. In: Proceedings of the 23rd National Conference on Artificial Intelligence, 2008. 1433–1438
171 Li J, Zheng J, Zheng Y, et al. DecisionNCE: embodied multimodal representations via implicit preference learning. In: Proceedings of Forty-first International Conference on Machine Learning, 2024
172 Ng A Y, Russell S. Algorithms for inverse reinforcement learning. In: Proceedings of ICML, 2000
173 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 2018
174 Van Moffaert K, Nowé A. Multi-objective reinforcement learning using sets of pareto dominating policies. J Mach Learn Res, 2014, 15: 3483–3512
175 Ng A Y, Harada D, Russell S J. Policy invariance under reward transformations: theory and application to reward shaping. In: Proceedings of the Sixteenth International Conference on Machine Learning, 1999. 278–287
176 Amodei D, Olah C, Steinhardt J, et al. Concrete problems in ai safety. 2016. ArXiv:1606.06565
177 Swamy G, Wu D, Choudhury S, et al. Inverse reinforcement learning without reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2023. 33299–33318
178 Ma Y J, Liang W, Wang G, et al. EUREKA: human-level reward design via coding large language models. 2023. ArXiv:2310.12931
179 Eysenbach B, Zhang T, Levine S, et al. Contrastive learning as goal-conditioned reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 35603–35620
180 Mazoure B, Eysenbach B, Nachum O, et al. Contrastive value learning: implicit models for simple offline RL. In: Proceedings of Conference on Robot Learning, 2023. 1257–1267
181 Schmidhuber J. Curious model-building control systems. In: Proceedings of IEEE International Joint Conference on Neural Networks, 1991. 1458–1463
182 Sutrop M. Challenges of aligning artificial intelligence with human values. Acta Balt, 2020, 8: 54–72
183 Han S, Kelly E, Nikou S, et al. Aligning artificial intelligence with human values: reflections from a phenomenological perspective. AI&Soc, 2022, 37: 1–13
184 Gabriel I. Artificial intelligence, values, and alignment. Minds Machines, 2020, 30: 411–437
185 Cath Y. The Oxford Handbook of Philosophical Methodology. Oxford: Oxford University Press, 2015
186 Paulo N. The unreliable intuitions objection against reflective equilibrium. J Ethics, 2020, 24: 333–353
187 Jonker C M, Siebert L C, Murukannaiah P K. Reflective hybrid intelligence for meaningful human control in decision-support systems. In: Proceedings of Research Handbook on Meaningful Human Control of Artificial Intelligence Systems, 2024. 188–204
188 Tarleton N. Coherent extrapolated volition: a meta-level approach to machine ethics. MIRI Working Paper, 2010. https://intelligence.org/files/CEV-MachineEthics.pdf
189 Robinson P. Moral disagreement and artificial intelligence. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021. 209–209
190 Stumborg M F, Roh B, Rosen M, et al. Dimensions of Autonomous Decision-making. DRM-2021-U-030642-1Rev. 2021
191 Lim H S M, Taeihagh A. Algorithmic decision-making in AVs: understanding ethical and technical concerns for smart cities. Sustainability-Basel, 2019, 11: 5791
192 Mei A, Zhu G N, Zhang H, et al. ReplanVLM: replanning robotic tasks with visual language models. IEEE Robot Autom Lett, 2024, 9: 10201–10208
193 Skreta M, Zhou Z, Yuan J L, et al. Replan: robotic replanning with perception and language models. 2024. ArXiv:2401.04157
194 Zhou H, Ji T, Sommerhalder L, et al. RoboGolf: mastering real-world minigolf with a reflective multi-modality vision-language model. 2024. ArXiv:2406.10157

195 Chow Y, Nachum O, Duenez-Guzman E, et al. A lyapunov-based approach to safe reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 31

196 Chow Y, Nachum O, Faust A, et al. Lyapunov-based safe policy optimization for continuous control. 2019. ArXiv:1901.10031

197 Dawson C, Gao S, Fan C. Safe control with learned certificates: a survey of neural lyapunov, barrier, and contraction methods for robotics and control. IEEE Trans Robot, 2023, 39: 1749–1767

198 Wang T, Bao X, Clavera I, et al. Benchmarking model-based reinforcement learning. 2019. ArXiv:1907.02057

199 Janner M, Fu J, Zhang M, et al. When to trust your model: model-based policy optimization. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 32

200 Ji T, Luo Y, Sun F, et al. When to update your model: constrained model-based reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 23150–23163

201 Luo Y, Xu H, Li Y, et al. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In: Proceedings of International Conference on Learning Representations, 2018

202 Zhang H, Yu H, Zhao J, et al. How to fine-tune the model: unified model shift and model bias policy optimization. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 36

203 Levine S, Abbeel P. Learning neural network policies with guided policy search under unknown dynamics. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 27

204 Tassa Y, Erez T, Todorov E. Synthesis and stabilization of complex behaviors through online trajectory optimization. In: Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012. 4906–4913

205 Deisenroth M, Rasmussen C E. PILCO: a model-based and data-efficient approach to policy search. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011. 465–472

206 Rao A V. A survey of numerical methods for optimal control. In: Proceedings of Advances in the Astronautical Sciences, 2009. 135: 497–528

207 Chua K, Calandra R, McAllister R, et al. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 31

208 Yang M, Du Y, Ghasemipour K, et al. Learning interactive real-world simulators. 2023. ArXiv:2310.06114

209 Zhu F, Wu H, Guo S, et al. Irasim: learning interactive real-robot action simulators. 2024. ArXiv:2406.14540

210 Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay. 2015. ArXiv:1511.05952

211 Zha D, Lai K H, Zhou K, et al. Experience replay optimization. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019. 4243–4249

212 Sun P, Zhou W, Li H. Attentive experience replay. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 5900–5907

213 Yin H, Pan S. Knowledge transfer for deep reinforcement learning with hierarchical experience replay. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2017

214 Andrychowicz M, Wolski F, Ray A, et al. Hindsight experience replay. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 30

215 D'Oro P, Schwarzer M, Nikishin E, et al. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In: Proceedings of the Eleventh International Conference on Learning Representations, 2023

216 Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods. In: Proceedings of International Conference on Machine Learning, 2018. 1587–1596

217 Hasselt H. Double Q-learning. In: Proceedings of Advances in Neural Information Processing Systems, 2010. 23

218 Nachum O, Dai B, Kostrikov I, et al. AlgaeDice: policy gradient from arbitrary experience. 2019. ArXiv:1912.02074

219 Ho J, Ermon S. Generative adversarial imitation learning. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 29

220 Ross S, Gordon G, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011. 627–635

221 Hansen N, Lin Y, Su H, et al. MoDem: accelerating visual model-based reinforcement learning with demonstrations. In: Proceedings of the Eleventh International Conference on Learning Representations, 2023

222 Qu G, Wierman A, Li N. Scalable reinforcement learning for multiagent networked systems. Operations Res, 2022, 70: 3601–3628

223 Hansen N, Su H, Wang X. TD-MPC2: scalable, robust world models for continuous control. In: Proceedings of the Twelfth International Conference on Learning Representations, 2024

224 Ma C, Li A, Du Y, et al. Efficient and scalable reinforcement learning for large-scale network control. Nat Mach Intell, 2024, 6: 1006–1020

225 Kordabad A B, Wisniewski R, Gros S. Safe reinforcement learning using Wasserstein distributionally robust MPC and chance constraint. IEEE Access, 2022, 10: 130058–130067

226 Zhao W, Sun Y, Li F, et al. Guard: a safe reinforcement learning benchmark. 2024. ArXiv:2305.13681

227 Kwon R, Kwon G. Safety constraint-guided reinforcement learning with linear temporal logic. Systems, 2023, 11: 535

228 Junges S, Jansen N, Dehnert C, et al. Safety-constrained reinforcement learning for mdps. In: Proceedings of International Conference on Tools and Algorithms for the Construction and Analysis of Systems, 2016. 130–146

229 Zheng Y, Li J, Yu D, et al. Safe offline reinforcement learning with feasibility-guided diffusion model. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR), 2024

230 Liu Z, Mao Z, Liu Z, et al. Constrained decision transformer for offline safe reinforcement learning. In: Proceedings of International Conference on Machine Learning (ICML), 2023. 21611–21631

231 Vertovec N, Mathiesen F B, Badings T, et al. Scalable verification of neural control barrier functions using linear bound propagation. 2025. ArXiv:2511.06341

232 Yang Y, Hu H, Wei T, et al. Scalable synthesis of formally verified neural value function for Hamilton-Jacobi reachability analysis. J Artif Intell Res, 2025, 83: 19

233 As Y, Qu C, Unger B, et al. SPiDR: a simple approach for zero-shot safety in sim-to-real transfer. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2025

234 Duan J, Pumacay W, Kumar N, et al. AHA: a vision-language-model for detecting and reasoning over failures in robotic manipulation.

2024. ArXiv:2410.00371
235 Han M, Zhang L, Wang J, et al. Actor-critic reinforcement learning for control with stability guarantee. IEEE Robot Autom Lett, 2020, 5: 6217–6224
236 Cheng R, Orosz G, Murray R M, et al. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In: Proceedings of AAAI, 2019
237 Zhang H, Lei Y, Gui L, et al. CPPO: continual learning for reinforcement learning with human feedback. In: Proceedings of the Twelfth International Conference on Learning Representations, 2024
238 Dohare S, Hernandez-Garcia J F, Lan Q, et al. Loss of plasticity in deep continual learning. Nature, 2024, 632: 768–774
239 Luo J, Xu C, Wu J, et al. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. 2024. ArXiv:2410.21845
240 Omaisan A Y A, Mohamed I S. Towards accessible physical AI: lora-based fine-tuning of VLA models for real-world robot control. 2025. ArXiv:2512.11921
241 Lei Y, Mao S, Zhou S, et al. Dynamic mixture of progressive parameter-efficient expert library for lifelong robot learning. 2025. ArXiv:2506.05985
242 Shu J, Lin Z, Wang Y. RFTF: reinforcement fine-tuning for embodied agents with temporal feedback. 2025. ArXiv:2505.19767
243 Nhu A N, Son S, Lin M. Time-aware world model for adaptive prediction and control. In: Proceedings of the 42nd International Conference on Machine Learning, 2025
244 Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet, 2019, 10: 524
245 Zeng Y, Shimizu S, Cai R, et al. Causal discovery with multi-domain LINGAM for latent factors. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021
246 Jonas P, Dominik J, Bernhard S. Elements of Causal Inference. Cambridge: MIT Press, 2017
247 Imbens G W, Rubin D B. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge: Cambridge University Press, 2015
248 Zeng Y, Cai R, Sun F, et al. A survey on causal reinforcement learning. IEEE Trans Neural Netw Learn Syst, 2025, 36: 5942–5962