

Adaptability in Multi-Agent Reinforcement Learning: A Framework and Unified Review

Siyi Hu^{*1}, Mohamad A. Hady¹, Jianglin Qiao¹, Jimmy Cao¹, Mahardhika Pratama¹, and Ryszard Kowalczyk¹

¹Adelaide University, South Australia, Australia
 {siyi.hu, mohamad.hady, jianglin.qiao, jimmy.cao, dhika.pratama, ryszard.kowalczyk}@unisa.edu.au

Preprint Version

Abstract

Multi-Agent Reinforcement Learning (MARL) has shown clear effectiveness in coordinating multiple agents across simulated benchmarks and constrained scenarios. However, its deployment in real-world multi-agent systems (MAS) remains limited, primarily due to the complex and dynamic nature of such environments. These challenges arise from multiple interacting sources of variability, including fluctuating agent populations, evolving task goals, and inconsistent execution conditions. Together, these factors demand that MARL algorithms remain effective under continuously changing system configurations and operational demands. To better capture and assess this capacity for adjustment, we introduce the concept of *adaptability* as a unified and practically grounded lens through which to evaluate the reliability of MARL algorithms under shifting conditions, broadly referring to any changes in the environment dynamics that may occur during learning or execution. Centred on the notion of adaptability, we propose a structured framework comprising three key dimensions: learning adaptability, policy adaptability, and scenario-driven adaptability. By adopting this adaptability perspective, we aim to support more principled assessments of MARL performance beyond narrowly defined benchmarks. Ultimately, this survey contributes to the development of algorithms that are better suited for deployment in dynamic, real-world multi-agent systems.

1 Introduction

Multi-Agent Reinforcement Learning (MARL) extends reinforcement learning (RL) to settings involving multiple learning agents. It has become a key framework for addressing sequential decision-making problems in real-world scenarios [1, 2, 3, 4, 5, 6, 7, 8]. Compared to single-agent RL [9, 10, 11], the multi-agent setting introduces additional complexities, such as agent interaction structures [12, 13], partial observability [14, 15], and heterogeneous roles [16, 17], among others.

To address these complexities, MARL algorithms are often developed under specific structural and operational assumptions. Prominent examples include cooperative MARL under centralised

^{*}Corresponding author: siyi.hu@unisa.edu.au

training with decentralised execution (CTDE) [14, 18, 19], mean-field methods for large agent populations [13, 20], offline MARL based on static datasets [21, 22, 23], networked MARL with local communication [24, 25, 26], model-based methods that incorporate predictive models of the multi-agent systems [27, 28, 29], and constrained MARL frameworks for safe coordination [30, 31, 32]. While these paradigms demonstrate strong performance within their respective domains, they are often evaluated under narrowly defined or fixed conditions that align with their specific assumptions. As a result, it is difficult to assess generality or cross-paradigm robustness of a MARL algorithm.

In this work, we argue that *algorithmic-level assumptions must be treated as first-class considerations by evaluating the applicability of a MARL algorithm*. A strong learning paradigm should generalise across a range of environments and interaction patterns with minimal adjustment. This requirement is motivated by the fact that real-world multi-agent systems are inherently dynamic. Agent populations fluctuate, objectives evolve, and execution-time constraints vary across deployments. Algorithms designed around fixed training conditions frequently experience degraded performance—or even outright failure—when exposed to deployment-time variation. It is therefore essential to assess whether a method can maintain its effectiveness under such changes, rather than evaluating it solely within the confines of its original design assumptions. We highlight this issue through three representative cases:

1. *CTDE: Coordination Without Flexibility?* CTDE methods enable a group of agents to coordinate by optimizing centralized critics [33, 18, 34], joint value functions [14, 35, 36], and trust-region updates [37, 38, 19]. However, they are typically restricted to fixed agent sets and full observability during training. As a result, these approaches often struggle to adapt when population structures or observability patterns change at deployment [39, 40, 41].
2. *Mean-field or Networked MARL: Scalable But Still Centralized?* Mean-field approximations [42, 43, 13] or networked local communication [24, 25, 26] enable scalable learning by reducing agent interactions to statistical aggregates, but they all assume access to the full population during training. This assumption limits their effectiveness when centralized learning and synchronous policy update is infeasible [44, 45, 46].
3. *Offline MARL: Trained Once, Deployed Anywhere?* Offline MARL offers a compelling alternative by training policies on fixed datasets [47, 22, 23]. Yet, these methods often fail to generalize in environments with unseen agent combinations or evolving dynamics. Without updated trajectories, offline policies become susceptible to extrapolation errors [48, 49, 47].

Existing literature has attempted to characterize desirable properties that can mitigate the issues above. Terms such as *scalability*, *robustness*, *generalization*, and *transferability* appear frequently across papers and surveys [8, 3, 50, 7, 2]. However, these notions are often defined inconsistently and emphasize different, sometimes orthogonal, aspects of algorithm behaviour. Crucially, there is no unified conceptual framework for understanding how these properties relate to the core challenge of real-world deployment.

For example, *scalability* typically focuses on performance as the number of agents increases, but often neglects behavioural stability under changing agent relationships or asynchronous execution [7]. *Transferability*, on the other hand, emphasizes reusing learned knowledge across different tasks or environments, but tends to assume consistent agent configurations or fixed coordination protocols [2]. Similarly, *robustness* often refers to resilience under noise or perturbations, without accounting for structural shifts in agent interactions or task goals [51]. While each of these perspectives captures a critical facet of the problem, they remain fragmented and fail to provide a comprehensive view of the challenges posed by dynamic multi-agent systems. Figure 1 illustrates the conceptual overlap among these properties and highlights their intersection.

To address these limitations, we introduce *adaptability* as a unifying perspective for analysing

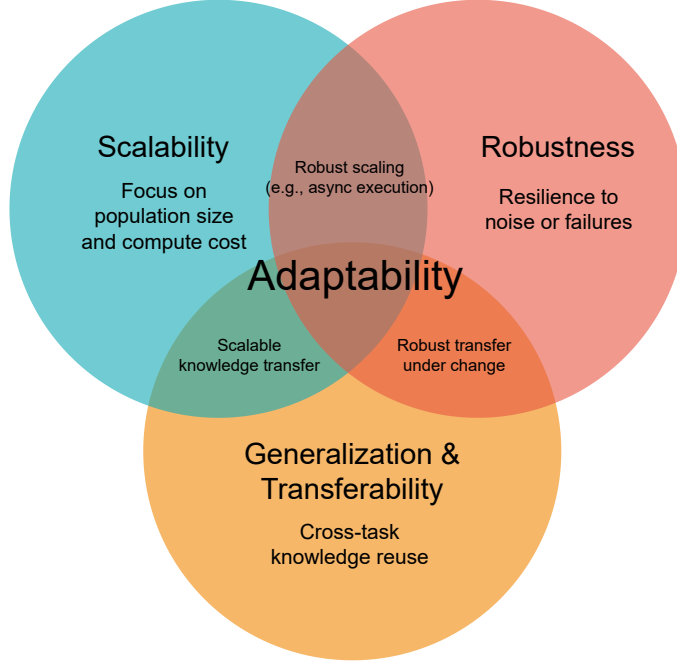


Figure 1: Conceptual overlap between scalability, robustness, and generalization/transferability in MARL. While each concept captures a critical facet of algorithm behaviour, only their intersection reflects the broader demands of adaptability in both MARL training and deployment.

and organizing MARL algorithms under dynamic and uncertain conditions. Unlike prior concepts that isolate specific dimensions of performance, adaptability captures the overarching requirement for agents to remain effective amid changes in system dynamics, task structure, or partner behaviour. We decompose this concept into three interrelated dimensions, each reflecting a distinct facet of the challenge:

- *Learning Adaptability* focuses on how robustly a learning paradigm performs under variations in agent populations, execution assumptions, and coordination structures, without changing the core learning mechanism.
- *Policy Adaptability* targets whether a policy trained on a subset of tasks or agent configurations can generalize to novel tasks, unseen objectives, or unfamiliar partner agents.
- *Scenario-Driven Adaptability* emphasizes whether environments used for training and evaluation support diverse and representative challenges that reflect the demands of real-world deployments.

Together, these perspectives offer a unified lens for evaluating whether a MARL algorithm or environment is capable of moving beyond isolated algorithmic advances toward broader usage.

The remainder of this survey is organized as follows. Section 2 reviews relevant literature and situates our contribution within existing surveys. Section 3 introduces the concept of *learning adaptability*, examining how different training paradigms respond to structural variation. Section 4 explores *policy adaptability*, analysing how policies generalize across tasks, roles, and agent populations. Section 5 presents *scenario-driven adaptability*, discussing how benchmark environments support or hinder algorithmic robustness. Finally, Section 6 outlines open research challenges and future directions toward more adaptive, transferable, and deployment-ready MARL systems, and Section 7 concludes with a summary of key insights.

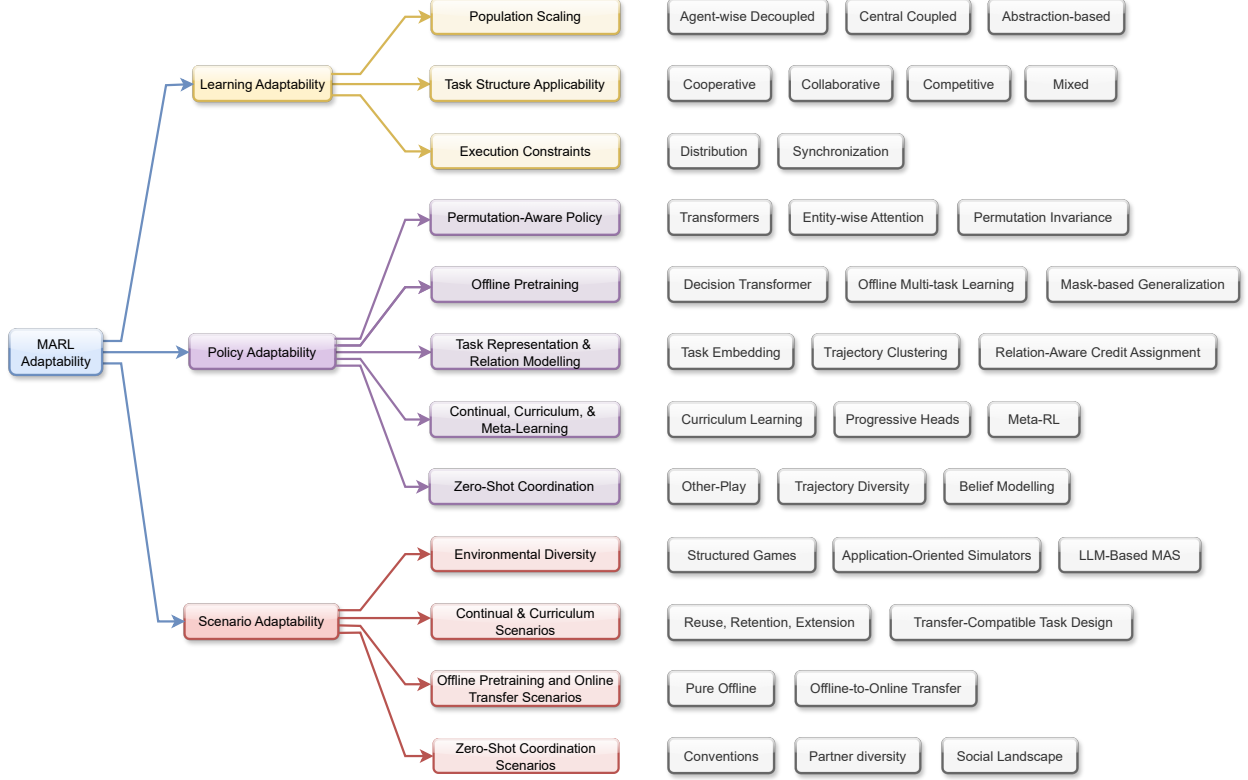


Figure 2: A structured overview of MARL adaptability across three dimensions with key topics.

2 Background and Motivation

2.1 A Primer on MARL Paradigms

A number of foundational paradigms structure the design of MARL algorithms. *Centralized Training with Decentralized Execution (CTDE)* is the dominant paradigm in cooperative MARL. It assumes access to global information during training, often via centralized critics [33, 18, 34] or joint value functions [14, 35, 36], while preserving decentralized execution at test time. *Independent Learning (IL)* removes inter-agent coupling by training each agent in isolation based only on local observations [12, 52]. Despite its scalability, IL often suffers from non-stationarity and poor coordination in cooperative tasks [53]. *Centralized Critics (CC)* such as MADDPG [33] or MAPPO [18] stabilize training by conditioning policy gradients on joint observations and actions, enabling global coordination across agents [54, 34]. *Value Decomposition (VD)* methods, including VDN [35], QTRAN [55], QMIX [14], WQMIX [56], and QPLEX [36], decompose a global value function into per-agent utilities using constraints such as monotonicity to retain coordination signal while training individual agents. *Heterogeneous-Agent (HA)* methods like HATRPO [19], HAPPO [38], and HASAC [37], stabilize training by updating agents sequentially and adjusting for changes in co-agent policies. This trust-region-based approach enables effective gradient updates even when agents have distinct roles and observation spaces. *Offline MARL* methods [21, 22, 23] learn policies from pre-collected datasets without active exploration. They promise fast deployment but face extrapolation errors in unseen scenarios. *Mean-field MARL* [13, 20, 57] approximates the influence of many agents by aggregating their behaviours into statistical summaries. This supports tractable training in large populations but often assumes homogeneity and synchronized execution [15, 58]. Other paradigms

include *model-based MARL* [27, 59, 60, 28, 61] and *safe MARL* [30, 31, 62, 63, 64], which prioritize sample efficiency and safety constraints, respectively. These foundational paradigms form the basis for algorithmic exploration in MARL and serve as anchors for understanding MARL adaptability in learning and execution.

2.2 From Scalability to Adaptability

Among the various terms used to describe a MARL algorithm’s ability to handle diverse tasks, *scalability* has long been a central focus, particularly in domains involving large agent populations or limited communication bandwidth. A broad range of methods has been proposed to address this challenge, spanning four primary directions. *Graph-based factorizations* decompose global objectives into local components using factored MDPs [65, 66], coordination graphs [67], and sparse Q-learning variants [68, 69], with extensions to partially observable settings via ND-POMDPs [70] and factored Dec-POMDPs [71]. Recent efforts have advanced scalability via decentralised actor-critic consensus [24, 25], correlation decay techniques [72, 73], and hypergraph-based coordination [74, 75]. *Mean-field approximations* simplify learning by modelling interactions through population-level statistics [13], enabling tractable coordination in large-scale systems [76], with extensions to partially observable [77] and heterogeneous-agent environments [58]. Theoretical advances in mean-field control [78, 79] further recast the MARL problem as a single-agent control problem in the large-agent limit. *Swarm intelligence* approaches, inspired by biological collectives, rely on decentralised heuristics for search and coordination [80]. Benchmarks such as MAgent [76] and Neural MMO [81] demonstrate emergent behaviours among thousands of agents. While inherently scalable, swarm-based methods typically require heavy domain-specific reward shaping and are often task-specific, with limited generalisability across scenarios. Although these methods have advanced scalability in terms of sample efficiency and population size, they often assume static settings in other aspects such as agent roles, task objectives, and interaction patterns—conditions that rarely hold in real-world application [39, 82, 83]. When settings become dynamic and learning objective starts shifting, these algorithms frequently exhibit performance degradation or behavioural instability [84, 85, 86].

To address this gap, we advocate for a shift in focus from scalability alone to a broader concept we term *adaptability*. Unlike scalability, which primarily addresses computational or representational efficiency as the agent count grows, adaptability concerns an algorithm’s ability to maintain performance when broader task setting changes. In the following sections, we introduce a structured framework describe MARL adaptability via three interrelated dimensions: *Learning Adaptability*, *Policy Adaptability*, and *Scenario-Driven Adaptability*, which offers more comprehensive lens for assessing real-world readiness of MARL methods.¹

3 Learning Adaptability

We begin by examining *learning adaptability*—the capacity of a MARL algorithm to remain stable and effective under diverse training conditions, even when those conditions diverge from its initial design assumptions. This dimension addresses three fundamental questions:

¹This survey on MARL adaptability is based on algorithms published in top-tier venues such as NeurIPS, ICML, ICLR, AAAI, TNNLS, JMLR, TPAMI, and AAMAS, primarily from 2017 onward. Rather than providing an exhaustive enumeration, we focus on well-cited foundational works and representative recent approaches.

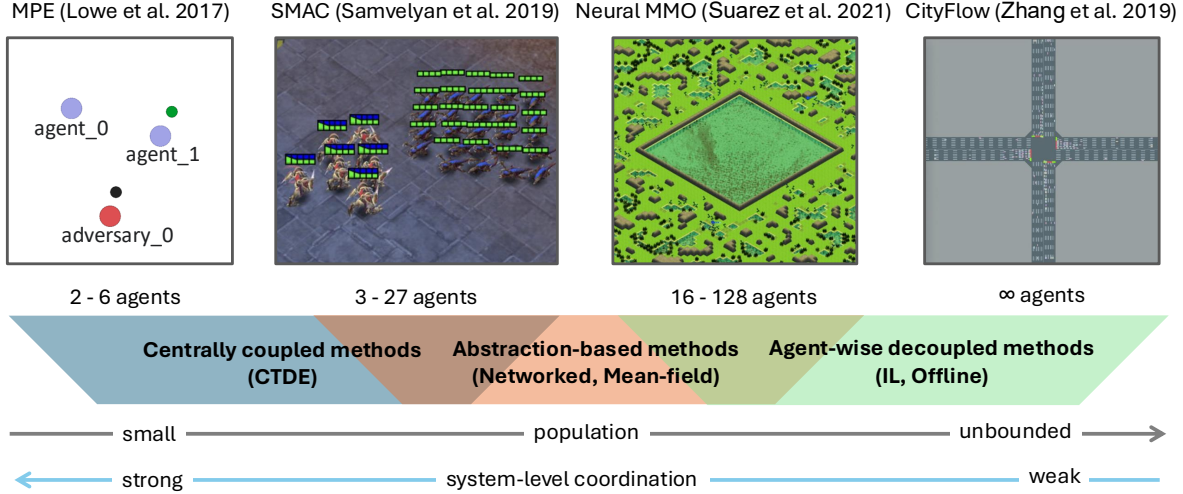


Figure 3: MARL environments exhibit significant variability in agent population scale, ranging from small-scale settings like MPE [33] (2–6 agents) and SMAC [87] (3–27 agents), to large-scale simulations such as Neural MMO [81] (16–128 agents) and CityFlow [88] (unbounded). As agent populations increase, the degree of system-level coordination typically decreases, as interactions become sparser and more local. This trend also reflects the strengths and limitations of different MARL paradigms: centrally coupled methods offer high coordination fidelity but struggle to scale; agent-wise decoupled methods scale well but exhibit weaker coordination; abstraction-based methods offer a middle ground, supporting moderate scalability while preserving coordination structure.

Key Questions for Learning Adaptability

1. Can the algorithm maintain training stability and convergence as the agent population scales up or down?
2. Is it robust across a range of task learning objectives—cooperative, collaborative, competitive, or mixed?
3. Can it be trained under real-world execution constraints, such as distributed infrastructure or asynchronous decision-making?

To systematically evaluate these questions, we analyse representative MARL paradigms across three shifting axes:

- *Population Scaling* (Sec. 3.1): Examines how algorithms cope with growing or shrinking agent populations and the associated scalability-robustness trade-offs.
- *Applicability Across Task Structures* (Sec. 3.2): Investigates the extent to which learning paradigms generalize to cooperative, collaborative, competitive, and mixed settings.
- *Execution Constraints on Distribution and Synchronization* (Sec. 3.3): Assesses the compatibility of MARL paradigms with distributed and asynchronous training environments.

Each axis highlights distinct challenges that arise when training is deployed beyond the confines of controlled benchmarks, emphasizing the importance of adaptable learning strategies in dynamic multi-agent systems.

3.1 Population Scaling

A critical dimension of learning adaptability in MARL is the ability to scale effectively with the number of agents. As agent populations grow, the dimensionality of the joint observation-action

space increases exponentially, making coordination increasingly difficult and expensive. In practice, environments with large agent populations often exhibit weaker system-level coordination, as the density of agent interactions becomes sparser and more localized. This variability introduces structural tensions between scalability and coordination fidelity, which must be carefully balanced by algorithmic design.

To analyse these trade-offs, we group MARL learning paradigms into three representative strategies based on how they structure agent interactions during training: (i) *centrally coupled optimization*, which enforces strong coordination via centralized critics or shared objectives; (ii) *agent-wise decoupling*, which emphasizes scalability by training agents independently or with minimal coupling; and (iii) *coordination via abstraction*, which approximates or restricts inter-agent dynamics using statistical or topological simplifications. These categories align with different points on the scalability-coordination spectrum, as illustrated in Fig. 3.

Centrally coupled methods embed inter-agent dependencies directly into the learning objective. Instantiated within the CTDE paradigm, they perform well in small to moderate populations [87, 89], where coordination is critical. Value Decomposition (VD) methods such as QMIX [14] and QPLEX [36] aggregate individual Q-values via monotonic mixing networks, enabling joint value estimation but incurring a scalability bottleneck. Reward Decomposition (RD) methods like COMA [34] or SHAQ [90] offer fine-grained credit assignment using counterfactual baselines or Shapley values, but scale poorly due to their combinatorial complexity. Centralized Critic (CC) methods such as MADDPG [33] and MAPPO [18] employ global critics that suffer from input explosion in large teams. Heterogeneous Agent (HA) approaches like HATRPO [19] reduce gradient interference via sequential updates but sacrifice sample efficiency in large populations.

Agent-wise decoupled strategies remove joint training dependencies, enabling scalable and distributed learning. Independent Learning (IL) methods like IQL [12] and IPPO [52] optimize policies based solely on local observations, allowing linear scalability. However, they typically struggle to learn coordinated behaviours in tightly coupled environments [54]. Offline MARL methods such as ICQ-MA [21] and OMAR [22] train from static datasets with no online interaction, but coordination must be implicitly embedded in the data. These methods scale well to large systems, yet often yield weaker global coordination.

Abstraction-based methods offer a middle ground, enabling partial coordination at scale via statistical or graph-based approximations. Mean-field MARL [13] represents agent interactions using population-level action statistics, achieving tractable learning in large homogeneous systems [76]. Networked MARL [24] introduces graph-structured communication or critic sharing, allowing agents to learn local coordination strategies. These abstractions reduce learning complexity but may limit coordination fidelity if assumptions on homogeneity or topology are violated [91].

In summary, population scaling in MARL reveals a fundamental trade-off: centrally coupled methods promote strong coordination but struggle with scalability; agent-wise decoupled methods scale to large systems but often underperform in cooperative tasks; abstraction-based methods strike a balance, offering scalable learning while maintaining a degree of coordination structure. These trade-offs highlight the need for adaptive designs that can modulate between scalability and coordination demands as agent populations grow.

3.2 Applicability Across Task Structures

While most MARL algorithms are designed with specific task structures in mind, real-world applications often present diverse or evolving inter-agent reward schemes. These settings span a spectrum from fully cooperative to competitive, collaborative, and mixed interactions. Although it is not essential for a method to perform optimally across all task types, the ability to generalize

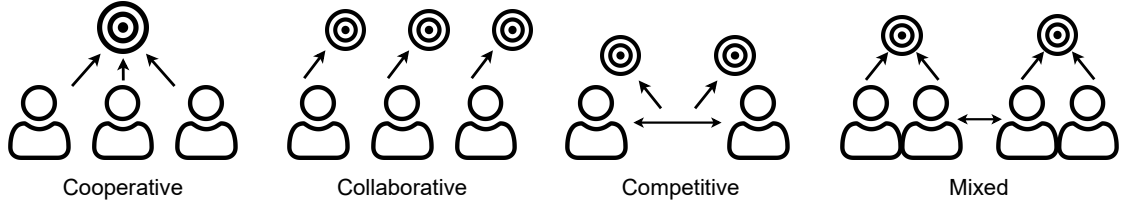


Figure 4: Illustration of four task modes in multi-agent systems categorized by reward structures: (a) Cooperative tasks where all agents share an identical global reward; (b) Collaborative tasks with similar but individual rewards for each agent; (c) Competitive tasks characterized by zero-sum individual rewards; and (d) Mixed tasks combining intra-team shared rewards and inter-team zero-sum competition. These distinctions highlight varying coordination and conflict dynamics fundamental to multi-agent learning.

without structural redesign is a hallmark of learning adaptability—particularly in open-ended or mission-driven domains.

To assess this axis of adaptability, we focus on four canonical task modes, summarized in Fig. 4. Each task mode imposes different coordination and conflict dynamics and highlights distinct inductive biases in algorithm design. We evaluate the alignment of each paradigm by categorizing its suitability as: (i) *natively suitable*, (ii) *broadly applicable*, (iii) *usable with task adjustment*, or (iv) *incompatible*.

Fully cooperative tasks require agents to optimize a shared global reward. CTDE paradigms including VD [14, 35, 92], RD [93, 90, 94], CC [33, 34, 18], and HA [19, 38, 37] are natively suitable due to their tight coupling between agent policies and joint objectives [87, 54]. Abstraction-based methods such as mean-field MARL [13, 15, 95] and networked architectures [96, 26, 24] also support effective cooperation in large-scale settings.

Collaborative tasks exhibit partially aligned objectives and often decentralized observability. IL methods [12, 52] and networked MARL naturally support such settings, leveraging local policy optimization [25, 24] and communication graphs [97, 91, 98]. CC and mean-field methods are broadly applicable but may require task-specific reward shaping. VD and RD methods, which rely on a globally shared reward signal, are generally unsuitable for collaboration without significant modifications.

Competitive tasks involve directly conflicting goals across agents or teams. Model-based MARL [60, 28, 61] is natively suited to these settings, offering capabilities for planning and opponent modelling. Offline MARL [21, 22] can be conditionally effective if adversarial interactions are well-represented in the data. IL and mean-field methods are usable only in restricted forms and typically lack mechanisms for anticipating adversarial strategies. CTDE methods are fundamentally misaligned with competitive settings due to their cooperative training assumptions.

Mixed cooperative-competitive tasks require intra-team coordination and inter-team competition. Model-based MARL again provides strong adaptability, supporting hierarchical reasoning across team boundaries. CTDE methods can be applied within teams but are limited in generalizing across teams with competing objectives. Networked MARL can support such hybrid scenarios when communication structures are carefully defined. Offline MARL retains conditional applicability, provided task transitions are present in the training data.

Overall, the suitability of a MARL paradigm is closely tied to the task structure. Although no paradigm is universally optimal, robust performance across diverse tasks without requiring structural or algorithmic changes remains one of the key markers of a learning paradigm’s adaptability.

Table 1: Learning adaptability of MARL paradigms across seven dimensions: Pop. Scal. = Population Scaling, Coop. = Cooperative, Collab. = Collaborative, Comp. = Competitive, Dist. Train. = Distributed Training, Async. Exec. = Asynchronous Execution.

Paradigm	Pop. Scal.	Coop.	Collab.	Comp.	Mixed	Dist. Train.	Async. Exec.
Value Decomposition (VD)	✗	✓	✗	✗	✗	✗	✗
Reward Decomposition (RD)	✗	✓	✗	✗	✗	✗	✗
Centralized Critic (CC)	△	✓	✓	✓	✓	△	△
Heterogeneous Agent (HA)	△	✓	✗	✗	✗	△	✗
Independent Learning (IL)	✓	✓	✓	✓	✓	✓	✓
Offline MARL	△	△	△	△	△	✗	△
Model-Based MARL	△	✓	✓	✓	✓	✗	△
Mean-Field MARL	✓	✓	✓	✗	✓	✗	✗
Networked MARL	✓	✗	✓	✗	△	✓	✗

Legend: ✓— natively suitable, △— partially suitable or task-dependent, ✗— incompatible.

3.3 Execution Constraints on Distribution and Synchronization

In real-world deployments, MARL algorithms must address not only the challenges of coordination and learning, but also the system-level constraints imposed by practical execution environments. Two constraints are particularly prominent. First, *distributed training infrastructure* is often required, where agents are deployed across physically or logically disjoint computational nodes. This arises in settings such as edge-computing robotics, networked sensors, and geographically dispersed systems, necessitating training paradigms that minimize centralized dependencies. Second, many applications involve *asynchronous execution*, where agents operate at different temporal resolutions or update frequencies due to hardware heterogeneity, task allocation, or communication latency. These constraints

Distributed Training Support. Distributed training enables agents to learn in parallel across decentralized systems, reducing computational bottlenecks and accommodating bandwidth or privacy constraints. IL algorithms [12, 52] are fully compatible with distributed training: agents optimize local policies using local observations, without parameter sharing or joint critics. Networked MARL [24, 96] also supports distributed learning through peer-to-peer message passing or consensus-based updates, requiring no centralized coordination. In contrast, most CTDE methods, such as VD, CC, and HA, are only partially compatible. These algorithms permit decentralized execution but require centralized training inputs or synchronized gradient updates. This limits their deployment in bandwidth-constrained or privacy-sensitive systems. Model-based MARL [99], mean-field MARL [13], and offline MARL [22] generally assume access to global state information or joint trajectories during training, making them incompatible with fully distributed infrastructure.

Asynchronous Execution Support. Asynchronous adaptability is crucial in environments where agents act independently, receive delayed observations, or are triggered by event-driven processes. IL methods are natively asynchronous: agents update independently and require no shared timing, making them well-suited for real-time robotic and sensor network applications [52]. CC methods, such as MADDPG [33], can be adapted to asynchronous settings by decoupling critic evaluations

from agent policy updates. Similarly, model-based MARL may support asynchronous dynamics if agent-specific transition functions are modelled separately. Offline MARL methods exhibit asynchronous compatibility during deployment, as policy inference does not require synchronized execution. However, their training-phase adaptability depends on the structure of the offline dataset: if trajectories reflect synchronized execution, learned policies may inherit synchrony assumptions [100]. Most other paradigms including VD, HA, mean-field, networked MARL, and safe MARL assume synchrony for value aggregation, communication, or constraint satisfaction, and thus fail in environments with variable update rates or latency [30, 32].

In summary, execution adaptability varies widely. As shown in Table 1, IL and networked MARL offer the greatest flexibility for distributed, asynchronous deployment. CTDE methods are effective under centralized infrastructure but require careful scheduling. Model-based and offline approaches provide partial compatibility, contingent on problem structure and data assumptions.

4 Policy Adaptability

While learning adaptability focuses on training-time resilience under varied conditions, it does not fully capture the *generalization behaviour of the resulting policy*. In real-world deployments, agents often face unforeseen circumstances, including evolving task specifications, dynamic team configurations, and unfamiliar partner behaviours. This motivates the second axis of our framework: *policy adaptability*, which we define as the ability of a learned policy to generalize effectively across related tasks, agent roles, or coordination structures without requiring retraining.

Figure 5 illustrates the conceptual distinction between learning and policy adaptability. Learning adaptability considers whether an algorithm can be independently trained on multiple diverse tasks. In contrast, policy adaptability evaluates whether a single policy, trained under a specific configuration, can be reused or adapted for other tasks with minimal adjustment. The following questions are central to understanding policy adaptability:

Key Questions for Policy Adaptability

1. Can a policy be reused in new tasks without requiring architectural changes?
2. Can a policy coordinate agents across tasks with differing objectives, settings, or previously unseen partners?
3. Can we learn a sufficiently general policy from offline or expert data to improve generalization?

These questions can also be viewed from the perspective of task structure. When task semantics lie within a well-defined space, the acquired knowledge can be effectively transferred across tasks. Representative types of task gaps include: changes in the number of agents (affecting the policy’s input-output structure), shifts in objectives or environmental settings (task generalization), sequential or incremental task progressions (lifelong learning), and the presence of novel partners with unfamiliar behaviours. As illustrated in Figure 6, different methods are suited to bridging different types of task gaps.

To explore these challenges, we organize the literature into five methodological categories that explicitly promote generalization across tasks and agents.

- *Permutation-Aware Policy* (Sec. 4.1): Investigates invariant and equivariant architectures that decouple policy learning from agent identity and ordering.
- *Offline Pretraining* (Sec. 4.2): Leverages static datasets and decision transformers to encode generalized behavioural priors.

- *Task Representation and Relation Modelling* (Sec. 4.3): Utilizes task embeddings and trajectory-based structure discovery to support transfer across task distributions.
- *Continual, Curriculum, and Meta-Learning* (Sec. 4.4): Employs progressive training schemes and fast adaptation mechanisms to support lifelong learning in evolving environments.
- *Zero-Shot Coordination* (Sec. 4.5): Focuses on social generalization, enabling agents to align with unfamiliar teammates without shared training history.

Collectively, these approaches form the foundation for building general-purpose, transferable MARL policies capable of adapting fluidly across tasks, agents, and deployment scenarios.

4.1 Permutation-Aware Policy

A foundational component of policy adaptability is the architectural capacity to represent coordination patterns independent of agent identities or ordering. Recent work has converged on three core principles for enabling permutation-aware generalization in multi-agent policy models: the use of transformer architectures [84], entity-wise attention mechanisms [85], and explicit enforcement of permutation invariance or equivariance [86, 101].

Transformers for Variable Agent Inputs. Transformer-based policy models offer a natural framework for processing variable-length, unordered agent observations. One representative approach is UPDeT [84], which introduces policy decoupling through transformers. By representing agent observations and actions as sets of entities, UPDeT processes these inputs via a shared attention backbone, allowing the same policy model to generalize across tasks with different numbers and types of agents. This design obviates the need for handcrafted task-specific modules and enables multi-task generalization from a unified architecture.

Entity-Wise Attention and Factorization. Entity-centric models extend this paradigm by isolating reusable interaction patterns across agents. For instance, Randomized Entity-wise Factorization [85] partitions input observations into semantic substructures, enabling policies to attend selectively to relevant context while maintaining awareness of task dynamics. Such designs decouple agent-specific variability from shared coordination signals, facilitating transfer across scenarios with heterogeneous inputs and team structures.

Permutation Invariance and Equivariance. Several models go further by explicitly encoding permutation symmetry into the policy architecture. Dynamic Permutation Networks [86] and HGAP [101] construct neural modules whose outputs are invariant or equivariant to reordering of agents. These models combine modular feature extractors with hypernetworks or graph attention layers to enforce symmetry constraints. The resulting policies maintain consistent behaviour under arbitrary agent re-indexing, a crucial property for generalization to unseen team compositions.

Together, these architectural strategies provide the structural groundwork for adaptable policies. By abstracting away agent identity and focusing on relational or set-based reasoning, permutation-aware policies form the basis for downstream generalization in offline learning, transfer learning, and zero-shot coordination scenarios.

4.2 Offline Pretraining

Offline pretraining offers a promising pathway to improve policy adaptability by learning from diverse multi-task datasets without requiring interactive environment access. These approaches aim to extract transferable structure from offline trajectories and deploy a unified model across tasks or agent populations. Recent work has introduced techniques that vary along three key axes: decision-transformer modelling [100], offline multi-task learning architectures [102, 103], and mask-based generalization mechanisms [104].

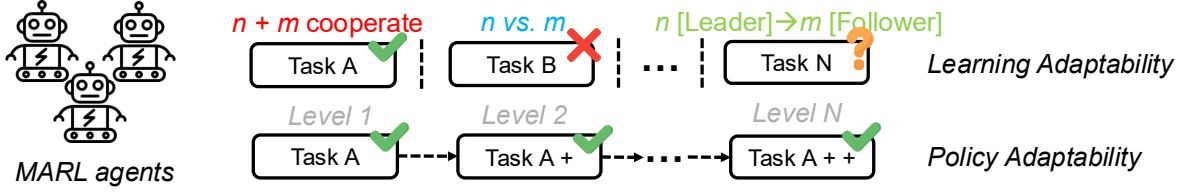


Figure 5: Illustration of the key distinction between learning adaptability and policy adaptability. Learning adaptability evaluates whether an algorithm can be trained and deployed on different tasks (e.g., Task A, B, C) under varying training conditions. Policy adaptability instead evaluates whether a trained policy (e.g., from Task A) can generalize or transfer to related tasks (e.g., Task A +, A + +) with minimal or no retraining.

Decision Transformer for Sequence modelling. A central innovation in this space is the application of decision transformer architectures [105] to multi-agent settings. Multi-Agent Decision Transformers (MADT) [100] extend autoregressive models to encode multi-agent trajectories conditioned on prior state, action, and return sequences. These models are pretrained over offline datasets collected from diverse tasks, allowing them to generalize behaviour without explicit fine-tuning. The sequence modelling framework allows policies to reason over temporal dependencies and supports compositional reuse of learned behaviours across task boundaries.

Offline Multi-Task Learning with Task Conditioning. Beyond raw sequence modelling, offline pretraining methods increasingly leverage modularity to enhance task transfer. M3 [102] introduces a task-conditioned architecture that incorporates explicit prompts and agent-invariant embeddings. It utilizes a vector-quantized variational autoencoder (VQ-VAE) [106] to encode heterogeneous agent roles and supports a decoupled representation of shared skills and task-specific behaviours. Similarly, hierarchical frameworks [103] learn decomposable sub-policies, enabling selection of both generic and task-tailored strategies at inference time. These models facilitate generalization to novel tasks by encoding structured latent representations of skill and role.

Mask-Based Generalization Across Agent Variability. To accommodate input and output variability across tasks, recent work introduces masking strategies that condition the policy on different subsets of observation-action dimensions. MaskMA [104] combines transformers with mask-based training to enable a single model to handle agents with differing roles, modalities, or interface structures. This approach supports strong zero-shot transfer, as the model implicitly learns a general action representation that is robust to agent-specific permutations and dimensionality shifts.

Collectively, these methods demonstrate that offline pretraining can serve as a foundation for adaptable MARL. By integrating decision sequence modelling, modular task representation, and input masking, they enable broad generalization across agent configurations and task domains with minimal reliance on online fine-tuning.

4.3 Task Representation and Relation Modelling

A growing body of work enhances policy adaptability by constructing explicit representations of tasks and inter-agent relationships. Rather than relying solely on architecture or data scale, these methods aim to encode transferable structure across tasks, enabling more efficient reuse and generalization. Key directions include task embedding [107, 108, 109], trajectory-based clustering [110, 111], and credit assignment strategies that model inter-task and inter-agent relations [112, 113].

Task Embedding for Policy Conditioning. Task embedding methods learn compact representations of task identity or context and use them to modulate policy execution. These embeddings

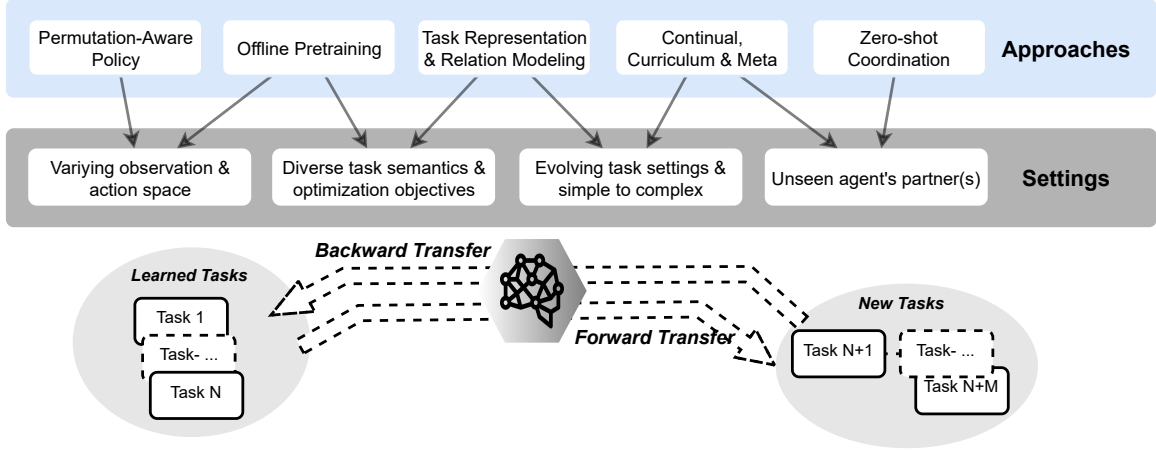


Figure 6: Relationship between adaptable policies and available approaches. An adaptable policy should handle new tasks or partners through forward transfer while retaining prior knowledge via backward transfer, all with minimal adjustment.

may capture global task features [107] or agent-specific roles [108], and can be inferred from initial observations or support trajectories. By conditioning policy outputs on these learned embeddings, agents can adapt behaviour with minimal retraining. Methods such as subtask encoders [109] decompose complex objectives into reusable latent components, supporting structured policy reuse across tasks with shared skill dependencies.

Trajectory Clustering for Transferable Structure. Another line of work clusters past trajectories to uncover recurring coordination patterns or latent task variations. This allows agents to identify and transfer strategies from similar prior experiences. For example, unsupervised task discovery via trajectory distributions [110] enables zero-shot generalization by matching new tasks with similar previously encountered scenarios. These approaches can also be integrated with transformer-based regret modelling [111], where clustering guides adaptive loss weighting across tasks of varying difficulty or progression rates.

Relation-Aware Credit Assignment. To further facilitate generalization, several methods incorporate relational reasoning across tasks and agents. By modelling similarity between task objectives or environmental dynamics [112], policies can selectively transfer knowledge between aligned domains while avoiding negative transfer. Relation-aware credit assignment [113] extends this idea to cooperative settings, ensuring reward attribution reflects task-specific dependencies and agent roles. This relational perspective supports more robust policy updates across heterogeneous or evolving task structures.

Together, these methods enhance policy adaptability by enabling agents to extract, represent, and reason over latent structure in multi-task settings. By leveraging task embeddings, trajectory-based clustering, and relation-aware reasoning, they provide scalable mechanisms for transfer and generalization in MARL.

4.4 Continual, Curriculum, and Meta-Learning

Another pathway to policy adaptability is through mechanisms that support dynamic policy evolution over time. These methods approach generalization as a process of continual refinement, where agents accumulate and adapt knowledge progressively across changing task distributions. Key strategies include structured curricula, modular continual learning, and meta-learning for fast

adaptation.

Curriculum Learning for Task Staging. Curriculum learning organizes the training process into sequences of tasks with increasing complexity, enabling agents to bootstrap performance in difficult scenarios from prior experience with simpler ones. Evolutionary curricula [114] gradually introduce harder coordination challenges by modifying team compositions or interaction structures. More adaptive schemes [115] employ contextual bandit models to dynamically construct curricula based on agent progress, supporting stable population-invariant learning under sparse rewards. Another approaches propose an auto-curriculum, where the training curricula is generated automatically. For sparse-reward environments, [116] employs task expansion and entity progression by effectively generating training curricula that adapt both the task configurations and the number of participating entities to promote scalable and adaptable multi-agent learning. In zero-sum games, the study in [117] introduces subgame curriculum learning to accelerate learning process, where agents progressively train on strategically selected subgames to build competence before tackling the full game complexity. PORTAL [118] automatically generates a curriculum by learning a shared feature space across tasks, enabling it to characterize tasks based on their feature distributions and prioritize those most similar to the target task for more effective transfer and adaptation. Moreover, curriculum learning has been used to construct policy with better scalability. A novel network architecture, Dynamic Agent-number Network (DyAN) [119] introduces transfer mechanisms across curricula, which efficiently handles varying numbers of agents through dynamic input sizing. Also, in [120], the auto-curriculum approach begins by training agents in multi-agent scenarios with a small number of agents and progressively increases the number of agents, allowing the learning process to adapt gradually to more complex interactions. These curricula promote sample-efficient generalization and reduce the risk of unstable learning in large or heterogeneous populations.

Progressive Heads for Continual Coordination. Continual learning strategies often adopt modular architectures to mitigate catastrophic forgetting. Progressive task heads [121] allocate separate output modules for each task while sharing a common feature backbone, allowing new skills to be acquired without overwriting prior ones. Complementary approaches prioritize experience replay or task sampling based on hindsight-derived importance metrics [122], maintaining performance across previously encountered tasks while encouraging exploration. These methods support lifelong adaptation by explicitly managing task-specific knowledge retention and reuse.

Meta-RL for Fast Generalization. Meta-reinforcement learning formulations target fast policy adaptation across tasks by learning how to learn. Collaborative meta-RL [123] encodes agent-type relationships to facilitate coordination transfer between heterogeneous teams, even under role or capability shifts. Transformer-based approaches [124] implement credit assignment at the coalition level, allowing agents to generalize coordination strategies under dynamic observation and action spaces. These meta-learning architectures accelerate adaptation to new tasks or partner configurations by internalizing transferable learning priors.

Collectively, curriculum learning, continual modularity, and meta-adaptation provide a framework for progressive coordination development. These methods extend policy adaptability beyond static generalization by enabling agents to grow and adapt their behaviour in response to evolving task structures and team compositions.

4.5 Zero-Shot Coordination

Beyond task generalization, a distinct facet of policy adaptability in MARL is an agent’s ability to coordinate with unfamiliar partners without additional retraining—referred to as *zero-shot coordination* (ZSC) [125, 126, 127, 128]. This setting reflects real-world deployments where agents may be developed independently or trained under different assumptions [129, 45]. Success in ZSC requires

policies that are robust to partner variability and capable of aligning with unknown conventions or behaviours.

Other-Play for Convention Robustness. The foundational approach to ZSC is the *Other-Play* (OP) framework [125], which addresses the failure of self-play to produce partner-compatible conventions. In symmetric cooperative settings, self-play policies often rely on arbitrary symmetry-breaking strategies (e.g., always going left), leading to coordination failure when paired with independently trained agents. OP combats this by randomizing symmetric factors during training, guiding policies toward shared, robust conventions. OP demonstrated significant gains in Hanabi, where conventional self-play policies failed to align with diverse partners.

Trajectory Diversity for Coordination Generality. While OP assumes known symmetries, subsequent work focuses on learning partner-compatible behaviour without explicit symmetry modelling. The *Trajectory Diversity* (TrajeDi) framework [126] encourages policies that generate diverse trajectory distributions while preserving coordination potential. Using a generalized Jensen-Shannon divergence objective, TrajeDi increases robustness to unseen policies even in partially observable environments. This approach enables ZSC in more complex or asymmetric domains where coordination structure is latent or unknown.

Belief Modelling and Open-Ended Coordination. Recent advances incorporate explicit reasoning over partner behaviour to improve ZSC performance. The *Any-Play* framework [127] trains agents to maximize cross-play success by optimizing compatibility with policies trained under different algorithms. It combines diversity-driven self-play with policy augmentation, producing strategies resilient to inter-algorithm mismatches. Similarly, the *Cooperative Open-ended Learning* (COLE) framework [128] models agent compatibility as a graph-theoretic problem, iteratively refining strategies via preference graphs and best-response dynamics. These methods move beyond symmetry-breaking to consider partner belief modelling and distributional alignment in cooperative learning.

Together, these approaches redefine policy adaptability through the lens of social compatibility. Instead of adapting to new tasks, ZSC emphasizes generalization across agent identities, a critical capability for deployment in decentralized or mixed-agent environments.

5 Scenario-Driven Adaptability

While learning and policy adaptability address algorithmic robustness and generalization capacity, they ultimately depend on the availability of environments that faithfully capture real-world variability. We refer to this third axis as *scenario-driven adaptability*, the ability of MARL systems to be evaluated and trained in environments that reflect diverse, dynamic, and structurally complex settings.

This dimension emphasizes the role of *environment design* in fostering reliable assessments of MARL adaptability. Unlike fixed-task benchmarks, scenario-driven environments allow researchers to vary agent populations, communication structures, reward types (cooperative, competitive, mixed), and execution constraints (asynchrony, decentralization). Such flexibility is essential for simulating real-world challenges, from dynamic teaming to long-horizon deployment.

Accordingly, this section addresses the following questions central to scenario-driven adaptability:

Key Questions for Scenario-Driven Adaptability

1. Do benchmarks allow configurable agent setups (e.g., population size, heterogeneity, asynchrony)?
2. Can they support progressive adaptation across task sequences (e.g., curriculum or continual learning)?
3. Do they enable generalization from offline data and coordination with novel partners or conventions?

To explore these questions, we structure our review around three major themes:

- *Survey of MARL Benchmarks* (Sec. 5.1): Reviews structured games, application-oriented simulators, and emerging LLM-based systems, with attention to their configurability and evaluation fidelity.
- *Continual and Curriculum Scenarios* (Sec. 5.2): Discusses benchmark affordances for transfer-compatible design and representational continuity across evolving task sequences.
- *Offline Pretraining, Online Transfer, and Zero-Shot Scenarios* (Sec. 5.3 and Sec. 5.4): Examines environment support for evaluating offline-to-online transfer, generalization from fixed datasets, and coordination with novel partners.

Together, these components provide a foundation for scenario-driven evaluations that complement algorithmic innovations, ensuring that MARL research is grounded in practical, scalable, and dynamic deployment contexts.

5.1 Survey of MARL Benchmarks

We review a broad range of benchmarks commonly used to evaluate MARL. Through summarization and comparison, we aim to highlight the configurability and diversity of environments within each category, as well as to analyse their limitations in capturing the complexities of real-world deployment conditions.

Structured Games. Structured games constitute the foundation of many MARL algorithmic developments, providing minimal yet expressive settings for studying inter-agent coordination, credit assignment, and policy generalization. Benchmarks such as MPE [33], SMAC [87], GRF [89], and RWARE [130] offer tractable environments with 2–27 agents, well-defined action-observation structures, and support for cooperative or mixed objective formulations. For example, SMAC enables heterogeneous unit control under cooperative goals with partial observability, while MPE introduces mixed-sum scenarios with optional communication and varying team sizes. More scalable environments such as MAgent [76], Neural MMO [81], and MAPF [131] support populations exceeding 100 agents, albeit often without full observability or agent heterogeneity.

The diversity of structured games extends further across dimensions such as communication modality (e.g., Overcooked [129], MACO [132]), asynchronous interactions (e.g., Matrix Games [133], MARL"O [134]), and the degree of environment customizability. Environments like Hide-and-Seek [135] and Hallway [136] provide sparse or emergent coordination structures, while Hanabi [45] uniquely emphasizes belief modelling and implicit coordination. Collectively, structured games support controlled experimentation and comparative analysis, though they often abstract away complexities inherent to real-world deployments.

Application-Oriented Simulators. Application simulators advance beyond abstract coordination to incorporate real-world constraints, offering a higher-fidelity testbed for evaluating MARL

algorithms under deployment-oriented assumptions. These include robotic manipulation domains (e.g., MAMuJoCo [92], Bi-DexHands [137]), smart city and mobility simulators (e.g., CityFlow [88], SUMO [138]), and emerging space and sustainability applications (e.g., BSK-RL [82], SustainDC [139]). These simulators typically support larger populations (e.g., MAPDN [83], MetaDrive [39]), asynchronous execution (e.g., SMARTS [40], WFCRL [140]), and diverse agent roles or morphologies (e.g., MARBLER [141], MaMo [142]).

Many application simulators prioritize task realism and configurability. MetaDrive includes procedural generation of traffic and road topology, facilitating generalization to novel driving conditions. BSK-RL and Flatland [143] support mission composability and dynamic task scaling. Meanwhile, LAG [144], MABIM [145], and SMARTS offer fine-grained agent-environment interactions with mixed-motive tasks and temporal constraints. Although such environments introduce higher variance and computational demands, they are indispensable for stress-testing policy robustness.

LLM-Based Multi-Agent Systems. Recent work has introduced language-enabled multi-agent environments that leverage large language models (LLMs) to support open-ended, natural language-driven interactions. These settings emphasize high-level reasoning, task modularity, and human-aligned communication protocols. Environments such as Welfare [146], AgentVerse [147], and Collab-Overcooked [148] focus on cooperation, negotiation, and task decomposition via structured prompts. Llmarena [149] and BattleAgentBench [150] extend this paradigm to adversarial and mixed-motive interactions, incorporating multi-round strategy evolution and language-based negotiation. These environments typically support 2–8 agents under partial observability and rely on prompt-based control interfaces, offering a unique lens into emergent coordination and role specialization. However, they currently lack standardized protocols and evaluation metrics, limiting their utility for systematic comparison.

Benchmark Characterization. To enable a systematic comparison, Table 2 provides an overview of representative MARL benchmarks across seven key dimensions. Although Structured Games, Application Simulators, and LLM-based environments differ in their underlying assumptions, task abstractions, and complexity, they exhibit considerable overlap in core features. Across all categories, one can find environments that support large-scale agent populations, partial observability, inter-agent communication, diverse reward structures, and asynchronous execution or agent heterogeneity. Notably, heterogeneity and task customizability are not confined to any single category but instead manifest differently depending on the simulation context and design emphasis. Rather than implying a strict progression or superiority among these categories, this taxonomy highlights the importance of aligning benchmark selection with the specific goals of a study. Structured games provide controlled, interpretable environments well-suited for analysing coordination strategies and algorithmic components in isolation. Application-oriented simulators introduce realistic dynamics, environmental stochasticity, and mission-driven variability, thereby enabling evaluation under more deployment-relevant conditions. In contrast, LLM-based environments foreground natural language interfaces, emergent role specialization, and high-level reasoning, offering a unique lens into communication and generalization in language-mediated multi-agent systems. Collectively, these benchmarks form a complementary suite, and their utility should be assessed in terms of research intent like whether to isolate algorithmic contributions, test robustness in realistic domains, or explore language-grounded agent interaction.

Table 2: Environment diversity and configurability. Each benchmark is assessed across seven dimensions: population range, communication and observability structures, learning objectives, support for asynchronous execution, agent heterogeneity, task customizability, and the number of available tasks. Note: if an environment does not explicitly provide predefined scenarios but allows them to be generated within a continual learning setup, we assign a value of 1 for its task count.

	Environment	Pop. Scale	Comm./Obs.	Objective	Async.	Hetero.	Customize.	Tasks
Structured Games	Matrix Game [133]	2	No / Full	Mixed	✓	✓	✓	1
	RWARE [130]	2–4	No / Partial	Coop	✗	✗	✓	1
	MPE [33]	2–6	Yes / Partial	Mixed	✓	✓	✓	6
	SMAC [87]	2–27	No / Partial	Coop	✗	✓	✓	23
	GRF [89]	2–22	No / Full	Mixed	✗	✓	✗	7
	MAgent [76]	>100	No / Partial	Mixed	✓	✗	✓	6
	GoBigger [151]	>100	No / Partial	Mixed	✗	✗	✓	1
	Overcooked [129]	2–4	No / Full	Coop	✓	✓	✓	5
	Pommerman [152]	2–4	No / Full	Mixed	✗	✗	✗	3
	SISL [153]	3–8	No / Full	Coop	✗	✗	✗	3
	Hanabi [45]	2–5	No / Partial	Coop	✓	✗	✗	4
	MACO [132]	5–15	Yes / Partial	Mixed	✗	✗	✓	6
	MARLÖ [134]	2–8	No / Partial	Mixed	✓	✗	✓	14
	Hallway [136]	2	Yes / Partial	Coop	✗	✗	✓	1
	Hide-and-Seek [135]	2–6	No / Full	Mixed	✗	✓	✗	1
	Gathering [154]	2	No / Full	Coop	✗	✗	✓	1
	MAPF [131]	>100	No / Partial	Coop	✗	✓	✓	27
	DCA [155]	2–30	No / Partial	Mixed	✗	✗	✓	1
	Neural MMO [81]	>100	No / Partial	Mixed	✓	✗	✓	1
Application Simulators	Bi-DexHands [137]	2	No / Full	Coop	✗	✓	✓	17
	MATE [156]	2–16	Yes / Partial	Mixed	✗	✓	✓	5
	MAMuJoCo [92]	2–6	No / Full	Coop	✗	✓	✗	10
	SustainDC [139]	3	No / Partial	Coop	✗	✓	✓	1
	MAPDN [83]	>100	No / Partial	Coop	✗	✓	✗	3
	LAG [144]	2–8	No / Partial	Mixed	✗	✗	✓	3
	BSK-RL [82]	>100	Yes / Partial	Coop	✗	✗	✓	2
	WFCRL [140]	7–92	No / Partial	Mixed	✓	✗	✗	2
	CityFlow [88]	>100	No / Partial	Coop	✓	✗	✗	1
	MetaDrive [39]	20–40	No / Partial	Mixed	✓	✗	✓	7
	Flatland [143]	>100	No / Partial	Coop	✓	✗	✓	1
	SUMO [138]	2–6	No / Full	Mixed	✗	✗	✓	1
	MARBLER [141]	4–6	Yes / Partial	Mixed	✗	✓	✓	5
	MaMo [142]	2–4	No / Partial	Coop	✗	✓	✓	8
	MABIM [145]	>100	No / Partial	Mixed	✗	✗	✓	2
	SMARTS [40]	3–5	No / Partial	Mixed	✓	✓	✓	1
LLM-based Benchmark	Welfare [146]	2–7	Yes/Full	Coop	✗	✓	✓	1
	Magic [157]	3	Yes/Partial	Coop	✗	✗	✗	5
	Agentverse [147]	2–3	Yes/Partial	Coop	✓	✓	✓	3
	Avalonbench [158]	5	Yes/Partial	Mixed	✗	✓	✗	1
	Villageragent [159]	2–8	No/Partial	Coop	✗	✓	✓	3
	Llmarena [149]	2–5	Yes/Partial	Mixed	✓	✓	✓	7
	Battleagentbench [150]	1–6	Yes/Partial	Mixed	✗	✗	✓	3
	PokerBench [160]	6	No/Partial	Comp	✗	✗	✗	1
	Multiagentbench [161]	2–7	Yes/Partial	Mixed	✗	✓	✓	6
	Collab-Overcooked [148]	2	No/Full	Coop	✗	✓	✓	6

5.2 Continual and Curriculum Scenarios

Scenario-driven adaptability is not only about diversity in environment configurations, but also about how tasks evolve over time to systematically probe the generalization capacity of MARL policies. In this context, continual and curriculum learning paradigms serve as structured approaches to designing *task gaps*—the differences between successive tasks—that reveal an algorithm’s ability to adapt across changing conditions [114, 115, 116, 117, 120]. These gaps, whether arising from increasing population size, added agent heterogeneity, or modified reward semantics, are essential to testing how well agents can *reuse* prior knowledge, *retain* learned behaviours, and *extend* existing capabilities to meet new demands.

Specifically, *reuse* refers to the agent’s capacity to apply learned behaviours to similar but incrementally more complex settings; *retention* concerns the stability of earlier skills when new learning occurs; and *extension* captures the ability to build on previous knowledge to solve qualitatively novel tasks. The structure and semantics of task gaps determine whether these dimensions of adaptability can be meaningfully assessed. As such, continual and curriculum settings provide a principled framework for designing multi-agent scenarios that gradually introduce complexity while preserving meaningful continuity in learning signals and task structure.

Task Gaps and Principles for Transfer-Compatible Design. Designing transfer-compatible task gaps requires more than simply increasing difficulty. Effective curricula must preserve semantic alignment across tasks to ensure that policy improvements reflect generalization rather than task-specific tuning. Several principles have emerged as critical for structuring meaningful transitions in multi-agent settings:

1. *Incremental population scaling:* Gradually increasing the number of agents (e.g., $3 \rightarrow 5 \rightarrow 8$) supports the reuse of coordination strategies and exposes scalability constraints.
2. *Progressive role diversification:* Introducing new agent types or abilities in modular stages (e.g., Zealot \rightarrow Zealot+Stalker \rightarrow Zealot+Stalker+Colossus [87]) enables compositional policy learning and generalization to heterogeneous teams.
3. *Reward structure consistency:* Maintaining coherent reward objectives across tasks (e.g., always cooperative) reduces the need for strategy re-invention and promotes cumulative skill development.

Despite the promise of curriculum-based training, it remains underexplored in MARL. Most benchmarks lack support for structured task graphs [133, 130, 33], scaffolded skill progression [92, 83, 82], or staged evaluation protocols [142, 145, 40] that enable quantitative assessment of *forward transfer* (performance gains on future tasks due to earlier learning) or *backward transfer* (retention of earlier capabilities). These metrics are essential for characterizing the trajectory of learning across task gaps.

Representational Continuity Across Tasks. In addition to behavioural alignment, effective curricula must ensure representational consistency across tasks to facilitate transfer. Several constraints support this goal:

1. *Feature alignment:* Observation and action spaces should maintain consistent semantics. For example, stable slots for agent-local, opponent-specific, and global inputs to support shared encoders [45, 92, 33].
2. *Predictable dimensional scaling:* As task complexity increases (e.g., more agents or abilities), changes in the input/output space should follow structured patterns to avoid frequent architectural redesign [87, 39, 40].

3. *Scenario modularity*: Tasks should include reusable behavioural components (e.g., navigation, foraging, cooperation) to encourage the emergence of transferable sub-skills [130, 156, 144].

By adhering to these principles, scenario designers can structure task gaps that meaningfully evaluate reuse, retention, and extension. This not only supports rigorous benchmarking of continual MARL methods, but also reveals which forms of generalization are most critical and challenging under dynamic multi-agent settings.

5.3 Offline Pretraining and Online Transfer Scenarios

While curriculum learning emphasizes structured environment interaction, many practical MARL deployments operate under strict limitations on real-time data collection [140, 82, 83]. In such cases, agents must be trained from static datasets, often with no access to online rollouts. This motivates offline [47, 49, 105] and offline-to-online [162, 163, 164] learning paradigms, which offer a complementary lens on adaptability—focusing on generalization under restricted supervision rather than interactive feedback. These settings lie at the intersection of learning adaptability and policy adaptability: the learning process must be robust to fixed and potentially biased datasets, while the resulting policies must transfer to new scenarios without retraining.

Offline Scenario. Offline MARL entails learning policies entirely from pre-collected trajectories, with no further environment interaction during training. This paradigm is particularly relevant for domains where online sampling is impractical due to cost, safety, or logistical constraints. In the absence of interactive feedback, generalization depends heavily on dataset quality and diversity. Desirable properties include:

1. Coverage of diverse coordination behaviours, agent roles, and reward signals;
2. Inclusion of supervision levels ranging from expert to exploratory or random policies;
3. Scenario heterogeneity across different environment configurations and population structures.

The OG-MARL benchmark suite [23] exemplifies this setting, providing stratified datasets across canonical environments such as SMAC [87], SMACv2 [165], MAMuJoCo [92], Flatland [143], RWARE [130], and MPE [33]. In the meantime, many existing datasets are collected using hand-designed or static policies and lack mechanisms to promote behavioural diversity or balanced task coverage [166, 22, 167]. This can lead to under-representation of critical coordination scenarios and hinder robust evaluation.

Offline-to-Online Transfer. The offline-to-online setting bridges the gap between static offline training and fully interactive learning. It evaluates whether policies pretrained on fixed datasets—or supplemented with additional offline data—can accelerate learning in the online version of the same or similar tasks [162, 163, 164]. This setting centres on continued policy optimization using a combination of offline and online experience.

In contrast to curriculum learning, where task complexity typically increases in a structured progression (e.g., $3 \rightarrow 5 \rightarrow 8$ agents), offline-to-online transfer can involve transitions across tasks that are unordered or non-monotonic in complexity. For example, an agent trained offline on SMAC tasks such as 3m and 8m may be evaluated online on 5m, requiring compositional generalization rather than stepwise skill accumulation. This makes offline-to-online transfer a more flexible—yet also more challenging—test of policy adaptability. From the perspective of task gaps, offline tasks should be designed to complement each other in ways that support generalization. Specifically, knowledge acquired from offline behavioural datasets A and B should be composable or synergistic, enabling

more effective adaptation when fine-tuning on a distinct but related online task C. To design a good offline-to-online MARL setting, the following principles are essential:

1. *Complementary offline tasks*: Offline datasets should span distinct but related tasks to support compositional generalization during online adaptation.
2. *Representational consistency*: Observation and action spaces must be aligned across offline and online tasks to enable direct policy reuse without architectural changes.
3. *Behavioural diversity*: Offline data should include a mix of expert and suboptimal behaviours to encourage robust initialization and flexible fine-tuning.

Overall, Offline MARL highlights a unique angle in the adaptability landscape: its success hinges on whether agents can generalize from fixed data distributions, regardless of the environment’s immediate configurability. When equipped with sufficiently rich offline data, a flexible learning algorithm may implicitly absorb a wide spectrum of interaction dynamics and task structures—mirroring the benefits of environmental diversity without explicit exploration. As such, offline MARL represents a promising bridge between sample-efficient learning and policy-level transfer, especially when embedded within diverse, modular benchmarks. However, multi-agent offline-to-online transfer remains underexplored in both algorithm design and system development compared to its single-agent counterpart [162, 163, 164].

5.4 Zero-Shot Coordination Scenarios

Zero-shot coordination (ZSC) scenarios assess an agent’s capacity to collaborate with unfamiliar partners at test time—without prior joint training, shared parameters, or explicit coordination protocols [125, 126, 127, 128]. Rather than optimizing for task performance alone, these scenarios foreground *social adaptability*: the agent’s ability to interpret novel conventions, align with diverse behaviours, and act coherently in uncertain multi-agent settings.

In practical deployments, such as modular robotics, autonomous driving, or simulation-based multi-agent platforms, agents frequently encounter partners developed independently under different assumptions, inductive biases, or design paradigms. This variability challenges agents to operate not within a fixed policy ecosystem, but across a dynamic *social landscape* composed of heterogeneous conventions and strategies.

Empirical studies show that independently trained agents often converge on idiosyncratic or brittle conventions that fail under cross-play [125, 126]. Recent approaches mitigate this by explicitly injecting *partner diversity* during training, encouraging robustness to novel coordination styles via latent belief modelling [127, 128]. These findings motivate the need for benchmarks that expose agents to a wide range of conventions and promote generalization across unseen social configurations.

Effective ZSC benchmarks should therefore be designed with three core criteria:

1. *Multiple viable conventions*: Tasks must permit a spectrum of valid coordination equilibria, avoiding overfitting to a single dominant solution.
2. *Held-out partner diversity*: Evaluation should involve interaction with previously unseen agents exhibiting distinct learning trajectories, inductive priors, or reward shaping.
3. *Implicit interaction mechanisms*: Partial observability and restricted communication ensure that agents must infer partner intent from behaviour, rather than rely on explicit signalling.

Prominent examples include Hanabi [168], which enforces theory-of-mind reasoning [169] under strict communication limits, and Overcooked-AI [129], where agents must disambiguate spatial-temporal plans in environments with multiple emergent conventions. These domains illustrate how structured randomness in the social landscape can reveal coordination failures or successes under zero-shot constraints.

6 Future Directions

We conclude by outlining open challenges and research opportunities for improving adaptability in MARL. To reflect the structure of this survey, we organize this discussion along three axes corresponding to learning adaptability, policy adaptability, and scenario-driven adaptability.

On the learning adaptability, a key challenge lies in balancing scalability with coordination fidelity. Current methods such as value decomposition and centralized critics offer effective coordination but scale poorly to large or asynchronous systems due to communication overhead and growing input dimensionality. Future work may explore partially coupled architectures that enable localized coordination within agent subsets, while maintaining global coherence through compositional critics, graph-based learning, or dynamic communication structures. Additionally, real-world systems often involve transitions between cooperative, competitive, or mixed-motive dynamics. This calls for adaptable objective representations, reward disentanglement, and regularization techniques that support generalization across task interaction types. Supporting distributed and asynchronous execution remains another open problem. Hybrid approaches that distil centralized critics into decentralized policies, or apply minimal synchronization protocols in networked settings, present promising directions.

At the policy level, adaptability requires architectures that generalize across tasks, roles, and teammates. Transformer-based models and permutation-invariant networks provide useful inductive biases but often lack mechanisms for semantic modularity or task-aware decomposition. Future designs should incorporate subtask abstractions, interpretable attention patterns, and modular heads aligned with agent roles or latent strategies. Offline pretraining can further support policy transfer, yet current approaches are constrained by dataset diversity and limited reuse. Developing shared policy libraries through distillation, policy interpolation, or embedding-based indexing may help enable fast adaptation and in-context learning. Lifelong learning also remains underexplored. More flexible adaptation strategies such as online embedding updates, latent-conditioned policy generation, and regret-aware updates could improve robustness across evolving task sequences. Zero-shot coordination with unfamiliar partners introduces additional challenges, requiring mechanisms for belief modelling, convention inference, and partner profiling to align behaviour under uncertainty.

Advancing these capabilities depends on the availability of structured and diagnostic evaluation environments. Many existing benchmarks consist of static tasks with limited configurability, offering coarse-grained assessments of generalization. To enable fine-grained analysis, future environments should support parametrized variations in agent count, role heterogeneity, reward structure, and communication topology. Such continuous scenario spaces would facilitate the study of curriculum learning, continual adaptation, and transfer across structured complexity gradients. There is also a pressing need for standardized offline MARL datasets that span diverse tasks and agent types, supporting reproducible research on offline-to-online transfer and pretraining. To evaluate social adaptability, benchmarks should include multiple viable conventions, diverse agent roles, and limited observability or communication. Evaluation protocols should prioritize cross-play with independently trained agents using different architectures or learning objectives. Finally, diagnostic toolkits such as transferability matrices, skill composition graphs, and failure mode taxonomies are essential for understanding where and how current algorithms succeed or fail.

To address these gaps, we advocate for future research on more adaptable learning paradigms, more generalizable policy transfer mechanisms, and more realistic and structured testing environments. By aligning algorithmic development with flexible training regimes and diagnostic benchmarks, MARL can move toward more robust, general-purpose multi-agent systems capable of operating in highly dynamic real-world settings.

7 Conclusion

This survey has presented a structured analysis of adaptability in multi-agent reinforcement learning, organized along three core axes: learning adaptability, policy adaptability, and scenario-driven adaptability. We examined how existing methods address population scaling, task variation, and execution constraints during training; how policies generalize across tasks, roles, and partners; and how environment design supports meaningful evaluation of adaptability. Despite recent progress, our review highlights significant gaps in both algorithmic approaches and benchmark support. Key challenges in algorithm design include enabling transfer across tasks, sustaining coordination under distributional shifts, and generalizing to unseen agents. Scenario-driven adaptability remains particularly underdeveloped, with a lack of system-level support for evaluating continual learning, offline-to-online transfer, and zero-shot coordination. In the future, we encourage the development of scalable training paradigms, modular and transferable policy architectures, and structured evaluation environments that reflect the complexities of real-world multi-agent systems. Addressing these challenges is critical for advancing MARL from controlled benchmarks toward practical, adaptable deployment.

References

- [1] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multi-agent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.
- [2] Lei Yuan, Ziqian Zhang, Lihe Li, Cong Guan, and Yang Yu. A survey of progress on cooperative multi-agent reinforcement learning in open environment. *arXiv preprint arXiv:2312.01058*, 2023.
- [3] Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):4, 2024.
- [4] Amal Feriani and Ekram Hossain. Single and multi-agent deep reinforcement learning for ai-enabled wireless networks: A tutorial. *IEEE Communications Surveys & Tutorials*, 23(2):1226–1252, 2021.
- [5] James Orr and Ayan Dutta. Multi-agent deep reinforcement learning for multi-robot applications: A survey. *Sensors*, 23(7):3625, 2023.
- [6] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.
- [7] Kai Cui, Anam Tahir, Gizem Ekinici, Ahmed Elshamashory, Yannick Eich, Mengguang Li, and Heinz Koepl. A survey on large-population systems and scalable multi-agent reinforcement learning. *arXiv preprint arXiv:2209.03859*, 2022.
- [8] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.
- [9] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

- [10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [11] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.
- [12] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the international conference on machine learning*, pages 330–337, 1993.
- [13] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pages 5571–5580. PMLR, 2018.
- [14] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 4295–4304, 2018.
- [15] Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Partially observable mean field reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 537–545, 2021.
- [16] Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. Rode: Learning roles to decompose multi-agent tasks. In *International Conference on Learning Representations*, 2021.
- [17] Jakub Grudzien Kuba, Xidong Feng, Shiyao Ding, Hao Dong, Jun Wang, and Yaodong Yang. Heterogeneous-agent mirror learning: A continuum of solutions to cooperative marl. *arXiv preprint arXiv:2208.01682*, 2022.
- [18] Chao Yu, Akash Velu, Eugene Vinitzky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, pages 24611–24624, 2022.
- [19] JG Kuba, R Chen, M Wen, Y Wen, F Sun, J Wang, and Y Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *ICLR 2022-10th International Conference on Learning Representations*, page 1046. The International Conference on Learning Representations (ICLR), 2022.
- [20] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in neural information processing systems*, 32, 2019.
- [21] Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:10299–10312, 2021.
- [22] Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *International conference on machine learning*, pages 17221–17237. PMLR, 2022.

- [23] Claude Formanek, Asad Jeewa, Jonathan Shock, and Arnu Pretorius. Off-the-grid marl: Datasets and baselines for offline multi-agent reinforcement learning. In *Extended Abstract at the 2023 International Conference on Autonomous Agents and Multiagent Systems*. AAMAS, 2023.
- [24] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Networked multi-agent reinforcement learning in continuous spaces. In *2018 IEEE conference on decision and control (CDC)*, pages 2771–2776. IEEE, 2018.
- [25] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International conference on machine learning*, pages 5872–5881. PMLR, 2018.
- [26] Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph convolutional reinforcement learning. *arXiv preprint arXiv:1810.09202*, 2018.
- [27] Daniël Willemsen, Mario Coppola, and Guido CHE de Croon. Mambpo: Sample-efficient multi-robot reinforcement learning using learned world models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5635–5640. IEEE, 2021.
- [28] Zhiwei Xu, Bin Zhang, Yuan Zhan, Yunpeng Baiia, Guoliang Fan, et al. Mingling foresight with imagination: Model-based cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35:11327–11340, 2022.
- [29] Vladimir Egorov and Aleksei Shpilman. Scalable multi-agent model-based reinforcement learning. *arXiv preprint arXiv:2205.15023*, 2022.
- [30] Wenbo Zhang, Osbert Bastani, and Vijay Kumar. Mamps: Safe multi-agent reinforcement learning via model predictive shielding. *arXiv preprint arXiv:1910.12639*, 2019.
- [31] Chenyi Liu, Nan Geng, Vaneet Aggarwal, Tian Lan, Yuan Yang, and Mingwei Xu. Cmix: Deep multi-agent reinforcement learning with peak and average constraints. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, pages 157–173. Springer, 2021.
- [32] Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319:103905, 2023.
- [33] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.
- [34] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2974–2982, 2018.
- [35] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 2085–2087, 2018.

- [36] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. {QPLEX}: Duplex dueling multi-agent q-learning. In *International Conference on Learning Representations*, 2021.
- [37] Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, QIANG FU, Xiaojun Chang, and Yaodong Yang. Maximum entropy heterogeneous-agent reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.
- [38] Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32):1–67, 2024.
- [39] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3461–3475, 2022.
- [40] Ming Zhou, Jun Luo, Julian Vilella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadarar, Zheng Chen, Aurora Chongxi Huang, Ying Wen, Kimia Hassanzadeh, Daniel Graves, Dong Chen, Zhengbang Zhu, Nhat Nguyen, Mohamed Elsayed, Kun Shao, Sanjeevan Ahilan, Baokuan Zhang, Jiannan Wu, Zhengang Fu, Kasra Rezaee, Peyman Yadmellat, Mohsen Rohani, Nicolas Perez Nieves, Yihan Ni, Seyedershad Banijamali, Alexander Cowen Rivers, Zheng Tian, Daniel Palenicek, Haitham bou Ammar, Hongbo Zhang, Wulong Liu, Jianye Hao, and Jun Wang. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving, 11 2020.
- [41] Zhaoyue Xia, Jun Du, Jingjing Wang, Chunxiao Jiang, Yong Ren, Gang Li, and Zhu Han. Multi-agent reinforcement learning aided intelligent uav swarm for target tracking. *IEEE Transactions on Vehicular Technology*, 71(1):931–945, 2021.
- [42] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187, 1999.
- [43] Alain Bensoussan, Jens Frehse, Phillip Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.
- [44] Masanao Aoki. Optimal control of partially observable markovian systems. *Journal of The Franklin Institute*, 280(5):367–386, 1965.
- [45] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.
- [46] Eric Steinberger. Pokerrl. <https://github.com/TinkeringCode/PokerRL>, 2019.
- [47] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.
- [48] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

- [49] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.
- [50] Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.
- [51] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.
- [52] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviyshuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.
- [53] Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz De Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.
- [54] Georgios Papoudakis, Filippas Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2022.
- [55] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International conference on machine learning*, pages 5887–5896. PMLR, 2019.
- [56] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:10199–10210, 2020.
- [57] Washim Uddin Mondal, Mridul Agarwal, Vaneet Aggarwal, and Satish V Ukkusuri. On the approximation of cooperative heterogeneous multi-agent reinforcement learning (marl) using mean field control (mfc). *Journal of Machine Learning Research*, 23(129):1–46, 2022.
- [58] Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E Taylor, and Nidhi Hegde. Multi type mean field reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 411–419, 2020.
- [59] Barna Pásztor, Andreas Krause, and Ilija Bogunovic. Efficient model-based multi-agent mean-field reinforcement learning. *Transactions on Machine Learning Research*, 2021.
- [60] Qizhen Zhang, Chris Lu, Animesh Garg, and Jakob Foerster. Centralized model and exploration policy for multi-agent rl. *arXiv preprint arXiv:2107.06434*, 2021.
- [61] Xiaopeng Yu, Jiechuan Jiang, Wanpeng Zhang, Haobin Jiang, and Zongqing Lu. Model-based opponent modeling. *Advances in Neural Information Processing Systems*, 35:28208–28221, 2022.
- [62] Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Başar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8767–8775, 2021.

- [63] Daniel Melcer, Christopher Amato, and Stavros Tripakis. Shield decentralization for safe multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13367–13379, 2022.
- [64] Donghao Ying, Yunkai Zhang, Yuhao Ding, Alec Koppel, and Javad Lavaei. Scalable primal-dual actor-critic method for safe multi-agent rl with general utilities. *Advances in Neural Information Processing Systems*, 36:36524–36539, 2023.
- [65] Carlos Guestrin, Daphne Koller, and Ronald Parr. Multiagent planning with factored MDPs. In *Proc. NIPS*, volume 14, pages 1523–1530. MIT Press, 2001.
- [66] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.
- [67] Carlos Guestrin, Michail Lagoudakis, and Ronald Parr. Coordinated reinforcement learning. In *Proc. ICML*, volume 2, pages 227–234, 2002.
- [68] Jelle R Kok and Nikos Vlassis. Sparse cooperative q-learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 61, 2004.
- [69] Jelle R. Kok and Nikos Vlassis. Collaborative multiagent reinforcement learning by payoff propagation. *J. Mach. Learn. Res.*, 7(65):1789–1828, 2006.
- [70] Ranjit Nair, Pradeep Varakantham, Milind Tambe, and Makoto Yokoo. Networked distributed pomdps: A synthesis of distributed constraint optimization and pomdps. In *AAAI*, volume 5, pages 133–139, 2005.
- [71] Frans A Oliehoek, Shimon Whiteson, Matthijs TJ Spaan, et al. Approximate solutions for factored dec-pomdps with many agents. In *AAMAS*, pages 563–570, 2013.
- [72] Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control*, pages 256–266. PMLR, 2020.
- [73] Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. Multi-agent reinforcement learning in stochastic networked systems. *Advances in neural information processing systems*, 34:7825–7837, 2021.
- [74] Yunpeng Bai, Chen Gong, Bin Zhang, Guoliang Fan, and Xinwen Hou. Value function factorisation with hypergraph convolution for cooperative multi-agent reinforcement learning. *arXiv:2112.06771*, 2021.
- [75] Bin Zhang, Yunpeng Bai, Zhiwei Xu, Dapeng Li, and Guoliang Fan. Efficient cooperation strategy generation in multi-agent video games via hypergraph neural network. *arXiv:2203.03265*, 2022.
- [76] Lianmin Zheng, Jiacheng Yang, Han Cai, Ming Zhou, Weinan Zhang, Jun Wang, and Yong Yu. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [77] Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Partially observable mean field reinforcement learning. In *Proc. AAMAS*, volume 20, pages 537–545, 2021.

- [78] René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *arXiv:1910.12802*, 2019.
- [79] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field multi-agent reinforcement learning: A decentralized network approach. *arXiv:2108.02731*, 2021.
- [80] Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, and Marco Dorigo. Swarm robotics: a review from the swarm engineering perspective. *Swarm Intelligence*, 7:1–41, 2013.
- [81] Joseph Suarez, David Bloomin, Kyoung Whan Choe, Hao Xiang Li, Ryan Sullivan, Nishaanth Kanna, Daniel Scott, Rose Shuman, Herbie Bradley, Louis Castricato, et al. Neural mmo 2.0: A massively multi-task addition to massively multi-agent learning. *Advances in Neural Information Processing Systems*, 36:50094–50104, 2023.
- [82] Mark A Stephenson and Hanspeter Schaub. Bsk-rl: Modular, high-fidelity reinforcement learning environments for spacecraft tasking. In *75th International Astronautical Congress, Milan, Italy, IAF*, 2024.
- [83] Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. *Advances in Neural Information Processing Systems*, 34:3271–3284, 2021.
- [84] Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. Updet: Universal multi-agent rl via policy decoupling with transformers. In *International Conference on Learning Representations*, 2021.
- [85] Shariq Iqbal, Christian A Schroeder De Witt, Bei Peng, Wendelin Böhmer, Shimon Whiteson, and Fei Sha. Randomized entity-wise factorization for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4596–4606. PMLR, 2021.
- [86] HAO Jianye, Xiaotian Hao, Hangyu Mao, Weixun Wang, Yaodong Yang, Dong Li, Yan Zheng, and Zhen Wang. Boosting multiagent reinforcement learning via permutation invariant and permutation equivariant networks. In *The Eleventh International Conference on Learning Representations*, 2022.
- [87] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.
- [88] Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *The world wide web conference*, pages 3620–3624, 2019.
- [89] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4501–4510, 2020.
- [90] Jianhong Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. Shaq: Incorporating shapley value theory into multi-agent q-learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 5941–5954, 2022.

- [91] Yali Du, Bo Liu, Vincent Moens, Ziqi Liu, Zhicheng Ren, Jun Wang, Xu Chen, and Haifeng Zhang. Learning correlated communication topology in multi-agent reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 456–464, 2021.
- [92] Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. In *Advances in Neural Information Processing Systems*, pages 12208–12221, 2021.
- [93] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley q-value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7285–7292, 2020.
- [94] Yugu Li, Zehong Cao, Jianglin Qiao, and Siyi Hu. Nucleolus credit assignment for effective coalitions in multi-agent reinforcement learning. *arXiv preprint arXiv:2503.00372*, 2025.
- [95] René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *The Annals of Applied Probability*, 33(6B):5334–5381, 2023.
- [96] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.
- [97] Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29, 2016.
- [98] Junjie Sheng, Xiangfeng Wang, Bo Jin, Junchi Yan, Wenhao Li, Tsung-Hui Chang, Jun Wang, and Hongyuan Zha. Learning structured communication for multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 36(2):50, 2022.
- [99] Weinan Zhang, Xihuai Wang, Jian Shen, and Ming Zhou. Model-based multi-agent policy optimization with adaptive opponent-wise rollouts. *arXiv preprint arXiv:2105.03363*, 2021.
- [100] Linghui Meng, Muning Wen, Chenyang Le, Xiyun Li, Dengpeng Xing, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, Yaodong Yang, et al. Offline pre-trained multi-agent decision transformer. *Machine Intelligence Research*, 2023.
- [101] Bor-Jiun Lin and Chun-Yi Lee. Hgap: boosting permutation invariant and permutation equivariant in multi-agent reinforcement learning via graph attention network. In *Forty-first International Conference on Machine Learning*, 2024.
- [102] Linghui Meng, Jingqing Ruan, Xuantang Xiong, Xiyun Li, Xi Zhang, Dengpeng Xing, and Bo Xu. M3: Modularization for multi-task and multi-agent offline pre-training. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023.
- [103] Sicong Liu, Yang Shu, Chenjuan Guo, and Bin Yang. Learning generalizable skills from offline multi-task data for multi-agent cooperation. In *International Conference on Learning Representations*, 2025.

- [104] Jie Liu, Yinmin Zhang, Chuming Li, Zhiyuan You, Zhanhui Zhou, Chao Yang, Yaodong Yang, Yu Liu, and Wanli Ouyang. Maskma: Towards zero-shot multi-agent decision making with mask-based collaborative learning. *Transactions on Machine Learning Research*, 2023.
- [105] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.
- [106] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [107] Lukas Schafer, Filippos Christianos, Amos Storkey, and Stefano Albrecht. Learning task embeddings for teamwork adaptation in multi-agent reinforcement learning. In *NeurIPS 2023 Workshop on Generalization in Planning*, 2023.
- [108] Chao Li, Shaokang Dong, Shangdong Yang, Yujing Hu, Tianyu Ding, Wenbin Li, and Yang Gao. Multi-task multi-agent reinforcement learning with interaction and task representations. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [109] Zikang Tian, Ruizhi Chen, Xing Hu, Ling Li, Rui Zhang, Fan Wu, Shaohui Peng, Jiaming Guo, Zidong Du, Qi Guo, et al. Decompose a task into generalizable subtasks in multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 2023.
- [110] Hyungho Na, Kwanghyeon Lee, Sumin Lee, and Il-Chul Moon. Trajectory-class-aware multi-agent reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [111] Yuanheng Zhu, Shangjing Huang, Binbin Zuo, Dongbin Zhao, and Changyin Sun. Multi-task multi-agent reinforcement learning with task-entity transformers and value decomposition training. *IEEE Transactions on Automation Science and Engineering*, 2024.
- [112] Rongjun Qin, Feng Chen, Tonghan Wang, Lei Yuan, Xiaoran Wu, Yipeng Kang, Zongzhang Zhang, Chongjie Zhang, and Yang Yu. Multi-agent policy transfer via task relationship modeling. *Science China Information Sciences*, 2024.
- [113] Yang Yu, Likun Yang, Zhourui Guo, Yongjian Ren, Qiyue Yin, Junge Zhang, and Kaiqi Huang. Relation-aware learning for multi-task multi-agent cooperative games. *IEEE Transactions on Games*, 2024.
- [114] Qian Long, Zihan Zhou, Abhinav Gupta, Fei Fang, Yi Wu, and Xiaolong Wang. Evolutionary population curriculum for scaling multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2020.
- [115] Rundong Wang, Longtao Zheng, Wei Qiu, Bowei He, Bo An, Zinovi Rabinovich, Yujing Hu, Yingfeng Chen, Tangjie Lv, and Changjie Fan. Towards skilled population curriculum for multi-agent reinforcement learning. *arXiv preprint arXiv:2302.03429*, 2023.
- [116] Jiayu Chen, Yuanxin Zhang, Yuanfan Xu, Huimin Ma, Huazhong Yang, Jiaming Song, Yu Wang, and Yi Wu. Variational automatic curriculum learning for sparse-reward cooperative multi-agent problems. *Advances in Neural Information Processing Systems*, 34:9681–9693, 2021.

- [117] Jiayu Chen, Zelai Xu, Yunfei Li, Chao Yu, Jiaming Song, Huazhong Yang, Fei Fang, Yu Wang, and Yi Wu. Accelerate multi-agent reinforcement learning in zero-sum games with subgame curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11320–11328, 2024.
- [118] Jizhou Wu, Jianye Hao, Tianpei Yang, Xiaotian Hao, Yan Zheng, Weixun Wang, and Matthew E Taylor. Portal: Automatic curricula generation for multiagent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15934–15942, 2024.
- [119] Weixun Wang, Tianpei Yang, Yong Liu, Jianye Hao, Xiaotian Hao, Yujing Hu, Yingfeng Chen, Changjie Fan, and Yang Gao. From few to more: Large-scale dynamic multiagent curriculum learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7293–7300, 2020.
- [120] Tianle Zhang, Zhen Liu, Zhiqiang Pu, and Jianqiang Yi. Automatic curriculum learning for large-scale cooperative multiagent systems. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(3):912–930, 2022.
- [121] Lei Yuan, Lihe Li, Ziqian Zhang, Fuxiang Zhang, Cong Guan, and Yang Yu. Multiagent continual coordination via progressive task contextualization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [122] Yang Yu, Qiyue Yin, Junge Zhang, and Kaiqi Huang. Prioritized tasks mining for multi-task cooperative multi-agent reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023.
- [123] Hongda Jia, Yong Zhao, Yuanzhao Zhai, Bo Ding, Huaimin Wang, and Qingtong Wu. Crmrl: Collaborative relationship meta reinforcement learning for effectively adapting to type changes in multi-robotic system. *IEEE Robotics and Automation Letters*, 2022.
- [124] Tianze Zhou, Fubiao Zhang, Kun Shao, Zipeng Dai, Kai Li, Wenhan Huang, Weixun Wang, Bin Wang, Dong Li, Wulong Liu, et al. Cooperative multi-agent transfer learning with coalition pattern decomposition. *IEEE Transactions on Games*, 2023.
- [125] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. “other-play” for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.
- [126] Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *International conference on machine learning*, pages 7204–7213. PMLR, 2021.
- [127] Keane Lucas and Ross E Allen. Any-play: An intrinsic augmentation for zero-shot coordination. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 853–861, 2022.
- [128] Yang Li, Shao Zhang, Jichen Sun, Yali Du, Ying Wen, Xinbing Wang, and Wei Pan. Cooperative open-ended learning framework for zero-shot coordination. In *International Conference on Machine Learning*, pages 20470–20484. PMLR, 2023.

- [129] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.
- [130] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
- [131] Roni Stern, Nathan Sturtevant, Ariel Felner, Sven Koenig, Hang Ma, Thayne Walker, Jiaoyang Li, Dor Atzmon, Liron Cohen, TK Kumar, et al. Multi-agent pathfinding: Definitions, variants, and benchmarks. In *Proceedings of the International Symposium on Combinatorial Search*, 2019.
- [132] Tonghan Wang, Liang Zeng, Weijun Dong, Qianlan Yang, Yang Yu, and Chongjie Zhang. Context-aware sparse deep coordination graphs. In *International Conference on Learning Representations*, 2022.
- [133] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998(746-752):2, 1998.
- [134] Diego Perez-Liebana, Katja Hofmann, Sharada Prasanna Mohanty, Noburu Kuno, Andre Kramer, Sam Devlin, Raluca D Gaina, and Daniel Ionita. The multi-agent reinforcement learning in malm\ " o (marl\ " o) competition. *arXiv preprint arXiv:1901.08129*, 2019.
- [135] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *International conference on learning representations*, 2019.
- [136] Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly decomposable value functions via communication minimization. In *International Conference on Learning Representations*, 2020.
- [137] Yuanpei Chen, Yaodong Yang, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen Marcus McAleer, Hao Dong, and Song-Chun Zhu. Towards human-level bimanual dexterous manipulation with reinforcement learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [138] Daniel Krajzewicz, Georg Hertkorn, Christian Rössel, and Peter Wagner. Sumo (simulation of urban mobility)-an open-source traffic simulation. In *Proceedings of the 4th middle East Symposium on Simulation and Modelling (MESM20002)*, pages 183–187, 2002.
- [139] Avisek Naug, Antonio Guillen-Perez, Ricardo Luna Gutierrez, Vineet Gundecha, Cullen Bash, Sahand Ghorbanpour, Sajad Mousavi, Ashwin Ramesh Babu, Dejan Markovikj, Lekhapriya Dheeraj Kashyap, et al. Sustaindc: Benchmarking for sustainable data center control. *Advances in Neural Information Processing Systems*, 37:100630–100669, 2024.
- [140] Claire Bizon Monroc, Ana Bušić, Donatien Dubuc, and Jiamin Zhu. Wfcr1: A multi-agent reinforcement learning benchmark for wind farm control. *arXiv preprint arXiv:2501.13592*, 2025.

- [141] Reza J Torbati, Shubham Lohiya, Shivika Singh, Meher S Nigam, and Harish Ravichandar. Marbler: An open platform for standardized evaluation of multi-robot reinforcement learning algorithms. In *2023 International Symposium on Multi-Robot and Multi-Agent Systems (MRS)*, pages 57–63. IEEE, 2023.
- [142] Ke Xue, Jiacheng Xu, Lei Yuan, Miqing Li, Chao Qian, Zongzhang Zhang, and Yang Yu. Multi-agent dynamic algorithm configuration. *Advances in Neural Information Processing Systems*, 35:20147–20161, 2022.
- [143] Sharada Mohanty, Erik Nygren, Florian Laurent, Manuel Schneider, Christian Scheller, Nilabha Bhattacharya, Jeremy Watson, Adrian Egli, Christian Eichenberger, Christian Baumberger, et al. Flatland-rl: Multi-agent reinforcement learning on trains. *arXiv preprint arXiv:2012.05893*, 2020.
- [144] Qihan Liu, Yuhua Jiang, and Xiaoteng Ma. Light aircraft game: A lightweight, scalable, gym-wrapped aircraft competitive environment with baseline reinforcement learning algorithms. <https://github.com/liuqh16/CloseAirCombat>, 2022.
- [145] Xianliang Yang, Zhihao Liu, Wei Jiang, Chuheng Zhang, Li Zhao, Lei Song, and Jiang Bian. A versatile multi-agent reinforcement learning benchmark for inventory management. *arXiv preprint arXiv:2306.07542*, 2023.
- [146] Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation. *arXiv preprint arXiv:2310.08901*, 2023.
- [147] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2023.
- [148] Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, Hao Fu, Fangkun Zhao, Caixia Yuan, and Xiaojie Wang. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. *arXiv preprint arXiv:2502.20073*, 2025.
- [149] Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. *arXiv preprint arXiv:2402.16499*, 2024.
- [150] Wei Wang, Dan Zhang, Tao Feng, Boyan Wang, and Jie Tang. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. *arXiv preprint arXiv:2408.15971*, 2024.
- [151] Ming Zhang, Shenghan Zhang, Zhenjie Yang, Lekai Chen, Jinliang Zheng, Chao Yang, Chuming Li, Hang Zhou, Yazhe Niu, and Yu Liu. Gobigger: A scalable platform for cooperative-competitive multi-agent interactive simulation. In *The Eleventh International Conference on Learning Representations*, 2023.
- [152] Cinjon Resnick, Wes Eldridge, David Ha, Denny Britz, Jakob Foerster, Julian Togelius, Kyunghyun Cho, and Joan Bruna. Pommerman: A multi-agent playground. *arXiv preprint arXiv:1809.07124*, 2018.

- [153] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 66–83. Springer, 2017.
- [154] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *International Conference on Autonomous Agents and Multiagent Systems*, 2017.
- [155] Qingxu Fu, Tenghai Qiu, Jianqiang Yi, Zhiqiang Pu, and Shiguang Wu. Concentration network for reinforcement learning of large-scale multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
- [156] Xuehai Pan, Mickel Liu, Fangwei Zhong, Yaodong Yang, Song-Chun Zhu, and Yizhou Wang. Mate: Benchmarking multi-agent reinforcement learning in distributed target coverage control. *Advances in Neural Information Processing Systems*, 35:27862–27879, 2022.
- [157] Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See Kiong Ng, and Jiashi Feng. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. *arXiv preprint arXiv:2311.08562*, 2023.
- [158] Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. URL <https://arxiv.org/abs/2310.05036>, 2023.
- [159] Yubo Dong, Xukun Zhu, Zhengzhe Pan, Linchao Zhu, and Yi Yang. Villageragent: A graph-based multi-agent framework for coordinating complex task dependencies in minecraft. *arXiv preprint arXiv:2406.05720*, 2024.
- [160] Richard Zhuang, Akshat Gupta, Richard Yang, Aniket Rahane, Zhengyu Li, and Gopala Anumanchipalli. Pokerbench: Training large language models to become professional poker players. *arXiv preprint arXiv:2501.08328*, 2025.
- [161] Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. *arXiv preprint arXiv:2503.01935*, 2025.
- [162] Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.
- [163] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.
- [164] Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.
- [165] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Nicolaus Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

- [166] Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. Offline multi-agent reinforcement learning with implicit global-to-local value regularization. *Advances in Neural Information Processing Systems*, 36:52413–52429, 2023.
- [167] Jianzhun Shao, Yun Qu, Chen Chen, Hongchang Zhang, and Xiangyang Ji. Counterfactual conservative q learning for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:77290–77312, 2023.
- [168] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.
- [169] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.