



From individual decisions to team emergence: a survey on explainable cooperative multi-agent reinforcement learning

Lei Sheng¹ · Xiliang Chen² · Zhiqiang Pan¹ · Fei Cai¹ · Honghui Chen¹

Received: 28 October 2025 / Accepted: 19 May 2026 / Published online: 13 June 2026
© The Author(s) 2026

Abstract

Multi-Agent Reinforcement Learning (MARL) holds significant promise for cooperative decision-making, yet its reliance on deep neural networks (DNNs) creates “black-box” characteristics that impede trustworthy deployment in high-stakes scenarios. This lack of transparency complicates tracing decision logic and raises concerns about safety and accountability. This survey systematically reviews Explainable MARL (XMARL) for cooperative settings, deconstructing the decision-making chain from individual agent policies to collective team behavior. To address the absence of a unified framework, we introduce a novel multi-level taxonomy encompassing microscopic individual behavior, interaction mechanisms, team strategy emergence, and system-level performance. We organize core explanatory questions and technical approaches within this structure, summarize the principles and limitations of representative methods, and critically discuss key challenges such as evaluation standards, causal reasoning integration, and deployment adaptability. Our goal is to provide both theoretical foundation and technical guidance for building transparent, trustworthy, and verifiably cooperative multi-agent systems (MASs).

Keywords Multi-agent reinforcement learning · Explainable artificial intelligence (XAI) · Collaborative autonomous systems · Explainable reinforcement learning (XRL) · Trustworthy AI

✉ Honghui Chen
248856416@qq.com

Lei Sheng
shenglei24@nudt.edu.cn

Xiliang Chen
lgd_chenxiliang@aeu.edu.cn

Zhiqiang Pan
panzhiqiang@nudt.edu.cn

Fei Cai
caifei08@nudt.edu.cn

¹ National Key Laboratory of Information Systems Engineering, National University of Defense Technology, 109 Deya Road, Changsha 410073, Hunan, China

² School of Command and Control Engineering, Army Engineering University, 88 Houbiaoying Road, Nanjing 210007, Jiangsu, China

1 Introduction

Reinforcement Learning (RL), which optimizes decision-making policies through sequential interaction with the environment, has emerged as a core paradigm for addressing control problems in complex dynamic systems (Li et al. 2026; Wang et al. 2023). In recent years, thanks to the powerful representational capacity of DNNs, Deep Reinforcement Learning (DRL) has achieved breakthrough progress in domains such as game playing (Barros e Sá and Madeira 2025) and robotic control (Qiu et al. 2022; Mnih et al. 2015), demonstrating its immense potential in handling high-dimensional state and action spaces. However, the inherent “black-box” nature of DRL models makes their decision logic difficult for humans to comprehend and trust, severely limiting their reliable deployment in safety-critical scenarios such as medical diagnosis (Thakre et al. 2025; Romdhana et al. 2022; Otto et al. 2023). To address this challenge, the field of XRL (or XDRL) has emerged, aiming to enhance the transparency and comprehensibility of RL agents through various methods, thereby building a bridge between model performance and decision trust (Milani et al. 2022a; Liu et al. 2023a).

Nevertheless, many critical tasks in the real world, such as coordinated traffic signal optimization (Zhang et al. 2024b; Ren et al. 2024a, b), unmanned aerial vehicle swarm formation (Kim et al. 2024a, b; Luvisutto et al. 2025), and smart grid management (Wang et al. 2021; Hu et al. 2024a; Jin et al. 2025a), are inherently problems of interaction and cooperation among multiple autonomous agents in a shared environment. Cooperative MARL extends the RL paradigm to collective intelligence scenarios, training a set of agents to learn collaborative policies that maximize team rewards, and has become the framework of choice for solving such problems (Ning and Xie 2024; Jin et al. 2025c; Li et al. 2024a). However, MARL not only inherits the “black-box” complexity of DRL but also introduces intrinsic challenges such as environmental non-stationarity, partial observability, and credit assignment (Ning and Xie 2024; Jin et al. 2025c; Li et al. 2024a). These factors are coupled together, rendering the decision-making process of MARL systems a multi-dimensional “compound black-box”: its opacity stems both from the internal workings of individual neural networks and from the dynamic, high-order interactions among agents. This complexity poses significant challenges for providing meaningful transparency, making the direct application of single-agent explanation methods (Xu et al. 2024; Xiong et al. 2024) often inadequate for elucidating inter-agent coordination, communication semantics, and the emergence of team-level behavior (Domènech I Vila et al. 2022). This signifies a paradigm shift from explaining individual decision rationality to explaining group interactive intelligence, thereby giving rise to the dedicated and emerging field of XMARL (Alahmari 2025; Kazhdan et al. 2020).

Within the broad spectrum of XMARL, explainable cooperative MARL has garnered significant attention due to its application prospects in safety-critical collective systems (e.g., collaborative robots, disaster rescue teams) and presents particularly severe and hierarchical demands for transparency (Kim et al. 2024a; Jin et al. 2025c). The decision-making process in cooperative MARL naturally exhibits a hierarchical structure: agents first make micro-decisions based on local observations (individual level); these decisions are coordinated through communication, attention, or explicit coordination mechanisms (interaction level); the coordinated behaviors of individuals further integrate to give rise to team-level strategies and tactics (team level); finally, the overall system’s performance, robustness,

and adaptability constitute the top level of evaluation (system level). This cognitive chain from micro to macro, from local to global, is key to understanding how “team intelligence” emerges from “individual decisions”. However, current XMARL research is notably fragmented and unfocused: a large body of work consists either of performance-driven auxiliary explanations for specific algorithms (Han et al. 2022b) or tailored solutions for specific tasks (Crespi et al. 2023). The field lacks a unified analytical framework that aligns with this intrinsic decision hierarchy, leading to dispersed research questions, inconsistent evaluation standards, and difficulties in systematic comparison and integration of technologies.

To fill this gap, this paper proposes a novel four-dimensional taxonomy framework for explainability in cooperative MARL, grounded in the inherent “individual → interaction → team → system” decision hierarchy. This framework serves not as a subjective classification but as a formalization of the intrinsic cognitive structure of collaborative agents, aiming to provide a unified analytical lens and discourse system for the field. Under this framework, we conduct a systematic survey, with core contributions as follows:

- **Constructing a Hierarchical Explanation Framework:** We introduce a structured taxonomy that systematically deconstructs the explainability landscape into four interconnected analytical tiers: from *micro-level* (individual behavior and intent) and *interactive* (collaboration mechanisms) to *team-level* (emergent behavior) and *system-level* (global performance and robustness diagnosis).
- **Conducting a Forward-Looking Systematic Review:** Based on this framework, we systematically organize and critically compare key recent research, analyzing core technical pathways and clarifying their principles, boundaries, and limitations to chart the technological landscape.
- **Building a Multi-Dimensional Evaluation Perspective and Identifying Challenges:** We propose a multi-dimensional evaluation system integrating faithfulness, causality, practicality, and efficiency, and systematically elaborate on core methodological challenges and future directions.

The remainder of this survey is structured to guide the reader through a logical progression of topics (see the research roadmap in Fig. 1). To ensure the rigor and reproducibility of this survey, we will first detail the literature collection and analysis methods in Sect. 2. Then, Sect. 3 lays the necessary theoretical foundations. Sect 4, as the core, systematically reviews XMARL techniques according to the four-level taxonomy. Sect 5 discusses the evaluation system and open challenges. Finally, Sect. 6 concludes the paper and outlines future research directions. This survey aims to provide a solid theoretical foundation and a clear technical blueprint for building transparent, trustworthy, and collaborative MASs.

2 Methodology: a systematic approach to surveying the literature

To ensure the systematicity, transparency, and reproducibility of this survey, we adhered to the established principles of systematic literature reviews. This section delineates our comprehensive methodology, encompassing the literature search strategy, explicit inclusion/exclusion criteria, and the structured analytical framework employed for synthesis. Our

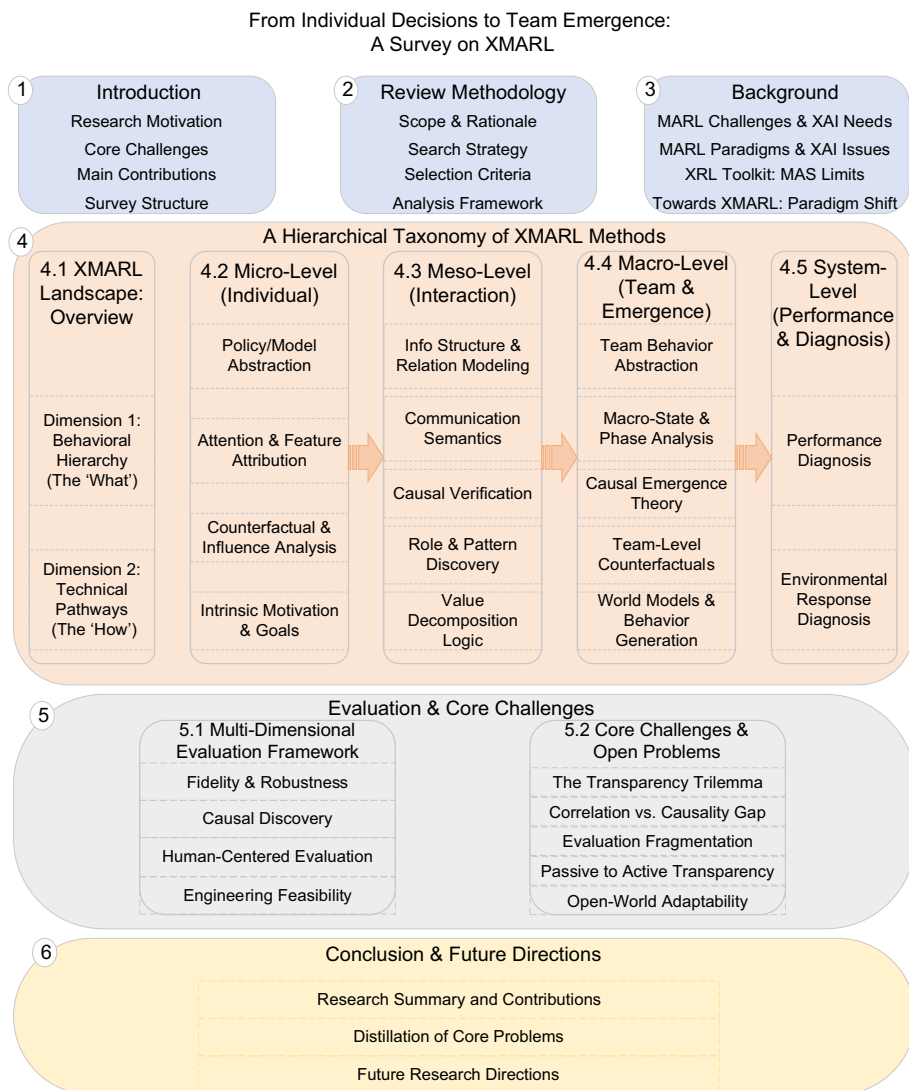


Fig. 1 A research roadmap for this survey

objective is to present a complete and auditable decision chain from literature identification to knowledge synthesis.

2.1 Temporal scope and rationale

This survey is dedicated to a systematic review and critical analysis of the latest advancements in XMARL for cooperative settings. The core temporal boundary for our systematic search spans from January 2022 to October 2025. This four-year window is deliberately

chosen and justified by several converging academic and technological trends that define the current epoch of XMARL research, rather than being arbitrary.

First, it reflects the period of consolidated academic focus and community formation. Since approximately 2022, dedicated workshops on themes such as “Trustworthy MASSs” and “Human-AI Collaboration with Explainability” have become staple fixtures at premier AI conferences (e.g., NeurIPS, ICML, ICLR, AAMAS). This institutionalization signifies a broad consensus within the research community that explainability is a *core bottleneck* hindering the transition of cooperative MARL from theoretical research to real-world, safety-critical deployment. Consequently, the volume and depth of dedicated XMARL publications have seen marked growth from this period onward.

Second, it captures the critical phase of methodological specialization. The last three to four years represent a pivotal shift where XMARL has evolved from primarily adapting general-purpose XAI tools (e.g., post-hoc attribution methods) towards developing its own native, specialized technical repertoire. This period has witnessed the focused emergence of techniques explicitly designed to address the unique core challenges of multi-agent settings: interpretable credit assignment in value decomposition, semantic parsing of emergent communication protocols, causal analysis of inter-agent influence, and the validation of explanation fidelity for collaborative emergent behaviors. Simultaneously, the active integration of XMARL with adjacent fields like causal inference, symbolic reasoning, and foundation models has become a dominant research trajectory, a trend best observed within this recent window.

Third, it addresses a specific gap in the existing survey landscape. While high-quality surveys exist for eXplainable AI (XAI) and single-agent Explainable RL (XRL), comprehensive treatments that systematically cover *cooperative* XMARL are notably absent. Often, MARL is addressed only peripherally in a single section of broader XRL reviews (Cao et al. 2024). Therefore, a dedicated survey focusing on post-2021 literature is necessary to fill this niche, providing readers with a panoramic and current view of the field’s technological frontier.

Fourth, it balances topical timeliness with theoretical coherence. To offer readers insights into the most current research dynamics, our analytical emphasis is placed on works published from 2022 onward in top-tier venues (the aforementioned conferences and journals like *IEEE Transactions on Neural Networks and Learning Systems*). To maintain a solid theoretical foundation, seminal and foundational works pre-dating 2022 are strategically incorporated and discussed within the integrated background presented in Sect. 3, ensuring a continuous and complete intellectual lineage.

2.2 Search databases and keyword strategy

To construct a robust initial literature corpus, we executed systematic searches across five major interdisciplinary databases: Web of Science, IEEE Xplore, ACM Digital Library, Scopus, and the arXiv preprint server. Our search strategy was designed to prioritize high recall, minimizing the risk of omitting relevant studies in this interdisciplinary domain.

The core Boolean search query was constructed using the field’s most prevalent terminology:

```
("multi-agent reinforcement learning" OR
"multiagent reinforcement learning" OR "MARL") AND
```

("explainable" OR "interpretable" OR "interpretability" OR "explainability")

This query was intentionally kept broad and did not include terms specifying particular scenarios (e.g., “cooperative”) or technical approaches (e.g., attention, causal). This design choice prevented the premature exclusion of potentially relevant papers that might use different nomenclature and avoided introducing excessive noise. The necessary focus on cooperative scenarios and specific technical pathways was rigorously enforced during the subsequent manual screening and content-coding phases (Sects. 2.3 and 2.5).

To further enhance comprehensiveness, we conducted backward and forward snowballing on the reference lists and citation networks of studies that progressed to the full-text assessment stage. Rather than serving as an independent primary retrieval channel, this procedure was used as a supplementary completeness check to verify that the database-based search had not overlooked highly relevant studies because of terminological heterogeneity. No additional studies satisfying all eligibility criteria were identified through snowballing. Accordingly, the quantitative selection flow reported in Table 1 is based solely on records retrieved from the five target databases.

2.3 Inclusion and exclusion criteria

To maintain a sharp focus on the survey’s central research question—*elucidating the decision-making transparency of cooperative MARL systems*—we defined and applied a strict set of inclusion and exclusion criteria during a multi-stage screening process (title/abstract, followed by full-text review).

Inclusion Criteria (all of the following dimensions were considered):

- IC1. **Research Topic:** The work must explicitly investigate the explainability or interpretability of a MARL model or system. Its contribution should manifest in at least one of the following ways: (a) proposing a novel explanation method (post-hoc or intrinsic); (b) designing a novel interpretable model architecture for MARL; (c) establishing a new framework or metric for evaluating explanations in multi-agent contexts; or (d) providing a profound analytical insight, critique, or survey of the field.
- IC2. **Scenario Focus:** Primary consideration was given to studies situated in fully cooperative MARL environments or mixed-motive settings where cooperative dynamics are central. This ensures high thematic relevance to “cooperative” XMARL.
- IC3. **Contribution Type:** We prioritized works that present clear methodological, theoretical, or empirical innovations related to explainability, rather than mere applications of existing methods without novel insight.
- IC4. **Publication Venue and Preprint Quality Assurance:** We included peer-reviewed journal articles, full conference papers from major venues, and preprints (primarily from arXiv) that meet a clearly defined threshold of scholarly credibility. Specifically, an arXiv preprint was deemed eligible for inclusion only if it satisfied **at least three** of the following objective indicators of “significant traction” and rigor:
 - **Formal Acceptance:** The work has been explicitly noted as accepted to a peer-reviewed conference or journal (e.g., *COLM 2025*, *AAAI XAI4DRL Workshop*), indicating community validation.

- **Digital Object Identifier (DOI) or E-ISSN:** The preprint possesses an active DOI (e.g., 10.48550/arXiv.xxxx) or an E-ISSN, signifying formal indexing in academic databases such as Engineering Village, Scopus, or IEEE Xplore.
- **Extensive Empirical Validation:** The manuscript exceeds 10 pages in length (excluding references and appendices) and contains comprehensive experimental sections with detailed analysis, rather than brief workshop abstracts or position papers.
- **Reproducibility Artifacts:** The authors provide a publicly accessible code repository (e.g., GitHub) or detailed supplementary materials that enable independent verification of the claimed results.
- **Institutional Affiliation:** The preprint is authored by researchers with verifiable academic or industrial affiliations, distinguishing it from anonymous or pseudonymous submissions lacking scholarly accountability.

This multi-criteria filter ensures that the 155 core studies comprising our analytical corpus—including 21 carefully vetted arXiv preprints—represent contributions of demonstrable quality and community impact, thereby safeguarding the survey’s integrity.

Exclusion Criteria (studies meeting any of the following were excluded):

- EC1. Works focused exclusively on fully competitive or adversarial MARL settings with no cooperative elements, as their objectives and challenges diverge significantly from our focus.
- EC2. Papers that utilize components with potential explanatory utility (e.g., attention mechanisms, prototype layers) purely as tools for enhancing task performance, without any accompanying analysis, visualization, or evaluation of their role in providing decision attribution or interpretability.
- EC3. Short papers, workshop abstracts, poster summaries, or any publication that does not provide sufficient methodological detail for critical assessment and synthesis.
- EC4. Literature for which the full text was inaccessible through our institutional subscriptions or reasonable academic channels.

2.4 Literature selection flow and temporal trend

To ensure the reproducibility of this systematic survey, we report the complete flow of records through the review process in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. This subsection provides a quantitative, stage-by-stage account of the literature selection procedure, from the initial database search to the final corpus of included studies, and supplements it with descriptive statistics that ground our high-level claims in empirical evidence.

Snowballing was additionally performed as a supplementary validation procedure and did not produce further eligible records; therefore, it is not represented as a separate identification branch in the selection counts. The selection process unfolded in five sequential stages, as detailed in Table 1:

Table 1 Quantitative summary of the systematic literature selection process. (Note: Snowballing was conducted as a supplementary completeness check and did not contribute additional eligible records; therefore, the counts reported in this table reflect only records retrieved from the five target databases.)

Stage	Description	Count	Removed/excluded
Identification	Records retrieved from database searches	1,247	–
Deduplication	Records after removing duplicates	891	356
Screening	Records retained after title/abstract screening	317	574
Full-text retrieval	Full-text articles sought for retrieval	317	–
– Not retrieved	Full text unavailable (EC4)	–	11
Eligibility assessment	Full-text articles assessed for eligibility	306	151
Final inclusion	Studies included in systematic synthesis	155	–

Identification:

The systematic query executed across the five target databases (Web of Science, IEEE Xplore, ACM Digital Library, Scopus, and arXiv) returned an initial pool of **1247** records.

Deduplication:

Duplicate entries arising from overlapping database coverage were removed using reference management software and verified manually. This yielded **891** unique records.

Screening (Title and Abstract):

The unique records were screened based on their titles and abstracts against the inclusion and exclusion criteria defined in Sect. 2.3. A total of **574** records were deemed irrelevant and excluded. The remaining **317** records advanced to full-text retrieval.

Full-Text Eligibility Assessment:

We sought to obtain the full text for all 317 records. Full texts were unavailable for **11** records (EC4). Consequently, **306** full-text articles were assessed in detail. Of these, **151** articles were excluded for the following primary reasons: (i) use of interpretable components exclusively for performance enhancement (EC2); (ii) insufficient methodological depth (EC3); or (iii) focus on fully competitive settings (EC1).

Final Inclusion:

A total of **155** studies satisfied all eligibility criteria and constitute the core corpus for the qualitative and thematic synthesis presented in Sects. 4 and 5.

This transparent quantification of the selection cascade not only enhances the methodological credibility of the survey but also provides readers with a clear understanding of

the scale and rigor of the evidence synthesis process. The subsequent subsection details the analytical framework applied to this final corpus of 155 studies.

2.4.1 Temporal trend of the final corpus

To ground the survey's high-level observations in empirical evidence, we first examine the temporal distribution of the included studies. Figure 2 plots the number of published papers per year from 2022 to 2025. The corpus shows a clear upward trend: 26 (2022), 23 (2023), 37 (2024), and 69 (2025, including early online preprints). This near-tripling over three years quantitatively confirms that XMARL has rapidly intensified as a specialized research area.

2.5 Analysis and synthesis framework

The final set of 155 studies meeting the above criteria formed the corpus for our in-depth analysis. We employed a structured, iterative qualitative content analysis method. The cornerstone of this analysis was the four-dimensional taxonomy framework introduced in Sect. 4 (Individual → Interaction → Team → System), which served as the guiding analytical lens. This process was designed to extract patterns, compare methodologies, and synthesize higher-order insights through the following steps:

2.5.1 Structured coding and categorization

The primary coding principle was to map each paper onto the intrinsic hierarchical decision structure of cooperative MARL that our taxonomy formalizes. This was not an exogenous imposition but a means to systematically chart the explanatory landscape across the complete chain from individual actions to system-level emergence. The coding process proceeded in two dimensions, with specific provisions for handling the interconnected nature of multi-agent explanations.

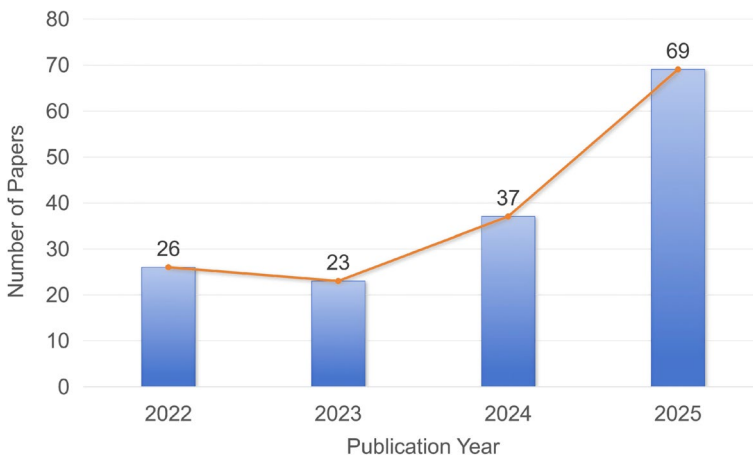


Fig. 2 Distribution of the 155 included papers by publication year

- **First-Level Coding (Primary Explanatory Focus):** Each paper's core explanatory objective was analyzed and assigned to the most appropriate of the four hierarchical levels (Micro, Meso, Macro, or System). *Handling Cross-Level Contributions.* Given the inherent interconnectedness of multi-agent systems, a number of studies necessarily span multiple behavioral hierarchies. To ensure analytical clarity and avoid inconsistent classification, we adhered to the following protocol:

(i) The Unexplained Modality Criterion. Classification was determined by identifying *the highest level of abstraction at which the explanation is generated or targeted*. For instance, if a method visualizes attention weights to reveal inter-agent focus, it is classified under **Meso-Level (Interaction)**. Although this visualization pertains to individual policies (Micro), the *explanatory content* intrinsically concerns interaction dynamics. Conversely, a method using Value Decomposition to assign individual credit is classified under **Meso-Level (Collaboration Logic)** because its *explanatory mechanism* requires analyzing the team's joint value function, a Meso-level construct, even though the output is a Micro-level score.

(ii) The Distinctive Multi-Agent Challenge Criterion. If a paper addresses a challenge unique to multi-agent settings (e.g., non-stationarity, emergent communication, inter-agent credit assignment), it is prioritized for Meso or Macro levels over the Micro level. An explanation of an individual's action *as a response to teammates* (Meso) is treated as distinct from an explanation of isolated behavior (Micro).

(iii) Explicit Cross-Referencing. For papers with substantial contributions at adjacent levels, we anchor the paper at its primary level and utilize explicit cross-references in the synthesis (e.g., "*As noted in the discussion of Value Decomposition (Sec. 4.3.5), this method also illuminates individual contributions...*"). This maintains the taxonomic integrity of Sect. 4 without fragmenting the holistic view of the work.

This protocol ensures that categorization is driven by the *semantic locus* of the explanation rather than merely the technical component used. A detailed operationalization of these boundary criteria, including a tabular summary aligned with the subsequent sections, is provided in Sect. 4.1.

- **Second-Level Coding (Technical Pathway):** Within each level, papers were further clustered based on methodological kinship. This resulted in identifiable technical pathways such as: *Attention-Based Feature Attribution*, *Causal and Counterfactual Analysis*, *Value Decomposition and Credit Assignment Interpretation*, *Policy/Behavior Abstraction*, and *Structured Communication and Relational Modeling*. This two-dimensional (Level $\times$ Pathway) coding matrix structured our knowledge base.

2.5.2 Taxonomic distribution of the corpus

Applying the above first-level coding scheme to the 155 included studies yields the taxonomic distribution displayed in Fig. 3. The distribution is highly non-uniform: the Meso level dominates with 99 papers (63.9%), followed by Micro (32, 20.6%), Macro (17, 11.0%), and System (7, 4.5%). This skewed pattern demonstrates that current XMARL

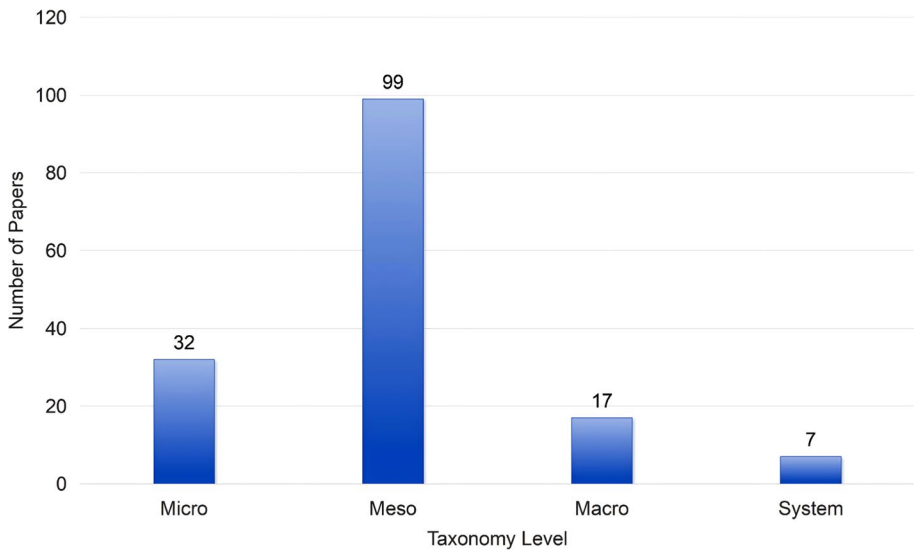


Fig. 3 Distribution of the 155 included papers across the four taxonomy levels (Micro, Meso, Macro, System)

research is heavily concentrated on explaining pairwise or group interactions, while strategic and system-level explanations remain nascent. Such quantitative evidence directly supports two central claims of this survey: (i) the field has become highly specialized around a few dominant explanatory modalities, and (ii) the proposed four-level taxonomy captures genuine structural divisions in the literature, thereby substantiating its novelty relative to prior, unstructured overviews.

2.5.3 Comparative analysis and critical synthesis

Following coding, we conducted detailed comparative analyses within each “Level-Pathway” cluster. We systematically contrasted studies across several key dimensions:

- **Core Mechanism:** The fundamental algorithm or process for generating explanations (e.g., input perturbation sensitivity analysis, learning a causal graph, decomposing a team value function).
- **Underlying Assumptions:** Critical prerequisites for the method’s application (e.g., differentiable agent policies, access to an environment simulator, availability of partial structural or causal priors).
- **Explanation Output Form:** The modality and presentation of the explanation (e.g., saliency heatmaps over observations/actions, natural language descriptions, contribution scores per agent, directed causal graphs).
- **Validation Paradigm:** The experimental testbeds used (e.g., StarCraft II Micromanagement, Multi-Agent Particle Environment, autonomous driving or traffic simulators) and the specific metrics employed for evaluation (e.g., task performance post-explanation, human participant accuracy in understanding, computational faithfulness metrics like infidelity or sensitivity).

This granular comparison allowed us to distill the distinctive advantages, inherent limitations, and domain-specific applicability of each technical pathway, including trade-offs related to explanation fidelity, computational overhead, generalizability, and dependency on prior knowledge.

2.5.4 Meta-synthesis of challenges and future directions

Finally, we conducted a cross-cluster meta-synthesis of the coded literature. By identifying recurrent methodological limitations, unresolved theoretical tensions, and convergent technological trends across the coded corpus, we derived the core challenges of the field. These higher-order themes—such as the trade-off among *explanation faithfulness*, *human interpretability*, and *generalizability*, as well as the challenge of relating *statistical regularities* in explanations to underlying *causal mechanisms*—directly inform the structured discussion of open challenges in Sect. 5 and the outlook on future research directions in Sect. 6.最后，我们进行了编码文献的跨群元综合。通过识别反复出现的方法论局限性、未解决的理论张力以及整个编码语料库中的趋同技术趋势，我们推导出了该领域核心、根本且驱动挑战的核心。这些高阶主题——如解释忠实性、人类可解释性与泛化性的三方权衡、可理解的简洁性与广泛泛化性的权衡，以及——即解释中统计规律性与因果机制之间桥接的挑战——直接影响了第5节开放挑战的结构化讨论以及第6节未来研究方向的展望。

This transparent, multi-layered analytical process helps ensure that the insights and conclusions presented in this survey are traceable, well-grounded, and based on a comprehensive and critical examination of the existing literature. In this way, the survey provides a structured reference point for future research on transparent cooperative MAS.

3 Background

This chapter establishes the theoretical foundation for the systematic review of XMARL. Section 3.1 begins with the standard formal model of cooperative MARL, provides a systematic analysis of its inherent core challenges, and articulates how these challenges necessitate a hierarchical approach to explainability. Section 3.2 then surveys the primary research paradigms developed to address these challenges, examining the distinct explainability issues inherent to each. The central argument of this chapter is that the intrinsic complexity and diversity of cooperative MARL logically motivate the “Micro–Meso–Macro–System” four-layer explanatory framework proposed in this survey.

3.1 Cooperative multi-agent reinforcement learning: model, challenges, and explainability drivers

Cooperative MARL focuses on training a team of autonomous agents to learn collaborative policies that maximize the team’s long-term cumulative return. The standard formal model for this problem is the Decentralized Partially Observable Markov Decision Process (Dec-POMDP) (Oliehoek et al. 2016), defined by the tuple $\langle N, S, A, P, R, \Omega, O, \gamma \rangle$. In this model, the policy of agent i , denoted $\pi_i(a_i | \tau_i)$, conditions on its action-observation his-

tory τ_i . The collection of all agent policies forms the joint policy π , and the objective is to maximize the expected cumulative discounted reward: $\mathbb{E}_{\pi}[\sum_t \gamma^t R(s_t, a_t)]$.

3.1.1 Systematic analysis of core challenges

The Dec-POMDP framework intrinsically embodies several fundamental challenges for cooperative MARL. These challenges not only define key difficulties in algorithm design but also fundamentally shape the scope and requisite granularity of explainability research.

Environmental non-stationarity and policy conflict

In single-agent RL, the environment dynamics $P(s'|s, a)$ are stationary. In contrast, from the perspective of any agent i in a MARL setting, the perceived state transition $P(s'|s, a_i, a_{-i})$ depends on both the environment's physical dynamics and the continually evolving policies π_{-i} of other agents. **This non-stationarity violates the Markov assumption and undermines the theoretical foundations of Bellman optimality principles central to classical RL** (Ning and Xie 2024). A direct consequence is that if agents learn independently, experience replay buffers contain data generated under different policy versions, leading to biased value function estimates and severe training instability (Papoudakis et al. 2019). Externally, this manifests as policy or action conflict: when agents optimize based on local information and individual objectives, their decisions may interfere or cancel each other out due to misaligned goals, resource competition, or faulty predictions of teammates' actions. Explaining such systems therefore requires methods that go beyond tracing individual decisions to uncover the root causes of mutual influence and coordination failure.

Partial observability and the resulting challenges of representation and coordination

A core tenet of the Dec-POMDP is that each agent i receives only a local observation o_i correlated with the global state s . While this partial observability is crucial for connecting MARL to real-world applications, it introduces significant difficulties: agents lack direct access to the complete environmental state, the precise intentions of teammates, and their actions.

To cope, agents must develop the capacity to extract decision-relevant information from their observation history. Theoretically, this can involve maintaining and inferring a belief state $b_i(s|\tau_i)$, i.e., a posterior distribution over the hidden state. In practice, most deep MARL algorithms eschew explicit probabilistic inference. Instead, an agent's policy network $\pi_i(a_i|\tau_i)$ or value network $Q_i(\tau_i, a_i)$ implicitly learns to integrate historical information, effectively functioning as a black-box mapping from history to action or value (Sunehag et al. 2017). Understanding what features an agent extracts from its history or what internal world model it forms remains a significant challenge.

Partial observability and non-stationarity are deeply intertwined and mutually exacerbating. An agent's perception of "environmental" dynamics—largely dictated by other agents' behaviors—is filtered through incomplete local observations. Consequently, its implicit estimates of others' policies π_{-i} are inherently error-prone. These estimation errors lead to inaccurate predictions of environment dynamics, increasing learning instability and potentially triggering conflicts rooted in misunderstanding (Daskalakis et al. 2023). Clarifying the "observation–representation–decision" chain is thus paramount, driving the need for explainability techniques that can peer into agents' internal information processing.

Reward sparsity, delayed feedback, and their intrinsic link to credit assignment

In complex sequential cooperative tasks, semantically meaningful team rewards (e.g., “task success”) are often sparse, appearing only at episode termination. Furthermore, the causal impact of an individual action on the final team outcome may be subject to long temporal delays (Itaya et al. 2023). These issues are amplified by the exponential complexity of the joint state-action space in multi-agent settings, leaving agents with little discriminative learning signal for most time steps, hampering exploration and favoring convergence to suboptimal cooperative patterns.

Sparsity and delay are intrinsically coupled with the credit assignment problem, which concerns how to attribute the global team reward $R(s, a)$ to individual agent contributions. This problem is challenging even with dense, immediate rewards. Under sparse and delayed rewards, it escalates into a severe spatio-temporal credit assignment dilemma: credit must be distributed not only among agents (spatial dimension) but also across the extended sequence of decisions that led to success (temporal dimension). Sparsity obscures *when* critical contributions occurred, while delay severs the direct link between actions and outcomes. Consequently, explaining cooperative behavior under these conditions fundamentally requires transparently deconstructing the complex mechanisms of spatio-temporal credit assignment.

3.1.2 Interplay of challenges and the rationale for a hierarchical framework

The challenges described above do not operate in isolation; rather, they form a complex and mutually reinforcing system. Their interplay gives rise to distinct explainability needs at different levels of analysis, directly corresponding to the four-layer framework proposed in this survey.

Environmental non-stationarity and partial observability jointly necessitate explanations at the **micro level** (individual decision logic) and **meso** (inter-agent coordination) levels. To understand decision-making in a dynamic and uncertain environment, we must explain an agent’s internal policy mapping π_i (motivating policy abstraction methods) and reveal its perceptual focus and internal beliefs (motivating attention visualization and goal analysis). To understand how agents coordinate under partial observability, we must make their interaction structures explicit, thereby motivating methods such as attention-based relational modeling, communication semantics analysis, and causal validation of interactions.

Reward sparsity, delayed feedback, and their coupling with credit assignment jointly drive the need for explanations at the **macro level** (team strategy) and **meso level** (credit flow). To extract the essence of team behavior from long, sparse-reward sequences, we need techniques for high-level behavior abstraction, goal identification, and task phase segmentation. More critically, addressing the spatio-temporal credit assignment problem requires methods for long-horizon causal attribution and equitable contribution quantification. This motivates techniques like counterfactual reasoning, value function decomposition analysis, and credit-assignment methods based on game-theoretic or causal contributions.

Ultimately, the combined effect of all challenges determines the system’s overall robustness, efficiency, and fairness, creating a need for diagnostic explanations at the **system level**. This necessitates comprehensive evaluation frameworks to identify performance bottlenecks, diagnose failures arising from unfair credit assignment or poor coordination, and assess overall robustness to non-stationarity and partial observability.

The challenges of cooperative MARL are thus inherently hierarchical, operating at the levels of individual agent cognition (micro), inter-agent interaction (meso), emergent

team strategy (macro), and overall system performance (system). Their complexity and interdependence render any single-level explanation inadequate. Therefore, constructing a corresponding hierarchical explainability framework is not only a practical response to transparency demands but also a useful conceptual basis for understanding collaborative intelligence. The “Micro–Meso–Macro–System” framework proposed in this survey is designed to meet this need, ensuring systematic coverage of the full explanation pipeline.

3.2 Research paradigms and their inherent explainability challenges

To tackle the challenges of cooperative MARL, researchers have developed diverse algorithmic paradigms that organize information flow, decision authority, and functional structure among agents in distinct ways. These paradigms, while improving performance, also create specific explainability challenges. We categorize these paradigms along two core dimensions: **Agent Organizational Architecture** and **Policy Representation & Coordination Mechanisms**.

3.2.1 Architectural paradigms: decentralized, centralized, and hybrid

This dimension classifies paradigms according to the information available during training and execution, defining the system’s basic organizational principle.

- **Decentralized Training with Decentralized Execution (DTDE):** Each agent learns and acts based solely on its local history τ_i , treating others as part of the environment. While simple and decentralized, it struggles with non-stationarity and complex coordination. Explainability here is confined to the individual agent’s local perspective.
- **Centralized Training with Centralized Execution (CTCE):** A central controller has access to all agents’ information (joint observations or global state) and outputs joint actions during both training and execution. This reduces the problem to a large single-agent task but suffers from the curse of dimensionality in the joint action space and single-point failure risk. Explainability focuses entirely on the central controller’s logic.
- **Centralized Training with Decentralized Execution (CTDE):** The current mainstream paradigm. During training, it is allowed to utilize global information to learn a centralized evaluation network (Critic) or guidance signals to overcome non-stationarity and credit assignment issues; during execution, each agent makes independent decisions based solely on its local observation o_i . It achieves a balance between training efficiency and execution feasibility.

This dimension defines the basic rules of “who can see what”. Under the DTDE paradigm, explanations must be entirely based on the individual’s local perspective; in the CTCE paradigm, the explanation problem is focused on the internal logic of the central controller; while in the CTDE paradigm, a core explanatory gap is created: how does the global value function $Q_{tot}(s, a)$ or $V(s)$ learned by the centralized evaluation network form? And how does it implicitly shape the decision preferences of each decentralized strategy π_i through gradient updates? This “global–local” causal influence chain is the key starting point for subsequent multi-level explainability analysis (especially the connection between micro individuals and macro teams).

3.2.2 Structured organizational paradigms: hierarchy and roles

To enhance the management and decomposition efficiency of complex tasks, research has introduced explicit organizational structures on top of hybrid paradigms like CTDE (Ma et al. 2023). These structures transform agent collaboration from an implicit, flat model into an explicit, structured one through functional division and hierarchical management. This advancement directly corresponds to and deepens the theoretical understanding of the “meso-interaction” and “macro-team” levels in MASSs. Two prevalent and complementary paradigms have emerged: hierarchical organization and role-based organization.

Hierarchical organization simulates the logic of vertical management and task decomposition found in real-world systems. Its architecture typically revolves around two functionally stratified agent types (Ma et al. 2023). A high-level *policy coordination agent* acts as a central decision unit, periodically generating high-level directives—such as sub-goals or task assignments—based on global information, thereby fulfilling macro-planning functions. Its training often employs centralized methods. Low-level *tactical execution agents* operate as distributed executors, receiving high-level instructions and translating them into concrete actions based on their local observations. The low-level coordination frequently utilizes the CTDE paradigm to achieve decentralized collaboration (Xu et al. 2023). This architecture fundamentally constitutes a hybrid “planning-execution” separation paradigm.

Role-based organization, in contrast, focuses on horizontal functional specialization. This paradigm constrains an agent’s policy space, objective function, or communication permissions by assigning or learning specific functional roles (e.g., leader, attacker) (Hu et al. 2024d). Roles can be pre-defined or dynamically allocated and switched based on the team state to achieve flexible adaptation (Lan et al. 2025). By defining coordinating roles (e.g., a leader) and functional roles (e.g., an explorer), this paradigm encourages agents to develop deep expertise within their designated functions.

These structured paradigms introduce distinct explainability demands. For hierarchical architectures, explanations must address the basis for high-level decisions (i.e., how macro goals are decomposed), the mechanism for aligning high- and low-level instructions, and the resolution of inter-level conflicts. For role-based architectures, explanations should clarify the semantic rationale behind role division, the logic governing role dynamics and evolution, and how a role shapes the underlying individual policy. Investigating these questions is key to understanding the emergence of “macro-team strategies” and the formation of “meso-structured interactions”.

3.2.3 Policy representation and coordination mechanisms

This dimension focuses on how agent policies are parameterized and how coordination is architecturally implemented within the model structure. It emerges as a natural extension of the previous two dimensions, crystallizing the concrete forms that collaboration can take in algorithmic design.

Independent policy learning directly implements the DTDE paradigm, wherein each agent learns a policy network π_i that depends solely on its own history τ_i . The explainability target here is relatively straightforward—the individual policy π_i —yet this approach offers no inherent mechanism for explaining any emergent cooperative behaviors that may arise from the interaction of multiple such independent policies.

Value decomposition networks operate within the CTDE paradigm by factorizing the global action-value function $Q_{tot}(s, a)$ into a monotonic combination of individual value functions $Q_i(\tau_i, a_i)$ (e.g., VDN, QMIX). This formulation establishes consistency between individual policy optimization and global team optimality. It represents a model of *implicit coordination mediated through the value function*. The primary explainability challenge centers on the **semantic veracity of the decomposition**: while mathematically sound, does the factorization genuinely reflect the underlying collaborative logic and true credit assignment structure of the task? Moreover, can the individual Q_i functions be meaningfully interpreted as reliable indicators of an agent's contribution? Addressing these questions is fundamental to explaining both **meso-level interactions** (the flow of credit) and **micro-level individual behavior** (individual value attribution).

Communication models and coordination graphs introduce explicit architectural components to address partial observability. This category includes models with dedicated communication channels, allowing agents to exchange learned messages, and coordination graphs that explicitly model pairwise value dependencies among agents to facilitate approximate joint action selection. These represent models of *explicit coordination through structured interaction*. For communication models, the core explanatory demand lies in providing a **semantic interpretation of the interaction**: what is the content and intended purpose of the messages, and how are they understood by recipients? For coordination graphs, the challenge is to **validate the learned graph structure**, ensuring it authentically reveals the essential **meso-level functional dependencies** among agents rather than capturing spurious data correlations.

The choice of policy representation dimension directly determines the entry points and applicability of different explanation techniques. The distinction between implicit value-based coordination and explicit interaction-based coordination fundamentally shapes the landscape of explainability challenges in cooperative MARL.

3.2.4 Paradigm fusion and the composite challenge of explanation

In practice, state-of-the-art algorithms increasingly incorporate a fusion of the previously delineated dimensions. A contemporary role-based CTDE algorithm, for instance, may simultaneously integrate a role allocation network—defining the collaborative structure—with an attention-based communication mechanism. This profound, multi-level integration of diverse modules transforms the system's decision-making process into an intricate pipeline composed of multiple “black boxes” connected in series or parallel. Consequently, the causal pathway from raw sensory inputs to coordinated multi-agent actions becomes deeply embedded within a complex, layered architecture.

This synthesis of paradigms elevates the challenge of explainability to a new level of complexity. The complexity of cooperative MARL systems therefore comes not just from the Dec-POMDP model, but more importantly from the advanced multi-layered algorithms used to solve it. While these algorithms improve performance, they also create many levels of abstraction and specialized parts. These layers make it hard to trace the complete cause-and-effect process—from a single agent's action to the team's overall cooperation—resulting in a lack of transparency that reduces interpretability.

Thus, the development of explainability techniques capable of **penetrating these layers of abstraction, revealing the logic governing cross-module interactions, and validat-**

ing the rationality of underlying structural assumptions (such as roles or dependency graphs) has become an indispensable prerequisite. It is only through such rigorous explanatory scrutiny that cooperative MARL can transition from experimental algorithms to reliable deployment in safety-critical real-world applications.

3.3 The XRL toolkit and its fundamental shortcomings for MASs

Before delving into the explainability techniques for single agents, it is necessary to clarify two core and often conflated concepts within the field—*interpretability* and *explainability*—to establish a clear terminological foundation for subsequent methodological discussions. **Terminology Clarification: Interpretability vs. Explainability.** Interpretability refers to the inherent transparency of a model's structure or decision-making process, often achieved by design (e.g., decision trees and linear models). An interpretable model allows a human to directly understand its internal logic.

Explainability typically refers to the ability to provide post-hoc rationales for the behavior of a complex, often opaque model (e.g., a DNN). Explanations are generated through external methods (e.g., feature attribution, counterfactual analysis) to make the model's decisions comprehensible.

While the boundary between them can be fluid, this distinction helps frame our discussion: **intrinsic methods** often aim for interpretability, whereas **post-hoc methods** provide explanations.

The overarching goal of Explainable Reinforcement Learning (XRL) is to decipher the “black-box” nature of deep RL models, rendering their decision processes transparent and comprehensible to humans. Over time, XRL has matured into a field with a systematic taxonomy, where methods are typically categorized based on their coupling with the model into **intrinsic** and **post-hoc** explainability. This toolbox of methods provides an important foundation for explainability research in more complex scenarios.

However, it is crucial to recognize that these XRL methods are predominantly built upon single-agent design premises. A fundamental limitation arises when they are directly applied to multi-agent settings: their focus on explaining individual agent decisions often fails to capture the emergent behaviors stemming from agent interactions, collaboration, and competition. This limitation motivates the transition from a single-agent framework to the study of explainability in MASs.

3.3.1 Intrinsic interpretability methods

Intrinsic explainability methods, commonly referred to as “white-box explanations,” focus on designing model architectures and decision processes that are inherently transparent and interpretable by construction. These approaches fundamentally differ from post-hoc methods by embedding explainability directly into the learning framework, rather than applying explanatory techniques after training. This paradigm can be broadly categorized into two complementary strands: those employing **intrinsically interpretable models** and those ensuring **process transparency** throughout the decision-making pipeline.

Intrinsically interpretable model architectures

This category replaces complex DNNs with function approximators whose internal reasoning mechanisms are inherently accessible and human-parsable. The primary advantage

of these approaches lies in their ability to provide direct insight into the policy's decision logic without requiring additional explanatory machinery. Within this framework, two principal architectural strategies have emerged:

Transparent-by-design models utilize classically interpretable models such as linear models, decision trees (Vos and Verwer 2023; Dai et al. 2022), or fuzzy logic systems (Dhebar et al. 2022) to represent the policy. For instance, a decision tree policy maps state features (e.g., distance to target < 5 m, obstacle detected) to concrete actions (e.g., move forward), forming a comprehensive set of human-understandable “if-then” rules that can be directly inspected and validated by domain experts.

Memory-based architectures employ networks with explicit memory modules, such as Key-Value Memory Networks, to store and retrieve interpretable representations (Annasamy and Sycara 2019). These architectures allow the agent's decisions to be explained through the retrieval and weighting of semantically meaningful memory prototypes.

Process-transparent learning frameworks

While intrinsically interpretable models provide direct access to decision logic, an alternative approach maintains transparency throughout the learning and decision-making process itself. These frameworks incorporate structural elements that expose the reasoning pathway at multiple levels of abstraction, even when utilizing complex DNNs. Three prominent methodologies exemplify this strategy:

Hierarchical reinforcement learning (HRL) decomposes complex tasks into hierarchies of semantically meaningful subtasks or options (Gurses et al. 2024; Li et al. 2025c). This decomposition enables explanation by tracking the agent's progression through the task hierarchy (e.g., complete subgoal A, then execute subgoal B), which naturally aligns with human planning structures and provides intuitive insight into the strategic breakdown of complex behaviors.

Building on structural transparency, attention and saliency visualization integrate attention mechanisms directly into policy or value networks, generating saliency maps that visually indicate which components of the input observation the agent prioritizes during decision-making (Ma et al. 2025a; Zhuang et al. 2025; Kim et al. 2022). For example, in autonomous driving applications, saliency visualization might reveal that the agent focused primarily on pedestrian positioning when deciding to initiate braking, providing immediate visual validation of the policy's focus areas.

For extracting higher-level understanding, symbolic knowledge extraction distills the learned knowledge from trained DNN policies into compact sets of symbolic rules, typically expressed in formalisms such as first-order logic (Xie et al. 2025; Acharya et al. 2023; Duttagupta et al. 2025). A fundamental challenge in this approach is ensuring the fidelity of extracted rules (e.g., *IF 'red light detected' AND 'speed > 0' THEN 'brake'*) relative to the original policy's behavior, while maintaining comprehensibility and coverage across the state space.

3.3.2 Post-hoc explainability methods

In contrast to intrinsic methods, post-hoc explainability techniques aim to provide explanations for pre-trained black-box models—such as standard Deep Q-Networks—that are difficult or impractical to modify after training. These methods infer decision logic indirectly by systematically analyzing the model's inputs, outputs, and interaction history, without

requiring direct access to or understanding of the model's internal mechanics. The post-hoc paradigm primarily operates through three methodological families: **perturbation-based analysis**, **imitation learning**, and **trajectory analysis**.

Perturbation and Counterfactual Analysis

These methods infer decision dependencies by systematically modifying inputs or constructing hypothetical scenarios to isolate the influence of specific factors on the agent's behavior. This approach operates on the fundamental principle that causal relationships can be revealed through controlled variations in the input space, with two distinct but related techniques:

Input perturbation modifies specific features of the observation space—typically through techniques such as occlusion, noise injection, or feature masking—and observes corresponding changes in action probabilities to identify critical features driving the policy's decisions (Huber et al. 2022; Helmboldt 2025; Guo et al. 2024; Mehmeti et al. 2025). This approach enables researchers to establish causal relationships between input variations and behavioral changes, albeit within the limitations of the perturbation scheme.

Extending beyond simple perturbation, **Counterfactual Explanations** pose hypothetical questions of the form “How would the agent's decision change if a specific state feature were altered?” (Gajcin and Dusparic 2024; Wang et al. 2024; Amitai et al. 2024). This methodology helps verify the causal influence of particular features—such as detecting an “icy road” surface on braking decisions—by constructing minimally altered scenarios that would lead to different behavioral outcomes, thereby illuminating the decision boundaries of the policy and providing insight into alternative behaviors under modified conditions.

Behavioral imitation and policy extraction

While perturbation methods focus on understanding specific decision points, a more comprehensive approach involves creating global approximations of the entire policy. This methodology employs an intrinsically interpretable “student model”—such as a decision tree or linear model—to mimic the input–output behavior of the black-box “teacher” RL policy (Li et al. 2024e; Acero and Li 2024). The transparent student model subsequently provides an indirect, global explanation of the teacher's decision logic, effectively translating the opaque policy into a human-interpretable representation while attempting to preserve its functional characteristics across the entire state space, though with inherent trade-offs between fidelity and interpretability.

Trajectory analysis and pattern discovery

Moving beyond individual decisions or global policy approximations, trajectory analysis techniques extract high-level decision patterns and behavioral regularities from historical agent–environment interaction data. This approach focuses on understanding the temporal structure and strategic progression of decision-making over extended time horizons, with two complementary methodologies:

Reward decomposition operates from a motivational perspective, breaking down the global cumulative reward signal into semantically meaningful sub-rewards corresponding to different aspects of performance (e.g., survival reward, exploration reward, goal achievement reward) (Towers et al. 2024; Mahmud et al. 2023; Ameer et al. 2024; Roghani and Koyuncu 2025). This decomposition clarifies which specific objectives the agent was optimizing when taking particular actions, providing insight into the relative importance of different behavioral dimensions within the overall policy and revealing the underlying value structure driving agent behavior across time.

Complementing this motivational analysis, key decision point identification applies time-series analysis, clustering, or saliency detection techniques to long interaction trajectories to pinpoint critical moments that disproportionately influence final outcomes (Takagi et al. 2024; Cheng et al. 2023; Wang and Aouf 2024). By identifying these pivotal decision points within extended behavioral sequences, researchers can focus explanatory efforts on the most consequential aspects of the agent's behavior, enabling more efficient and targeted understanding of how specific decisions propagate through time to shape overall performance and strategic outcomes.

3.3.3 Fundamental limitations of explainable reinforcement learning in multi-agent systems

Although the XRL toolkit provides a foundational set of techniques, its core paradigm is built around a single agent operating in a stationary environment under assumptions of individual rationality. Directly transferring these methods to multi-agent settings reveals several fundamental limitations:

Insufficient granularity and scope. XRL methods typically focus on instantaneous, single-step decisions, which prove inadequate for capturing the long-term temporal dependencies, complex commitment mechanisms, and dynamically evolving collaboration logic that characterize multi-agent interactions.

Neglect of interactive context. The explanatory scope of single-agent methods excludes core multi-agent phenomena. They cannot elucidate the specific process of credit assignment, decipher the true semantics of communicated messages, or explain how social conventions and role specializations spontaneously emerge from agent interactions.

Difficulty in causal attribution. In a MAS, an agent's decision is the result of the interplay between its own policy, the actions of others, and environmental dynamics. Most XRL methods, which reveal statistical correlations, struggle to accurately distinguish true causal drivers from spurious associations within this complex network. They fail to answer critical questions such as, "Did agent A's action cause agent B's response, or did they both react to a common environmental change?"

Misaligned rationality assumption. The endpoint of a standard XRL explanation is "how this action maximizes individual cumulative reward." In cooperative MARL, however, the starting point for explanation must be "how this action promotes or hinders the achievement of the collective goal." This paradigm shift from individual to collective rationality demands that explainability methods treat team cooperation effectiveness as the primary object of explanation.

In summary, while single-agent XRL offers a substantial repository of techniques, its inherent design boundaries make it inadequate for addressing the core challenges of MASs. Consequently, the development of XMARL cannot be a simple transplantation of XRL; it necessitates a fundamental paradigm reconfiguration oriented towards interaction, collaboration, and emergence.

3.4 Towards explainable multi-agent reinforcement learning: a paradigm shift

The discussions in Sects. 3.2 and 3.3 clarify that the transition from XRL to XMARL is not merely a linear technical extension, but a fundamental paradigm shift. This shift involves the very objects of explanation, the basis of rationality, and the core questions being addressed.

3.4.1 Core dimensions of the paradigm shift

The explanatory paradigm of XRL is built upon assumptions of individual rationality and a static environment, aiming to trace the independent decision chain of a single agent. Key differences between this paradigm and the one required for XMARL are summarized in Table 2.

As previously established, the core of MARL lies in the interaction and collaboration among agents, which fundamentally alters the explainability problem landscape. This shift manifests in three critical limitations when XRL methods are directly applied:

Limitation 1: Inability to Explain Decision Synergy. Single-agent attention mechanisms can identify features an individual agent focuses on within its local observation, but they cannot account for how this focus coordinates with the anticipated behaviors or strategic intentions of other agents. Consequently, they fail to explain how decentralized individual decisions coalesce into a coherent cooperative action.

Limitation 2: Inadequacy for Dynamic Interaction Analysis. Single-agent counterfactual methods are well-suited for analyzing perturbations to environmental states but are poorly equipped to model strategic interdependencies among learning agents. They cannot

Table 2 Comparison between XRL and XMARL

Comparison dimension	XRL	XMARL
Rationality basis	<i>Individual rationality</i> : explaining why the agent takes action a_t to maximize its own long-term return G_t	<i>Collective rationality</i> : explaining how individual action a_t^i contributes to or hinders the team goal G_{team} ; requiring trade-offs between individual and cooperative interests
Environment characteristics	Environment dynamics $P(s' s, a)$ are <i>stationary</i> , depending only on the agent's own action and natural state transitions	Environment is <i>non-stationary</i> , with dynamics $P(s' s, a)$ changing due to other agents' policies π^{-i} , requiring explanation of interactive intentions and mutual influences
Core explanation object	<i>Single decision chain</i> : focusing on the mapping from state to action to value	<i>Interactive decision network</i> : covering the full spectrum of behavioral hierarchy from micro individual decisions, meso-level coordination, to macro cooperative emergence
Core challenges	Action selection basis; value estimation credibility; state attention	Credit assignment; communication semantics; role division and evolution; collaboration structure; system robustness

address critical questions regarding cascading effects within a team, such as how the joint policy would adapt in response to an alternative action taken by an agent.

Limitation 3: Failure to Abstract Collaboration Protocols. Techniques like trajectory-based decision tree extraction can produce interpretable models of an individual agent's policy but are inherently limited to state-action mappings localized to a single agent. They cannot represent the distributed, often implicit, collaboration protocols and conventions that emerge from repeated interactions among multiple agents, as these are properties of the group's interaction dynamics.

3.4.2 Current landscape: fragmented research and the need for a unifying framework

As MARL has advanced in complex tasks, the need for explainability has gained significant attention, prompting various exploratory studies. However, current XMARL research remains largely fragmented. It is characterized by isolated, technology-driven solutions and problem-specific methods. The field lacks a unified analytical framework to systematically integrate existing work and guide future development.

Technology-driven point solutions. Many studies adopt specific technical architectures, such as graph neural networks or attention mechanisms, to improve MARL performance. Explainability in these works is often a secondary feature. For example, a method might use graph structures to model agent interactions for better credit assignment (Kortvelesy and Prorok 2022). While the resulting attention-weight visualizations can hint at agent dependencies, explainability itself is not the primary goal. These approaches produce explanations that are local and tied to the specific technology used.

Problem-specific customized methods. Another category designs explanation methods for narrow task domains, like multi-robot navigation or social network modeling. For instance, a study might use predefined, interpretable parameters to simulate social network evolution (Miyake et al. 2024). Such methods offer strong explanations within their specific context but have limited generalizability. They are closely coupled with the task's reward structure and domain knowledge, making them hard to apply to other multi-agent problems.

This fragmented state is evident across different levels of explanation:

- **Micro Level (Individual Agents).** Current methods can explain a single agent's decisions using policy graphs (Gimenez-Abalos et al. 2025), interpretable decision trees (Milani et al. 2022b), or behavior trees (Wan et al. 2025a). However, these explanations are often shallow. They focus on the agent's local logic without connecting its actions to the team's overall goals. There is a clear need to link micro-level explanations to higher-level collaborative objectives.
- **Meso Level (Inter-Agent Coordination).** Research here is active but divided. Different studies focus on credit assignment (Kortvelesy and Prorok 2022), communication semantics (Li et al. 2024c), causal structure (Meli et al. 2025), or dynamic roles (Li et al. 2024d). Each perspective uses its own concepts and methods. They rarely connect with each other, leading to a partial understanding of collaboration. A unified theory of agent interaction is missing.
- **Macro and System Levels.** Work on explaining emergent team strategy or overall system robustness is scarce. Some studies analyze team stability (Barfuss et al. 2025) or quantify an action's global impact (Li et al. 2024f). However, these efforts are isolated.

They do not form a coherent theory for macro-level team intelligence or system-level performance. The explanatory chain from low-level mechanisms to high-level outcomes is broken.

Lack of standardized evaluation. The field also lacks common evaluation benchmarks, protocols, and metrics. Researchers often use task performance as a proxy for explanation quality or rely on custom visualizations. Different methods prioritize different aspects, such as fidelity or simplicity. This makes comparisons difficult and reinforces the isolation between research approaches. It slows down progress and consensus-building in the field.

In summary, XMARL research is currently fragmented. This fragmentation exists at every level, from individual agent explanations to system-wide analysis. To move forward, the field needs a unifying framework. This framework should transcend specific tools and tasks. It should connect different explanation levels and provide a common reference for evaluating methods. The four-dimensional framework proposed in this survey aims to meet this need. It seeks to organize and integrate diverse research, providing a clearer path toward transparent and trustworthy MASs.

3.4.3 The need for a systematic taxonomy in XMARL

The preceding analysis underscores that the XMARL field has reached a critical juncture. Moving beyond isolated, ad-hoc solutions requires a systematic taxonomy capable of integrating diverse research efforts and providing coherent guidance. Such a framework is not arbitrary; it directly reflects the hierarchical nature of decision-making in cooperative MARL (Sect. 3.1) and constitutes a necessary evolution from the tools offered by single-agent XRL (Sect. 3.3).

An effective taxonomy for XMARL must fulfill three core objectives:

1. **Achieve Comprehensive Coverage.** The taxonomy should seamlessly incorporate the full spectrum of explanation foci, spanning micro-level (individual agent decisions), meso-level (inter-agent coordination), macro-level (team strategy and emergence), and system-level (global performance and robustness) concerns. This breadth is essential for identifying and addressing gaps in current research.
2. **Enable Technical Deconstruction.** It must provide a structured understanding of how different explanation methods operate. The framework should facilitate clear comparisons by elucidating how specific techniques tackle particular challenges (e.g., credit assignment), while critically examining their underlying assumptions, strengths, and limitations.
3. **Offer Directional Guidance.** Acting as a “knowledge map,” the taxonomy should allow for the precise categorization of existing work. This will help identify well-explored areas and, more importantly, highlight critical yet underexplored challenges, thereby providing a clear roadmap for future research.

To meet these objectives, this survey proposes a novel classification framework structured along two primary dimensions: *Behavioral Hierarchy* (the Explanatory Focus) and *Technical Method*. The Behavioral Hierarchy dimension is derived from the intrinsic structure of cooperative MARL problems (Sect. 3.1). The Technical Method dimension adapts and extends techniques from single-agent XRL (Sect. 3.3) to the multi-agent setting. By integrating these two axes, this framework aims to establish a unified foundation and a struc-

tured analytical lens for the currently fragmented field of XMARL, thereby setting the stage for the systematic review in Sect. 4.

4 A taxonomy of XMARL methods across behavioral hierarchies

Explainability is increasingly recognized as an important requirement for deploying MARL in safety-critical domains. While XMARL encompasses a growing set of approaches, they can be systematically organized through a unified taxonomy based on two core dimensions: the **behavioral hierarchy** of the explanation target and the **technical approach** employed. This section surveys the XMARL landscape through this structured lens, building on the coding framework established in Sect. 2.5.

4.1 The XMARL landscape: a structured overview

The field of XMARL encompasses both post-hoc explanations to trained policies and intrinsically interpretable architectures intended to support explanation throughout learning. Drawing on the systematic corpus (Sect. 2.4), we organize the literature along two primary dimensions.

Dimension 1: Behavioral hierarchy

This dimension categorizes methods based on the granularity and scope of the behavior they aim to explain.

Micro-Level (Individual Behavior) This most granular level seeks to uncover the internal policy of a single agent, answering fundamental questions about action selection, behavioral preferences, and the mapping from states and observations to actions.

Meso-Level (Inter-Agent Coordination) This level focuses on the dynamics between agents, aiming to explicate the mechanisms underlying collaboration, such as communication protocols, multi-agent attention, and role assignment.

Macro-Level (Team-Level Strategies) Adopting a macroscopic view, this level is concerned with the global policies and tactical patterns that emerge from local interactions, treating the team as a unified entity with its own strategic behavior.

System-Level (Performance & Adaptation) The broadest level analyzes the MAS holistically, diagnosing global performance characteristics, robustness, and its capacity to adapt to non-stationary environments.

Dimension 2: Technical approach

This dimension classifies the core methodologies used to achieve interpretability.

Policy Abstraction and Extraction. These approaches distill complex deep network policies into interpretable symbolic models, such as decision trees (Milani et al. 2022b), behavior trees (Wan et al. 2025a), or policy graphs (Gimenez-Abalos et al. 2025), to render the decision rules transparent.

Attention and Feature Attribution. These methods leverage attention weights (Motokawa and Sugawara 2023a) and saliency maps (Motokawa and Sugawara 2023b, c) to identify and visualize the input features most critical to an agent's decision.

Causal and Counterfactual Analysis. Moving beyond correlation, these techniques employ causal discovery (Briglia et al. 2025; Meli et al. 2025) and counterfactual reasoning (Bogge et al. 2022; Gyevnár et al. 2025) to uncover the causal mechanisms that drive agent

behaviors and system outcomes.

Credit Assignment and Value Decomposition. These methods, including those based on Shapley values (Heuillet et al. 2022) and causal contribution analysis (Yao et al. 2024; Zhang et al. 2024a), aim to attribute global success or failure to individual agents' actions.

Structured Communication and Relational Modeling. This pathway enhances transparency in agent interactions by learning explicit communication protocols (Yang et al. 2024; Li et al. 2024b) or dynamic relational graphs (Fu et al. 2024; Fan et al. 2025) that define the structure of collaboration.

Semantic and Symbolic Integration. By incorporating high-level concepts (Zabounidis et al. 2023; Grupen et al. 2022) or symbolic logic (Ji et al. 2023; Mu et al. 2025) into the learning process, these methods align the agent's internal representations with human-understandable semantics.

Operational criteria for level assignment.

To resolve boundary ambiguity, especially between the Meso and Macro levels, we operationalise the hierarchy using the following heuristics, which extend the coding rules of Sect. 2.5.

(1) Explanation Target Granularity.

- *Micro*: Focuses on the **internal state-action mapping** of a single agent, independent of concurrent teammate influences. The explanation answers “*What rule does this agent follow?*” (e.g., decision trees in Sect. 4.2.1) or “*What does this agent perceive?*” (e.g., saliency maps in Sect. 4.2.2).
- *Meso*: Focuses on **pairwise or group-wise dependencies** between agents. The explanation targets the *mechanism of coordination*—encompassing information flow (Sect. 4.3.1), communication semantics (Sect. 4.3.2), causal relations (Sect. 4.3.3), role structures (Sect. 4.3.4), and credit allocation (Sect. 4.3.5)—rather than the individual policy itself.
- *Macro*: Focuses on **global, team-wide behavioral patterns** that emerge from local interactions. The explanation addresses “*What strategy does the team as a whole exhibit?*” (Sect. 4.4.1), “*How does this global pattern evolve over time?*” (Sect. 4.4.2), “*Does the team constitute an irreducible whole?*” (Sect. 4.4.3), “*What are the key strategic contingencies?*” (Sect. 4.4.4), and “*What generative rules underlie team behavior?*” (Sect. 4.4.5).
- *System*: Focuses on **performance properties of the entire MAS in its environment**. The explanation targets systemic robustness, adaptation limits, or failure modes (Sect. 4.5), rather than being limited to specific tactical patterns.

(2) Temporal and Structural Scope.

- *Micro* and *Meso* explanations typically concern **instantaneous or short-horizon** decision events (single timestep actions, immediate communication acts).
- *Macro* explanations inherently involve **long-horizon, temporally extended** phenomena—phase transitions (Sect. 4.4.2), strategic subgoal sequences (Sect. 4.4.1), or emergent formation stability over episodes.

- *System* explanations may span multiple episodes or environmental contexts, addressing **across-run** performance characteristics and sensitivity analyses (Sect. 4.5.2).

(3) Multi-Agent Specificity.

- **Meso vs. Micro:** Assignment to *Meso* or higher requires the explanation to address challenges unique to multi-agent settings (e.g., non-stationarity, inter-agent credit assignment). Explanations of isolated agent policies remain *Micro*.
- **Macro vs. Meso:** Assignment to *Macro* requires the explanatory construct to **aggregate over the entire team** (e.g., global phase labels, joint strategy embeddings), transcending pairwise relations.

These criteria align the taxonomy with the structural organization of the subsequent sections. Section 4.2 through Sect. 4.5 each correspond to one hierarchical level, with internal technical pathways further detailed therein (Table 3).

In summary, the field of XMARL constitutes an integrated research ecosystem that addresses multiple behavioral hierarchies through diverse technical approaches. The overarching goal is to develop MASs that are capable, transparent, trustworthy, and controllable. Persistent challenges include ensuring explanation faithfulness, maintaining consistency in non-stationary environments, balancing explanatory granularity with computational efficiency, and establishing standardised evaluation benchmarks. The subsequent sections present a systematic analysis of these technical approaches, organised according to the behavioral hierarchy defined above.

4.2 Explainability techniques for micro-level individual behavior

The explainability of MASs lies not only in understanding macro-level emergent group behaviors but also in deeply dissecting the internal decision logic and behavioral patterns of each micro-level agent constituting the system. Comprehensively understanding an agent's behavior requires systematically answering four progressive and complementary core questions (as illustrated in Fig. 4): *What rules does it follow?* (Decision Logic), *What does it focus on when making decisions?* (Perceptual Focus), *Why does it choose this action rather than an alternative?* (Causal influence), and *What is the fundamental pursuit behind its behavior?* (Intrinsic Motivation). These four questions provide an organizing framework for surveying *micro-level agent behavior explainability*. Accordingly, existing research has formed four major technical paradigms, each targeting the above core questions and providing explanations complementary analytical perspectives: **Policy Abstraction and Model Extraction**, aiming to recover the rules of “how” decisions are made; **Attention and Feature Attribution**, aiming to reveal “what” environmental information is focused on; **Counterfactual Reasoning and Influence Analysis**, aiming to explore the causal mechanism of “why” choices are made; **Intrinsic Motivation and Goal Analysis**, aiming to analyze the fundamental drives behind “why something is pursued”. These four categories of methods are not isolated but form a complete explanatory system from surface-level behavioral description to deep-level motivational inference. The following sections will systematically elaborate on how these four categories of methods focus on and synergistically cover the aforementioned problem space.

Table 3 Operationalized summary of the four-level XMARL behavioral hierarchy

Level	Explanation target & core question	Analytical granularity & temporal scale	Representative pathways & typical outputs	Inherent limitations (as discussed)
Micro (Sect. 4.2)	Individual decision logic. “What rules/perceptions drive this agent’s actions?”	Single-agent state-action mapping. Instantaneous or short-horizon	<i>Policy Abstraction</i> (Sec. 4.2.1): symbolic models (decision/behavior trees, policy graphs). <i>Feature Attribution</i> (Sec. 4.2.2): saliency/attention heatmaps, feature importance vectors. <i>Counterfactual Reasoning</i> (Sec. 4.2.3): contrastive action pairs, causal contribution scores. <i>Motivation Analysis</i> (Sec. 4.2.4): concept activation vectors, interpretable reward parameters	Poor scalability to high-dim states; correlation vs. causation ambiguity; reliance on manual concept/predicate design; explanation fidelity trade-offs
Meso (Sect. 4.3)	Inter-agent coordination mechanisms. “How do agents interact and assign credit?”	Pairwise/group dependencies. Timestep-level interactions	<i>Attention & Relational Modeling</i> (Sec. 4.3.1): interaction graphs, attention matrices. <i>Communication Protocol Analysis</i> (Sec. 4.3.2): learned message protocols, communication topologies. <i>Causal Discovery</i> (Sec. 4.3.3): causal graphs, counterfactual influence scores. <i>Role & Pattern Discovery</i> (Sec. 4.3.4): role assignments, behavioral pattern clusters. <i>Value Decomposition</i> (Sec. 4.3.5): Shapley values, contribution weights	Explanations often correlative; protocol generalization challenges; high computational cost for causal/game-theoretic methods; semantic gap in learned roles
Macro (Sect. 4.4)	Team-level emergent strategies. “What global patterns and phase transitions characterize the team?”	Aggregate team behavior. Long-horizon, episode-level	<i>Behavior Abstraction</i> (Sec. 4.4.1): high-level strategy labels, subgoal sequences. <i>Macro-State Modeling</i> (Sec. 4.4.2): phase transition diagrams, spatio-temporal evolution graphs. <i>Causal Emergence</i> (Sec. 4.4.3): emergence metrics, macroscopic causal models. <i>Team Counterfactuals</i> (Sec. 4.4.4): strategy-level natural language explanations. <i>World Models</i> (Sec. 4.4.5): semantic reward functions, generative team behavior rules	Static grouping assumptions; weak cross-scale causal links; idealized theoretical models; nascent evaluation frameworks
System (Sect. 4.5)	Systemic performance & environmental robustness. “Why does the system succeed/fail under perturbation?”	Whole-MAS in environment. Across episodes/contexts	<i>Performance Diagnosis</i> (Sec. 4.5.1): agent contribution scores, causal performance pathways. <i>Environmental Response</i> (Sec. 4.5.2): robustness heatmaps, verified failure conditions	Trade-off between computational efficiency and explanatory depth; lack of online diagnostic tools; scalability limits of formal verification

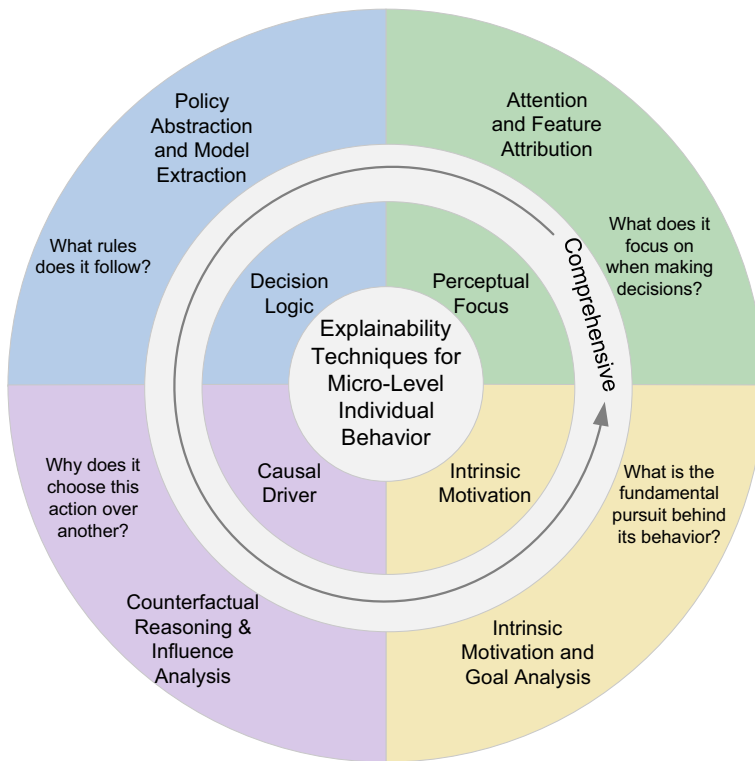


Fig. 4 Technical framework for explaining microscopic individual behaviors: core issues and four complementary technologies

4.2.1 Policy abstraction and model extraction

This category of methods aims to answer **“What decision rules does the agent follow?”** Its technical core lies in transforming trained black-box policies (e.g., DNNs) into human-intelligible symbolic or structured representations (e.g., decision trees, behavior trees, policy graphs). In the explainability hierarchy, these methods reside at the foundational logic layer, aiming to provide transparent traceability of the decision process and offer a foundational representation of behavioral logic for subsequent causal analysis (“why this choice”) and motivational inference (“why pursue”). Based on their implementation mechanisms and representation forms, this category can be further subdivided into the following four technical approaches:

Transparent model design

To avoid the “black box” problem, some research directly designs inherently interpretable policy representations. As a typical example, Luvisutto et al. (2025) adopt a tabular Expected SARSA algorithm for the collaborative pipeline inspection mission of underwater robots, and construct discrete state tuples that possess well-defined physical meanings. The learned Q-table can be directly visualized, clearly presenting collaborative group behaviors. Core tools are classical tabular RL and semantic state representations. The advantage is inherent model transparency, requiring no post-hoc explanation, suitable for simple

state spaces and real-time edge-deployment scenarios. Limitations include restricted policy expressiveness due to discretization and difficulty adapting to continuous/high-dimensional states and complex environments. With breakthroughs in Large Language Model (LLM) code generation capabilities, the paradigm of transparent model design is evolving from manually defined structures to machine-automated discovery. An emerging “**generative explainability**” path directly uses interpretable symbolic programs (e.g., Python code) as the model itself, leveraging LLM-driven program synthesis techniques to automatically search for optimal policy representations. Representative work by Castro et al. (2025b) proposes a two-level optimization framework: the outer level uses an LLM-enhanced evolutionary algorithm to search for policy structures in program space, and the inner level fits individualized parameters via gradient descent. The core idea is embedding explainability within the search space, where the final output program (containing semantic variables like `q_values`, `exploration_rate`) itself constitutes an inspectable transparent “white-box” model. This approach was applied to a rat reward-guided decision-making task. It successfully discovered programs that match or surpass the performance of state-of-the-art handcrafted models, with the generated code directly providing new, testable hypotheses for underlying behavioral mechanisms (Castro et al. 2025a). This suggests a possible transition from “*manual hypothesis-driven*” to “*data-and-algorithm co-discovery*”. Although currently limited by high computational costs and readability challenges of generated code, it provides a highly promising new avenue for automatically constructing inherently transparent and high-performance policy models in MASs.

Rule distillation from black-box policies

This is the most mainstream technical path, distilling interpretable symbolic models from pre-trained black-box neural network policies. Among these, decision trees are favored for their intuitive structure. Representative frameworks like MAVIPER (Milani et al. 2022b) use imitation learning to extract decision trees for all agents from a centrally trained expert policy, introducing a key-state resampling mechanism based on centralized Q-values to theoretically ensure that the extracted individual decision trees internalize collaborative logic. Its performance approaches that of the expert policy while being logically transparent but depends on expert policy quality and incurs high computational overhead. Chen et al. (2025c) builds upon this by introducing team Q-value resampling and a UCB-based policy selection mechanism to improve training efficiency. To enhance explanation readability, Xi-Jia et al. (2025) precisely distills a black-box policy into a decision tree via the DAGger algorithm, extracting decision paths as behavioral representations; subsequently, it uses LLMs to translate these paths into fluent natural language descriptions. The explanations are highly readable with low hallucination rates but are limited to instantaneous decision logic. In traffic signal control, Zhang et al. (2024b) combines decision tree rule distillation and SHAP feature attribution within the MADDPG actor-critic framework. It uses a multi-level visualization system to connect individual decisions to overall traffic patterns. While effective in simple networks, the method faces challenges in fidelity and scalability due to its reliance on proxy models.

Hierarchical and probabilistic policy representations

For decisions with complex temporal logic or requiring structured behavioral queries, behavior trees and policy graphs provide more structured representations. Wan et al. (2025a) proposes a method that discretizes states using time-aware clustering algorithms, such as K-Means, to construct modular behavior trees. This approach effectively captures task

decomposition and sequential execution logic. While it successfully balances performance and interpretability in complex multi-agent environments like StarCraft, its clustering and grouping processes depend on specific hyperparameter choices and require manual intervention. Domènech I Vila et al. (2022) adopts a probabilistic graphical model for the Overcooked-AI collaborative environment, mapping agent states into nodes based on discrete predicates and constructing a policy graph by recording interaction data. This approach supports behavior queries and is model-agnostic, yet the quality of explanation heavily depends on the appropriateness of state discretization. Gimenez-Abalos et al. (2025) further extends policy graphs by introducing hypothetical *desires* as assumed goals, computing *intent values* through reverse probability propagation on the graph structure, thereby providing teleological explanations for individual behavior. This method supports goal-driven semantic explanations and explicitly quantifies the trade-off between explainability and reliability but depends on sufficient observational data to build high-quality probabilistic graphs and requires manual assumption of goals.

Semantic mapping and policy comparison

This class of methods does not output rules directly but maps policies to a semantic space for deep analysis and comparison. Ji et al. (2023) proposed a neuro-symbolic method that fuses deep learning with symbolic reasoning. It uses attention mechanisms to softly select important predicates and multi-hop reasoning paths, ultimately generating first-order logic rules corresponding to agent-action-environment-state relationships. This method achieves simultaneous learning of decisions and explanations, but the rule space is limited by predefined predicates and reasoning steps. Reiestad (2025) focuses on automatically discovering interpretable semantic concepts. It utilizes concept-based explanation methods like TCAV and ConceptSHAP, identifying concepts with high cohesion, separation, and information gain from network activations through regularization constraints, and trains interpretable models based on these concept scores. This method reduces reliance on manually defined concepts and can reveal emergent collaborative behaviors but may suffer from information loss and depends on the type of surrogate model. Hu et al. (2024c) approaches from a metric perspective, mapping policies of different agents to a unified latent space via autoencoders and computing policy distances. This provides a general tool for quantitatively comparing differences between agents at overall or specific behavioral dimensions. The method is general, supports comparison across different action spaces and policy types, and can guide dynamic parameter sharing. However, the latent space itself is a black box, and computing policy distances requires Monte Carlo sampling, incurring significant overhead.

In summary, policy abstraction and model extraction techniques provide a direct path to achieving micro-level explainability. To systematically summarize their characteristics, Table 9 provides a comparative analysis of representative studies from the perspectives of technical approach, core mechanism, advantages, limitations, and applicable scenarios.

Notwithstanding their utility in providing intuitive, structured decision-logic explanations, these methods face inherent limitations. First, symbolic models like decision trees often lack the expressive power to handle continuous state spaces or high-dimensional perceptual data effectively. Second, the extraction process heavily relies on manually designed predicates and domain concepts, which introduces a dependency on expert knowledge and risks embedding designer bias. Finally, the model distillation process itself can introduce approximation errors, leading to a mismatch between the behavior of the extracted model and the original policy—a fundamental challenge known as the explanation fidelity problem.

To address these challenges, future research should pursue the following directions:

- **Automated Concept and Predicate Discovery.** Reducing reliance on manual design is crucial. Future work should develop automated and semi-automated discovery mechanisms for concepts and predicates, potentially by integrating techniques from representation learning and leveraging the semantic capabilities of LLMs. This would enhance the generalizability and reduce the bias of semantic mapping methods, building upon early efforts in automated concept discovery (Reiestad 2025).
- **Expressive Hybrid Explanation Models.** A promising path forward is the deeper integration of symbolic transparency with sub-symbolic expressiveness. Research should focus on developing hybrid models that retain interpretability without sacrificing performance in complex environments. This could involve refining pipelines like “concept discovery → semantic mapping → rule extraction” (Reiestad 2025) or advancing end-to-end neuro-symbolic frameworks that jointly learn policies and their logic, as preliminarily explored by Ji et al. (2023).
- **Rigorous Evaluation of Explanation Fidelity.** Establishing unified and rigorous evaluation frameworks is essential. To advance beyond simple performance matching, the community must develop quantifiable metrics that directly assess the faithfulness of explanations—for example, causality-based measures that systematically benchmark the behavioral consistency between an extracted model and the original policy. This would constitute a necessary refinement and standardization of the evaluation practices initiated in prior work (Zhang et al. 2024b; Reiestad 2025). Furthermore, the framework should incorporate considerations for the explanation’s intrinsic quality. This could involve borrowing from emerging research on quantifying the complexity and conciseness of generated symbolic models (e.g., using program analysis metrics) (Castro et al. 2025b). Such a comprehensive approach would enable scientific trade-offs and optimization among multiple objectives like performance, fidelity, and understandability, thereby driving explainability techniques toward greater standardization and practical application.

4.2.2 Attention and feature attribution

This category of methods focuses on answering “*What information does the agent focus on when making decisions?*” Its technical core lies in analyzing the distribution of attention mechanisms within the policy or quantifying the contribution of each input feature to the final decision through attribution algorithms. These methods provide shallow explanations, primarily revealing perceptual focus and input-decision correlations, suitable for diagnosing perceptual biases and understanding the policy’s information utilization patterns. They complement policy abstraction by forming an input-decision correspondence and provide a data foundation for causal analysis. Based on different technical paths, existing research can be classified into the following categories:

Conditional and instruction-modulated attention

This line of work focuses on how attention is dynamically shaped by high-level context. The DA6-X framework proposed by Motokawa and Sugawara (2023b) processes global state information via multi-head self-attention to generate saliency vectors. Dynamically modulated attention weight heatmaps explain how the agent’s strategy adjusts its focus on

local observations based on the global situation (e.g., its own position). This method is suitable for grid environment collaboration analysis, but its explainability lacks quantitative evaluation. Motokawa and Sugawara (2023c) extends this by introducing a pictorial “*destination channel*” as a high-level instruction. Through a similar modulation mechanism, it achieves controllable and explainable instruction-following behavior. Visualized object collection heatmaps explain the causal relationship between instructions and individual behavior. These methods provide transparent, condition-dependent decision processes but require predefined conditional states or instructions and have limited scope.

Multi-perspective analysis and strategy evolution tracking

To provide finer-grained explanations, some studies decompose or analyze attention temporally. In the DA3-X framework (Motokawa and Sugawara 2023a), two sets of hyperparameters are introduced to selectively mask observation information and positional embeddings at the state embedding layer, thereby separating “**positional attention**”, “**category attention**”, and standard attention within the same network. This representation-ablation-based attribution analysis allows researchers to discern whether an agent’s decision stems from fixed spatial attention patterns or semantic understanding of specific object categories. It enhances the depth of explanation for collaborative strategies in object collection games but still lacks quantitative evaluation of explainability. Chen et al. (2024) approaches from a temporal dimension, using self-attention mechanisms to compress observations from a variable number of agents into fixed-dimensional representations. By analyzing the temporal changes in attention weights, it reveals macro-level phase transitions in individual strategies (e.g., shifting from early collaboration to mid-game offense). This framework performs well in transfer learning but lacks systematic tools to quantify the relationship between these evolutions and instantaneous decisions.

Integrating prior knowledge and structured attribution

These methods incorporate external knowledge or structured prior models to produce more credible or intuitive attributions. Liu et al. (2025a) encodes human attention priors as fuzzy logic rules and dynamically generates parameters for the policy network via hypernetworks, thereby injecting prior knowledge into the agent in the form of interpretable attention weights. This method reduces reliance on expert demonstrations in StarCraft environments and improves coordination among heterogeneous agents, but its performance is sensitive to the design of fuzzy rules and attention quality. Yang et al. (2025b) learns task-relevant compact visual representations via contrastive variational bounds, avoiding direct reconstruction of high-dimensional pixels. The generated spatial attention heatmaps aim to explain the agent’s perceptual basis from pixel input. This method improves sample efficiency and generalization, but its explanations remain essentially correlational attention maps, and the model structure is complex.

Faithful attribution based on hybrid models

To overcome the potential disconnect between post-hoc attribution and the actual decision process, A prominent line of research focuses on constructing intrinsically interpretable policy architectures. One such method, proposed by Schier et al. (2025), employs a hybrid probabilistic model: it separately parameterizes the action distribution for each entity (or subspace) within the observation and then integrates them through a weighted combination to form the final policy. After a decision, the distribution corresponding to each entity and its weight can be directly visualized, clearly showing each observed object’s contribution to the action selection. This explanation is highly consistent with the policy’s generation

process, offering higher fidelity than post-hoc methods, and demonstrates computational efficiency and intuitive explanation in autonomous driving simulation environments. However, this method requires object-level structured observation input, and the mixture weight parameters can be sensitive.

Hard attention and efficiency improvement

Another line of research aims to improve efficiency and robustness through hierarchical frameworks. Representative work like Cheng et al. (2025) employs a localization agent to prune key visual regions, co-trained with a behavioral policy agent. The output sequence of localization coordinates provides spatially explicit explanations while significantly reducing computational costs and improving robustness to noise interference, especially suitable for visual tasks like Atari games and real robotic arms. However, this hierarchical training process is complex, and its advantages may not be significant in simple tasks.

Collectively, attention and feature attribution methods significantly lower the cognitive barrier to understanding complex policies by visualizing or quantifying the contribution of input features to decisions. To comprehensively compare their different implementation paths, Table 10 systematically summarizes the core mechanisms, key tools, and typical application scenarios of various methods.

Nevertheless, these approaches are often critiqued for providing “superficial” explanations, with three primary limitations standing out. First, attention weights and feature importance scores do not, by themselves, establish causal drivers of decisions—a concern raised by Schier et al. (2025) and Cheng et al. (2025)—which can render their practical interpretation misleading. Second, they generally fail to uncover the agent’s internal strategic reasoning process. Finally, they offer little insight into the complex coordination mechanisms that emerge among multiple agents.

To advance the field, a concerted effort is needed to move beyond correlation-based descriptions and toward causal, semantic, and rigorously validated explanations. Promising research directions include:

- **Developing Causal Interpretation Frameworks.** Future work could aim to formalize causal understanding by integrating preliminary causal explorations (e.g., Motokawa and Sugawara (2023c)) with established causal discovery models. Such integration would help test and characterize possible causal mechanisms underlying agent behavior, distinguishing them from mere correlational associations.
- **Achieving Semantically Grounded Explanations.** To deepen explanatory power, research must link attribution results to high-level, intervenable concepts (e.g., tactical objectives). Building on methods like those of Yang et al. (2025b) and Schier et al. (2025), the goal is to answer *why* certain information is attended to, moving beyond merely identifying *what* is attended to.
- **Establishing Standardized Evaluation Protocols.** The community urgently requires standardized benchmarks to rigorously assess the practical utility and robustness of post-hoc explanations (e.g., feature attribution and saliency maps) and to guard against “spurious explanations.” Efforts can build upon existing analyses of explanation fidelity (Cheng et al. 2025). Parallel to this, the pathway of designing *intrinsically interpretable models* (Schier et al. 2025) should be pursued as a complementary approach for building transparent and trustworthy decision-making systems.

4.2.3 Counterfactual reasoning and influence analysis

This category of methods aims to answer the core question: “*Why was this specific action chosen over alternatives?*” Their technical essence lies in constructing counterfactual scenarios (e.g., “what would have happened if a different action had been taken?”) to infer the causal impact of a decision or quantify an individual agent’s contribution to the global system state. Residing at the *causal layer* of the explainability framework, these methods are dedicated to revealing the contrastive nature and causal mechanisms underlying behavioral choices. However, they commonly face challenges such as high computational complexity and sensitivity to assumptions about the model and data. This approach serves both as a deepening of feature attribution (which deals with correlation) and as a provider of the causal foundation necessary for motivational analysis (goal-driven behavior). Currently, research in this domain primarily unfolds along the following three paths:

Latent variable inference and intention inference

This line of work employs probabilistic generative models to infer hidden drivers behind observed behavior. Jiang et al. (2024) introduces variational inference within a soft actor-critic framework, combined with Monte Carlo methods to derive a counterfactual policy. The influence of latent variables is quantified by computing the Kullback–Leibler (KL) divergence between the factual and counterfactual policies, which is then shaped into an intrinsic reward. This process provides a probabilistic mechanism for analyzing behavior in environments with non-stationary rewards and partial observability, enhancing exploration efficiency in tasks such as predator–prey. However, its performance is significantly affected by hyperparameters like the intrinsic reward weight, and the computational overhead exceeds that of baseline algorithms. Taghavi and Farnoosh (2025a) integrates the Expectation-Maximization (EM) algorithm with MARL. In the E-step, the posterior distribution of latent variables is estimated via variational inference or Monte Carlo sampling; in the M-step, agent policies are optimized based on this distribution. While this method improves decision-making in partially observable settings, the EM iterations increase computational costs, making it difficult to scale to large-scale scenarios involving hundreds or more agents.

Quantitative assessment of individual causal contribution

The core of this approach is to directly quantify the causal impact of a single agent’s action on the global state through counterfactual contrast, using this measure to drive individual learning. A representative method, proposed by Li et al. (2024f), employs a Conditional Variational Autoencoder (CVAE) to learn a compact latent representation of the global state. It separately models the “**factual**” and “**counterfactual**” state transition distributions. The individual causal contribution of each agent’s action is then measured by computing the Bayesian surprise (KL divergence) between these two distributions, which forms an intrinsic exploration scaffold. This method directly provides agents with exploration signals grounded in causal relationships, intended to mitigate the multi-agent credit-assignment problem and reduce sensitivity to environmental noise. It has been shown to improve efficiency in sparse-reward cooperative exploration tasks. Nonetheless, this approach requires training an additional generative model, leading to high computational complexity, and its counterfactual contrast is typically limited to single-step state transitions.

Automated explanation via simulation intervention and LLM fusion

An emerging frontier seeks to completely decouple explanation generation from model internals by constructing an external interaction framework for automated causal expla-

nation. A typical paradigm is presented by Gyevnár et al. (2025), which drives a LLM to engage in multi-round interactions with an environment simulator based on predefined query templates. The LLM performs counterfactual interventions on the simulator (e.g., using *what-if*, *remove* queries) to obtain comparative information. Combined with macro-action abstractions of agent behavior, it autonomously synthesizes natural language explanatory texts containing explicit causal relationships and counterfactual comparisons. This method generates explanations in autonomous driving simulation scenarios that are intuitive, causally strong, and human-friendly. Its model-agnostic nature makes it suitable for post-hoc auditing in complex interactive scenarios. However, its effectiveness heavily depends on the accuracy of the simulator and the LLM's reasoning and planning capabilities. Multi-round interactions may introduce robustness issues, and the computational cost remains high.

Collectively, these technical trajectories demonstrate the progressive maturation of counterfactual reasoning methods from theoretical causal inference frameworks toward practical explanatory tools capable of addressing real-world interpretability challenges. The evolution from latent variable modeling through quantitative causal assessment to narrative explanation generation reflects an increasing emphasis on producing not just technically sound but also human-meaningful causal explanations, representing a critical advancement in the quest for transparent and trustworthy autonomous decision systems. Table 11 systematically compares these methods.

Future research should therefore advance along the following critical paths:

- **Enhancing Computational Efficiency.** Develop optimized variational inference algorithms inspired by recent work (Jiang et al. 2024; Taghavi and Farnoosh 2025a) and create lightweight simulators based on emerging paradigms (Gyevnár et al. 2025) to mitigate computational overhead and improve accessibility.
- **Ensuring Robustness and Reliability.** Establish rigorous theoretical frameworks for assessing the plausibility of counterfactual scenarios and devise methods to reduce the dependency on pristine observational data in approaches such as Li et al. (2024f).
- **Expanding Explanatory Scope.** Transcend single-step, individual-centric analysis by developing frameworks capable of multi-step temporal reasoning and causal explanation of multi-agent interactions, thereby capturing the complex causal structures emerging from coordinated behaviors.

4.2.4 Intrinsic motivation and goal analysis

This category of methods addresses the profound question: “*What is the intrinsic driving force behind the agent’s behavior?*” The technical core lies in directly modeling and inferring an agent’s intrinsic motivation, high-level goals, or strategic intent by introducing human-understandable semantic concepts, interpretable reward parameters, or structured goal representations. Situated at the **semantic and intentional layer** of the explainability framework, these methods build complementary explanatory perspectives ranging from *behavioral logic* to *behavioral purpose* with the preceding three categories. They are particularly suited for scenarios such as human-agent collaboration, value alignment, and strategic explanation. Based on their implementation mechanisms, existing methods can be categorized into the following three mainstream paradigms:

Concept Bottleneck models

These methods embed interpretable concepts within the model architecture. The approach proposed by Zabounidis et al. (2023) inserts a “**concept layer**” into the policy network, forcing the network to first predict a set of expert-predefined semantic concepts from raw observations and then generate actions based on these concepts, forming a transparent “*observation* → *concept* → *action*” decision pipeline. It provides a human-understandable concept layer and supports real-time intervention but heavily relies on an expert-defined concept library, and the expressiveness of the concepts may limit policy performance. Gruen et al. (2022) extends this by introducing *test-time concept intervention analysis* techniques to quantitatively assess the causal impact of a concept on collaboration. This method can be used to identify key conceptual factors in multi-agent collaboration and detect lazy agents, providing a conceptual-level diagnostic tool for collaboration auditing.

Interpretable reward functions and parameterized personalities

Behavioral preferences or goal-related tendencies are represented through structurally interpretable reward functions or personalities with semantically explicit parameters. The *NetEvolve* method proposed by Miyake et al. (2024) learns a reward function for each node agent in a social network. Its parameters directly correspond to socio-psychological concepts, thereby providing an interpretable account of the modeled preferences underlying a node’s actions like forming connections or changing attributes. The model is highly interpretable, with few, semantically meaningful parameters, and reports strong predictive performance on both synthetic and real social network data. However, it relies on predefined sociological features and may fail to capture more complex network dynamics. Huang et al. (2025) proposes a *Multi-Persona Multi-Agent Meta RL* framework. Its core technology involves using meta-learning to train multiple personality policies with different “*exploration-exploitation*” preferences for each agent, quantifying personality as an adjustment to interpretable weight parameters in the reward function. Agents dynamically activate personalities via a persona selector, demonstrating high adaptability, coordination, and robustness to communication instability in dynamic traffic systems such as parking optimization and charging scheduling. Personality labels provide high-level semantic explanations. However, the number of personalities and their weights require careful design, and meta-training incurs high computational costs.

Motivation modeling based on structured external priors

This class of methods provides semantically grounded explanations for motivation by introducing strongly structured external knowledge frameworks (e.g., linguistic systems, physical laws, formal logic). Based on the type of prior knowledge, it can be further subdivided into three paths:

Semantic Alignment Path: Represented by Toquebiau et al. (2025), this path introduces structured language as a supervisory signal, forcing alignment between the agent’s internal representations and human semantic concepts through joint optimization, thereby achieving verbalizable explanations of motivation. Its advantages are intuitive explanations and potential generalization improvement, but it depends on high-quality linguistic prior data.

Physical Decomposition Path: Represented by Li et al. (2025b), this path introduces classical physical laws as a modeling framework, decomposing raw states into physics concepts like *force* and *distance* and synthesizing actions based on mechanical models. Its explanations are highly intuitive but strictly depend on the match between the task and the physical metaphor.

Formal Verification Path: Represented by Taghavi and Farnoosh (2025b), this path introduces formal logic and model checking tools, abstracting the policy into a mathematically verifiable model to provide explanations with formal guarantees regarding the achievability of specific motivations or goals. This path can provide strong guarantees for explicitly specified properties but is computationally complex and currently applicable to limited scenarios.

Collectively, intrinsic motivation and goal analysis methods lay the foundation for human-agent trust and value alignment by semantically modeling the intrinsic drives of agents. To systematically summarize their different implementation paradigms, Table 12 provides a comprehensive comparison from the perspectives of core mechanism, explanation form, and applicable scenarios.

However, the principal challenge confronting these methods is their substantial dependence on prior knowledge. Manually curated concept sets (Zabounidis et al. 2023; Toquebiau et al. 2025), reward function structures (Miyake et al. 2024), and personality definitions (Huang et al. 2025) not only demand significant expert effort but also risk introducing designer bias and constraining generalization.

Transformative advances will therefore require progress along several critical pathways:

- **Automated Semantic Concept Acquisition.** Future research should prioritize the dynamic acquisition of semantic concepts from human demonstrations, natural language feedback, or environmental interactions. This would reduce dependence on manual definitions, such as those employed in Zabounidis et al. (2023) and Toquebiau et al. (2025), while enhancing methodological generalization and scalability.
- **Dynamic and Hierarchical Explanation Frameworks.** Next-generation systems should generate context-aware explanations across multiple abstraction levels, adapting to different users and scenarios. Building upon comprehensive frameworks like Taghavi and Farnoosh (2025b), this includes investigating the dynamic evolution of motivations (Huang et al. 2025) to explain long-term preference shifts.
- **Unified Dynamic Motivation Modeling.** Ultimately, integrating these research threads could enable systems capable of interpreting the temporal evolution of intrinsic motivations, representing a paradigm shift from static interpretation to dynamic motivational understanding.

By pursuing these interconnected directions, the field may achieve a coherent framework for modeling and explaining the rich dynamics of intrinsic motivation in artificial agents.

4.2.5 Synthesis and future directions for micro-level individual behavior

This section provides a comprehensive summary and comparative analysis of the four technical paradigms for the explainability of individual agent behavior. It then synthesizes their intrinsic connections into a cohesive, multi--perspective interpretability framework. Finally, it discusses current limitations and outlines promising future research directions.

Comprehensive summary of technical paradigms

This section has systematically examined the four technical paradigms for micro-level agent behavior explainability. A comprehensive comparison of these paradigms is presented

in Table 4, which contrasts their respective technical categories, core explanatory questions, advantages, and limitations.

Policy abstraction and model extraction 4.2.1 interpretable *symbolic models* (e.g., decision trees, behavior trees) from black-box policies (Milani et al. 2022b; Wan et al. 2025a; Gimenez-Abalos et al. 2025), offering transparent logic for inspection (Hu et al. 2024c) but often struggling with high-dimensional states and manual design dependency. **Attention and feature attribution mechanisms** 4.2.2 visualize an agent's **perceptual focus** via heatmaps or feature attributions (Motokawa and Sugawara 2023b; Schier et al. 2025), lowering cognitive load but typically providing correlational, not causal, insights. **counterfactual reasoning** 4.2.3 constructs *hypothetical scenarios* to uncover causal mechanisms behind action choices (Gyevnár et al. 2025), though it is computationally costly and data-sensitive. **intrinsic motivation analysis** 4.2.4 analysis models drives using semantic concepts or interpretable rewards (Zabounidis et al. 2023; Miyake et al. 2024; Huang et al. 2025), delivering deep semantic explanations at the cost of heavy prior knowledge reliance and limited generalization.

Intrinsic Connections: A Multi-Perspective Explanatory Framework

The four paradigms are not isolated; they form a naturally complementary spectrum in explanatory depth and focus, constructing a holistic explanatory chain from “behavioral manifestation” to “behavioral motivation”. This systematic framework is manifested in three aspects:

- **Complementary Explanatory Perspectives:** Explanatory depth increases progressively. *Policy abstraction* describes “*what* the behavior is” (rules), *feature attribution* reveals “*on what* the behavior depends” (inputs), *counterfactual analysis* explains “*why* the behavior occurred” (causality), and *motivation analysis* clarifies “*for what purpose* the behavior is pursued” (goals).
- **Taxonomic Coverage:** The four methods provide the primary theoretical perspectives for explaining individual decision-making behavior from four dimensions: logic (*How*), perception (*What*), causality (*Why*), and motivation (*For What*).
- **Functional Complementarity:** They mutually supplement each other to approximate a complete understanding of agent behavior. For instance, explaining an agent's lane-changing decision at an intersection may require: first, consulting a policy graph (policy abstraction) to see its standard rule for such states (*logic*); second, using an attention heatmap (feature attribution) to inspect whether attention is assigned to the trailing vehicle (*perception*); third, employing counterfactual reasoning to verify if it would have stayed in lane *if* no trailing vehicle were present, assessing the causal necessity of that focus (*causality*); and finally, combining its reward function parameters (motivation analysis) to infer whether the behavior stems from a deep-seated preference for “efficiency-first” or “safety-first” (*motivation*).

Limitations and future directions

Current micro-level explanation methods (e.g., policy abstraction, attention mechanisms), while providing intuitive local insights, face fundamental constraints in explanatory depth and generalization. Most methods rely on manual prior design, making automated adaptation to complex tasks difficult. Moreover, they primarily reveal statistical associations between input features and decisions rather than causal mechanisms, compromising

Table 4 Comprehensive comparison of micro-level individual behavior explainability techniques

Technique category	Core explanatory question	Representative techniques	Explanatory focus	Core advantages	Inherent limitations & challenges
Policy Abstraction & Model Extraction (4.2.1)	What decision rules does the agent follow?	Transparent model design Luvisutto et al. (2025); LLM-based program synthesis for automated discovery (Castro et al. 2025a, b); Rule distillation (Milani et al. 2022b; Xi-Jia et al. 2025); Hierarchical representation (Domènech I Vila et al. 2022; Wan et al. 2025a); Semantic mapping (Ji et al. 2023; Hu et al. 2024c; Reiestad 2025)	Transparency and structure of decision logic	Intuitive explanations with clear logical chains, supporting decision tracing and comparison; Emerging automated methods may reduce reliance on manual design and generate inspectable symbolic models or testable hypotheses	Poor adaptability to high-dimensional/continuous states; Reliance on manually designed predicates/concepts; Model distillation may introduce errors; Automated methods face extremely high computational costs and generated code readability challenges
Attention & Feature Attribution (4.2.2)	What information does the agent focus on when deciding?	Conditional/instruction-modulated attention (Motokawa and Sugawara 2023b, c); Multi-perspective/evolution analysis (Motokawa and Sugawara 2023a; Chen et al. 2024); Prior fusion (Liu et al. 2025a; Yang et al. 2025b); Hybrid model attribution (Schier et al. 2025); Hard attention (Cheng et al. 2025)	Perceptual basis and information importance for decisions	Intuitive visualizations that reduce cognitive load and are useful for diagnosing perceptual errors	Mostly correlational rather than causal explanations; May produce misleading attributions; Difficulty in revealing deep reasoning logic
Counterfactual Reasoning & Influence Analysis (4.2.3)	Why was this action chosen over another?	Latent variable inference (Jiang et al. 2024; Taghavi and Farnoosh 2025a); Individual contribution quantification (Li et al. 2024f); Simulation+LLM automated explanation (Gyevnár et al. 2025)	Causal mechanism and criticality of actions	Provides causal and contrastive explanations to probe behavioral motives	High computational complexity; dependence on model assumptions and the quality or sufficiency of available data; challenges in ensuring reasonableness of counterfactual scenarios

Table 4 (continued)

Technique category	Core explanatory question	Representative techniques	Explanatory focus	Core advantages	Inherent limitations & challenges
Intrinsic Motivation & Goal Analysis (4.2.4)	What is the intrinsic driving force behind the agent's behavior?	Concept bottleneck models (Zabounidis et al. 2023; Grupen et al. 2022); Interpretable reward/personality (Huang et al. 2025; Miyake et al. 2024); Semantic supervision alignment (Toquebiau et al. 2025); Physics-inspired decomposition (Li et al. 2025b); Formal verification and model checking (Taghavi and Farnoosh 2025b)	Semanticized motivations and preferences behind behavior	Deep, semantically rich explanations that directly enhance human-AI trust	Strong dependence on prior knowledge (concepts, goal templates); limited generalization; automation remains limited in many approaches

reliability in safety-critical applications. To overcome these limitations, future research should focus on the following specific technical pathways:

- **Automated Concept Discovery:** Fusing representation learning with the semantic understanding capabilities of LLMs to enable automated or semi-automated generation of high-level predicates, subgoals, and intrinsic motivation functions. This reduces dependency on domain knowledge and enhances scalability, extending automation-related approaches such as those explored by Xi-Jia et al. (2025); Gyevnár et al. (2025).
- **Lightweight Causal Integration:** Developing low-overhead, online causal discovery modules that can be co-trained with policy networks. Deeply embedding theoretical frameworks like Structural Causal Models into the learning process can enable real-time, verifiable causal attribution for decisions, driving a fundamental shift in explanation paradigms from “correlation” to “causation” (Jiang et al. 2024).
- **Interactive Explanation Frameworks:** Designing interactive systems capable of dynamically generating multi-granularity explanations based on user roles and real-time queries (Zhang et al. 2024b). Establishing quantitative evaluation protocols covering fidelity, robustness, and cognitive efficiency is crucial to steer the field towards standardization and practical utility.

In conclusion, through coordinated advancement along these research directions, explainability techniques for micro-level individual behavior are poised to evolve from auxiliary analytical tools into core components for constructing transparent, trustworthy, and high-performance agent policies.

4.3 Explainability techniques for agent interaction and collaboration mechanisms

The intelligence of a MAS stems not only from the individual capabilities of its constituent agents but is also rooted in the complex and dynamic interactions and collaborations among them. Therefore, to truly understand and trust an MAS, the key lies in systematically examine its collaborative mechanisms. In this review, explanations of agent interaction and collaboration are organized around five complementary questions regarding collaborative

behavior (as shown in Fig. 5): (1) the **Information Structure**—on what information do collaborative decisions depend? Whom do agents attend to, and how is information structured, flowing, and integrated? (2) the **Communication Semantics**—how do agents engage in explicit ‘dialogue’? Through what specific protocols, content, and timing is micro-level coordination achieved? (3) the **Intrinsic Logic**—is an observed collaborative pattern driven by causality or merely a spurious statistical correlation? What are the causal chains between actions? (4) the **Organizational Form**—does the system exhibit stable functional social structures? What are the role divisions and the enduring patterns and norms of collaboration? and (5) the **Incentive Root**—How does the mechanism of rewarding collective achievement incentivize collaboration within the team?

These five questions provide an organizing framework for surveying explainability in agent interaction and collaboration, progressing from intuitive information flows to underlying incentive structures. To address them, research has coalesced around five complementary technical paradigms: **Attention and Relational Modeling** for revealing informational foundations, **Communication Protocol Parsing** for interpreting interaction semantics, **Causal Reasoning** for validating interaction logic, **Role and Behavioral Pattern Discovery** for uncovering organizational structures, and **Team Value Decomposition** for explaining incentive roots. Together, these paradigms provide complementary perspectives on interaction- and collaboration-level explanation. The following sections detail how they

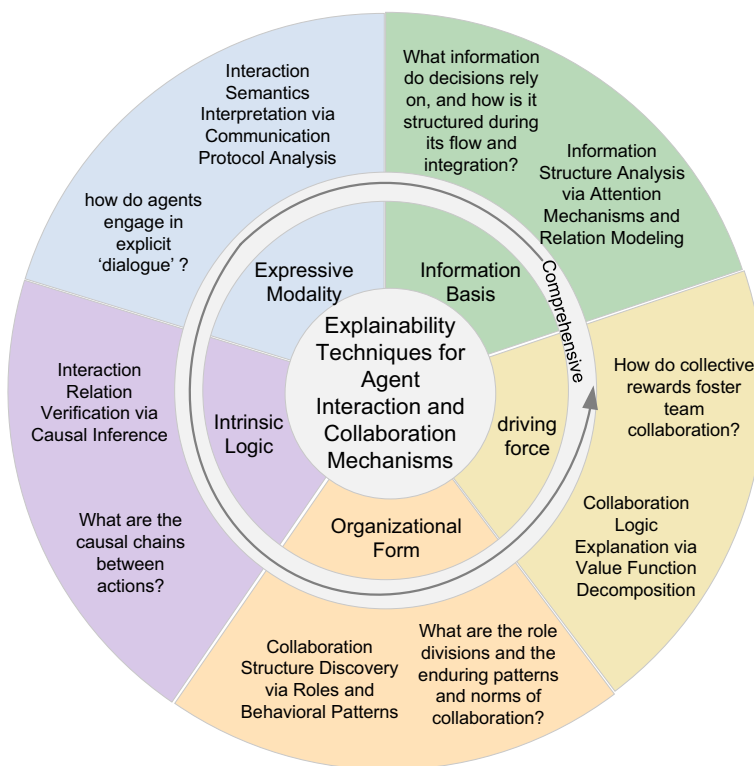


Fig. 5 Technical framework for explaining intelligent agent interaction and collaboration: core issues and five types of technologies

collectively address the above questions, providing a unified lens for explaining multi-agent collaboration.

4.3.1 Information structure analysis via attention mechanisms and relational modeling

This category of methods approaches the mechanistic interpretation of collaboration from an **information-centric perspective**, aiming to address the core questions: “How does information flow during collaborative decision-making, and whose information is more critical?” Its technical essence lies in visualizing the distribution of attention weights within neural networks or explicitly learning and constructing relational graph structures among agents. This reveals the underlying information dependencies and interaction topologies that govern collaborative behaviors. As an intuitive and commonly used technical paradigm, it provides a *structural snapshot* of the collaborative system, uncovering agents’ real-time attention preferences and interaction weights. It is therefore commonly regarded as foundational input for subsequent in-depth analyses, such as causal verification or semantic interpretation.

To systematically elucidate the informational foundations of collaboration, this line of inquiry is pursued along two primary pathways: first, the **multi-perspective analysis of attention weights**, which aims to translate implicit information focus into interpretable quantitative metrics; and second, the **structured construction of inter-agent relationships**, which aims to explicitly characterize interaction topologies through graphs or matrices. Together, these constitute the methodological pillars for understanding “*how information flows*” in collaborative settings.

Multi-perspective analysis of attention weights

This research direction focuses on analyzing or designing attention mechanisms to transform the implicit information preferences exhibited by agents during decision-making into explicit weight distributions. This reveals the informational structural basis of “who to attend to” and “why it matters” in collaboration. Based on differing research emphases, this can be further subdivided into three streams: the design and improvement of attention mechanisms, methodological research on the interpretability of attention itself, and visualization and analysis techniques for attention weights.

Design and improvement of attention mechanisms. Research in this stream is dedicated to proposing novel attention architectures whose core mechanisms are intrinsically interpretable. For instance, Sun et al. (2023) designed a *two-stage graph attention mechanism* that employs hard attention to filter communication partners into a sparse subgraph, followed by soft attention to compute continuous contribution weights. This architecture inherently provides explanations for communication topology and contribution levels in air combat simulation environments and significantly reduces communication overhead, though its computational complexity may become a bottleneck in large-scale swarms. Qi et al. (2022) proposed the *Adaptive Gated Graph Attention Network (AGGAT)*, which models interactions through a “coarse-to-fine” three-level cascaded attention mechanism (adaptive neighbor selection, standard weight attention, and gated attention). Its interpretability is manifested as multi-level binary neighbor matrices, continuous weights, and gating signals, effectively filtering information noise in environments like traffic intersections, albeit at a high computational cost. Li et al. (2022) focus on time-varying composite action spaces,

using Graph Attention Networks to explicitly model dependencies among sub-actions. The attention weights provide an explanation for heterogeneous action associations, offering a first systematic treatment of this problem, though the framework's complexity results in significant computational overhead.

To further enhance the transparency of credit assignment and improve policy performance, researchers have designed novel paradigms that integrate attention. Chen et al. (2025b) introduced attention mechanisms (dual attention) concurrently within an Actor-Critic framework, enabling the Actor to aggregate collaborative information via cross-attention while the Critic performs precise credit assignment. This mechanism provides an interpretable perspective on collaboration relevance in Multi-Agent Particle Environment (MPE) cooperative tasks and converges rapidly, though training time increases compared to baseline methods. Li et al. (2023) proposed constructing an *intrinsic reward network via attention mechanisms (AIR)*, which uses attention weights to calculate inter-agent correlations and generate intrinsic rewards. In heterogeneous StarCraft Multi-Agent Challenge (SMAC) scenarios, it provides an intuitive basis for credit assignment; however, its assumption that collaborating agents share similar observations and behaviors limits its generalizability. Shah et al. (2024) leveraged a Decision Transformer architecture, incorporating an *Agent Priority Module (APM)* that utilizes attention weights to compute individual contributions, thereby producing explicit attention-based contribution estimates. This approach balances safety and interpretability in SMAC and Multi-agent MuJoCo environments, though its scalability to very large-scale scenarios requires further validation.

Research on the interpretability of attention mechanisms. This stream systematically investigates the explainability of attention mechanisms themselves. Kailas et al. (2025) conducted a comprehensive exploration, evaluating the applicability of three post-hoc GNN explainers—*GraphMask*, *GNN-Explainer*, and *Attention Explainer*—in multi-agent collaboration. They also proposed *attention entropy regularization* to enhance explanation faithfulness. This method is simple, effective, and does not compromise task performance; however, existing explainers like GraphMask are limited by hard binary masks, making it difficult to capture diffuse interaction influences. Itaya et al. (2023) extended the Multi-Agent Actor-Critic (MAAC) framework by introducing independent attention heads for each Actor, enabling intuitive and comprehensible explanations for collaborative decisions without requiring additional backpropagation and without degrading task performance. A limitation is that its computation relies on the states of all agents, restricting its use in fully decentralized scenarios.

Visualization and analysis techniques for attention weights. This body of work focuses on transforming learned attention weights into human-intelligible explanatory outputs through various technical means, serving as a bridge between the model's internal state and user cognition. Common forms include heatmaps, evolution graphs, and topology overlay visualizations. For example, Hao et al. (2023) fused mean-field theory with Graph Attention Networks (GAT), supporting large-scale scenarios and providing an intuitive interpretation of interaction weights. A limitation is its reliance on the assumption of a static graph with fixed agent positions, making it unsuitable for scenarios with dynamic positional changes. Chen et al. (2025d) utilized GNN attention mechanisms to automatically discover interpretable dependencies among sports facilities, achieving low communication overhead and good scalability. However, it assumes perfect sensor data, and its interpretability is confined to facility associations rather than fine-grained decision logic. Lei et al. (2025)

employed a **multi-head attention mechanism** to capture multi-perspective relationships, offering clear collaborative interpretability and good scalability. Its drawbacks include reliance on a simplified 2D motion model and explanations that do not delve into fine-grained individual action logic. Du et al. (2022) revealed agent preferences for neighbors and relationship types through *hierarchical attention weights (agent-level and relationship-level)*, adapting well to heterogeneous agents and complex relational scenarios. However, its interpretability analysis tends to be qualitative, lacking systematic evaluation metrics.

Structured construction of explicit relationship graphs

This research direction emphasizes the explicit construction and visualization of interaction relationship structures among agents. Based on the representation form and construction mechanism, it is primarily divided into two major categories: **relationship construction based on graphs** and **relationship modeling based on weight matrices**.

Relationship construction based on graphs. These methods treat agents as nodes and model their interaction relationships as edges in a graph, intuitively presenting the collaboration topology through the graph structure. Based on construction methodology, they can be further classified into **static prior graphs**, **dynamic learned graphs**, and **complex structure graphs**.

Static graph construction based on domain knowledge yields structures with clear physical meaning. For example, Han et al. (2022b) designed a specialized *Air Combat Relationship Graph (ACRG)*, using a Graph Attention Network to explicitly model key synergistic relationships. This enhances policy interpretability and decision efficiency; however, the construction of the ACRG relies on a predefined graph attention mechanism, which may fail to capture more complex, multi-level dependencies.

Dynamic graph structures offer better adaptability to complex environments. Fu et al. (2024) constructed a *Hierarchical Scalable Collaboration Graph (ECG)* whose dynamic topology provides intuitive explanations for hierarchical synergy and supports zero-shot transfer. A drawback is that its collaborative actions require manual design, and the training complexity for graph operators is high. Peng et al. (2025a) integrated evolutionary algorithms with GNNs to model dependencies, enhancing exploration through evolution and improving collaborative interpretability via GNNs. This approach, however, incurs significant computational overhead, and the GNN's explanations are contingent on the accuracy of the learned graph structure. Pesce and Montana (2023) proposed Connection-Driven Communication (CDC), which uses heat kernels to simulate information diffusion. Its communication patterns are intuitive and adapt to varying numbers of agents, but the heat kernel computation is complex, and threshold parameters require careful tuning. Liu et al. (2022a) employed Temporal Convolutional Networks (TCNs) to learn dynamic edge weights, capable of dynamically capturing temporal collaboration intensity. The resulting weight heatmaps are intuitively interpretable; however, its receptive field parameters must be adjusted based on task characteristics, and its adaptability to tasks lacking distinct temporal features is limited. Duan et al. (2024b) dynamically infer collaboration relationships via an *Implicit Temporal Sparse Coordination Graph (LTS-CG)*, leveraging trajectories to capture temporal dependencies with strong interpretability. Limitations include the need for manual tuning of the graph loss weight and potential inefficiency in processing very long trajectories. Li et al. (2025a) formed "*information groups*" through a dynamic directed graph model, adapting well to large-scale agent scenarios with low training cost. However, the scalability of its adjacency matrix is limited, and its robustness in communication-constrained scenarios

requires further verification. Liu et al. (2023b) introduced a *Deep Hierarchical Communication Graph (DHCG)* whose acyclic structure ensures convergence and reduces redundant communication, with interpretable intention propagation. In large-scale scenarios, however, intention sharing can be time-consuming, and the Directed Acyclic Graph (DAG) projection adds extra computational cost.

To model group-level or higher-order collaboration, researchers have introduced complex graph structures. Zhu et al. (2024) were the first to introduce *hypergraphs* into MARL for communication modeling. The hypergraph structure offers strong modeling capacity and efficient communication, but hypergraph construction is computationally expensive. Fan et al. (2025) proposed a *Factor-based Multi-agent Transformer*, which uses factor graphs to explicitly represent group-level relationships, significantly reducing computational complexity. A limitation is that factor partitioning relies on prior knowledge. Zhang et al. (2025b) designed a learnable communication graph combined with collaborative coordinate attention for path planning, achieving high structural and semantic clarity. The technology stack is complex, however, leading to considerable computational overhead. Ren et al. (2023) employed a *Hierarchical Graph Attention Network (HGAT)* with Gated Recurrent Units to aggregate spatiotemporal information, adapting to heterogeneous environments with transparent exploration strategies. Its network structure is complex, and the depth of interpretability heavily depends on the quality of visualization.

Relationship and incentive sharing modeling based on weight matrices. These methods forgo explicitly constructing interaction graphs. Instead, they learn an interpretable weight matrix that directly defines the attention or dependency relationships among agents for reward or loss allocation. This embeds relationship modeling deeply within the incentive mechanism, achieving a synergistic optimization of collaboration efficiency and interpretability. For instance, Gemp et al. (2022) proposed an individual loss adjustment mechanism based on a learnable mixing matrix. Here, each agent maintains a row-stochastic matrix to form a convex combination of its own loss with the losses of other agents, dynamically optimizing system efficiency (measured via the price of anarchy) in mixed-motive settings. The weights of the mixing matrix directly and quantitatively reveal each agent's trade-off between self-interest and the interests of others, explaining the social preference structure within the collaboration. Similarly, the *creed model* proposed by Radke et al. (2023) uses a parameterized vector to explicitly regulate an agent's attention weights on individual, team, and system-level rewards. Its creed parameters provide a clear semantic explanation for the agent's collaborative motivations and can guide role specialization in scenarios with incomplete common interests. Such methods model and explain relationships from the very source of incentive allocation, offering more direct and quantitative analytical tools for understanding the interest structure and driving forces behind collaboration.

Collectively, **information structure analysis techniques based on attention mechanisms and relationship modeling** provide a direct pathway to interpretability concerning “*how information flows and focuses*” during collaboration, through the visualization of weights and the construction of explicit graph models. To systematically delineate their characteristics, a comparative analysis of representative studies is presented in Table 13, covering technical approaches, core mechanisms, advantages, limitations, and applicable scenarios. However, significant challenges persist. First, as indicated by Kailas et al. (2025), high attention weights or connections indicate correlation rather than causation, raising concerns about explanation faithfulness. Second, despite attempts like Hao et al. (2023)

to visualize large-scale systems, computational complexity presents substantial scalability challenges. Finally, current methods (Han et al. 2022b) often require manual mapping of low-level numerical weights to high-level tactical semantics, impeding effective collaboration with domain experts.

Future research should pursue several promising directions:

- **Develop Theoretical Foundations and Verification Methods:** Explore integrating theories such as causal inference (Briglia et al. 2025) to build theoretical explanation frameworks atop attention mechanisms, distinguishing correlation from causation, and enhancing the reliability of information structure explanations for models like Sun et al. (2023), Qi et al. (2022).
- **Design Lightweight and Efficient Algorithms:** To overcome the scalability challenges identified in Hao et al. (2023) and Lei et al. (2025), future work should investigate efficient attention approximation algorithms and simplified dynamic graph models to facilitate real-time analysis of information structures in large-scale, online systems.
- **Promote Semantic Fusion and Automatic Annotation:** Investigate the combination of Natural Language Processing (NLP) and LLMs with visualization techniques to automatically assign high-level semantic labels to interaction patterns discovered by methods like Han et al. (2022b), thereby bridging the gap from “numerical weights” to “tactical concepts”.

4.3.2 Interaction semantics interpretation via communication protocol analysis

This category of methods approaches the explanation of collaboration mechanisms from the perspective of **expression modalities**, aiming to answer the question: What specific ‘dialogue’ semantics do agents employ to coordinate? Its technical core involves analyzing the content, sender, receiver, and timing of communicated messages to interpret the **language** established among agents for achieving collaboration. It forms a complementary relationship with information structure analysis at the surface and deep levels: the former focuses on “*how information flows*”, while the latter focuses on “*what the information content is*”.

Centered on the interpretation of **dialogue** semantics, related research progresses along three main fronts: first, the information-theoretic optimization and explanation of communication content to enhance message utility and comprehensibility; second, the structural analysis of communication patterns and protocols to clarify the rules and frameworks of coordination; and third, promoting the transparency of communication semantics and integration with natural language to bridge the human-machine semantic gap. These fronts collectively aim to answer *how agents coordinate through dialogue*.

Information-theoretic optimization and explanation of communication content

This direction focuses on optimizing and interpreting the communication messages themselves. Message distillation based on the *information bottleneck* is a significant theoretical foundation for achieving efficient and robust communication. Yang et al. (2024) employ an information bottleneck to optimize communication, enabling dynamic team formation and heterogeneous channels. However, the high computational complexity of its team reasoning strategy may limit scalability in large-scale environments. Ding et al.

(2023b) applied the *Graph Information Bottleneck (GIB)* to multi-agent communication, resulting in strong robustness and efficient communication representation. A drawback is its strong dependency on graph structure, and the information bottleneck optimization may increase training complexity.

Dynamic weight analysis of communication content constitutes another important line of work. Guan et al. (2022) achieved efficient message aggregation and good scalability through *Self-supervised Information Aggregation (MASIA)*; however, its performance may be unstable in sparse reward scenarios. Zhang et al. (2025c) implemented information completion via a *Pluggable Adaptive Generation Network (PAGNet)*, significantly improving collaboration efficiency. Limitations include potential instability in training the generative model and limited scalability in extremely large-scale scenarios.

Intent-driven communication decisions explicitly link communication acts to internal states. He et al. (2025b) realized real-time intent-driven communication through an Intent-Correlation Network (ICN), enhancing training stability and efficiency. Its limitations include reliance on predefined influencing factors and high computational complexity. Yuan et al. (2022) generated incentivizing messages via decentralized teammate modeling, avoiding policy space explosion and improving efficiency through sparse communication. However, the teammate modeling scales poorly to very large numbers of agents, and it relies on simple IDs to distinguish agents.

Structured analysis of communication patterns and protocols

This direction focuses on providing a structured characterization of communication behavior itself. **Explicit communication matrices and graph structures** represent the most direct approach. Meng and Tan (2024) achieve personalized communication through linear-complexity modules, drastically reducing computational cost. Its drawbacks include a simplistic ID embedding method and reliance on a fixed number of communication rounds. Similarly, Hu et al. (2024b) employed a *bi-level optimization* to automatically search for an optimal communication topology, matching the performance of a fully-connected graph in several tasks. However, its training requires bi-level optimization, incurring high computational cost, and the static graph may not adapt to dynamic topology changes. Shao et al. (2022) enhanced zero-shot generalization ability and communication efficiency through a *self-organizing group* mechanism, but it relies on perfect communication channels, and the group size in random elections depends on a hyperparameter.

Some research aims for interpretable coordination by having agents follow pre-designed protocols. Sheng et al. (2022) proposed the *LSC framework* combined with a Cluster-Based Routing Protocol (CBRP), achieving high communication efficiency and strong scalability. However, its weight generator requires additional training, and the CBRP protocol is non-differentiable. Selimović et al. (2024) approached the problem from the perspective of distributed algorithms, designing a consensus-based collaboration mechanism where decisions are based on group consensus, thereby enhancing interpretability. Its assumptions of a fully connected communication graph may not hold in practice, and the perception confidence threshold requires manual adjustment.

Transparent communication semantics and natural language integration

This direction is dedicated to directly enhancing the semantic interpretability of the communication content itself. The advent of **LLMs** has brought revolutionary progress to semantic communication. For example, Li et al. (2024b) leveraged an LLM to generate human-readable natural language messages, producing human-readable communication

intended to improve semantic transparency. Limitations include the training's dependence on high-quality message labels generated by a teacher LLM, and the initial difficulty for agents to coordinate when generating messages based solely on local observations. The LangGround method proposed by Li et al. (2024c) significantly enhanced the human interpretability of communication and cross-team collaboration ability. However, it similarly relies on the quality of LLM-synthesized data and requires tuning alignment hyperparameters. Godfrey et al. (2024) utilize LLMs for negotiation to generate high-level plans, substantially reducing training time and offering a highly interpretable decision process. Drawbacks include high latency due to reliance on LLM APIs and limited applicability in continuous or complex environments.

To enhance the formality and computability of semantic communication, Gemp et al. (2024) modeled multi-round dialogues as an extensive-form incomplete information game, employing game-theoretic algorithms such as Counterfactual Regret Minimization (CFR) and Policy Space Response Oracles (PSRO) to solve for equilibrium strategies. The resulting strategic instructions were then injected as interpretable signals into the LLM's generation process, thereby improving the theoretical traceability and semantic transparency of interactive behaviors within natural language integration. Furthermore, a pioneering study by Daskalakis et al. (2024) employed an LLM as a *narrative parser* to automatically construct formal game trees from story text. Through game equilibrium analysis, it provided rational-choice-based motivational explanations for character behavior, demonstrating the potential of LLMs for semantic explanation and counterfactual reasoning in complex interactive behaviors. While the explanations generated by this method possess a formal foundation, it currently relies on prompt engineering and remains at the proof-of-concept stage.

In summary, **interaction semantic explanation techniques based on communication protocol analysis** provide a direct pathway to interpretability for understanding “*how agents coordinate through dialogue*” in collaboration, by dissecting message content, optimizing communication structures, and even integrating natural language. To systematically outline their characteristics, Table 14 presents a comparative analysis of representative studies, covering technical approaches, core mechanisms, advantages, limitations, and applicable scenarios.

Nonetheless, this field continues to face serious challenges:

- First, a fundamental tension exists between semantic richness and efficiency. Studies such as Li et al. (2024b) and Godfrey et al. (2024) demonstrate the powerful interpretability of natural language communication. However, this approach is characterized by high bandwidth and latency, which hinders its applicability in real-time systems. Conversely, Information Bottleneck-based methods, such as those presented by Yang et al. (2024) and Ding et al. (2023b), improve efficiency. Yet, the learned abstract representations often lack semantic transparency.
- Secondly, while promising, LLM-driven communication in studies such as Li et al. (2024b, 2024c) and Godfrey et al. (2024) is susceptible to inherent “hallucination” problems, potentially generating inaccurate or misleading content, which constitutes a critical flaw in safety-sensitive scenarios.
- Finally, concerning protocol flexibility, predefined structured protocols such as those of Selimović et al. (2024) and Meng and Tan (2024), often lack adaptability, whereas fully learned protocols, such as in Hu et al. (2024b), frequently exhibit limited generalization

capability.

Future research could seek breakthroughs in the following aspects:

- **Develop Hierarchical Communication Frameworks:** Build upon the information compression concepts from Yang et al. (2024) and Ding et al. (2023b) to design efficient low-level signals, while integrating the semantic generation capabilities of Li et al. (2024b) and Godfrey et al. (2024) to employ natural language at critical decision points. This approach strikes a balance between efficiency and semantics.
- **Enhance Reliability and Robustness:** Address the “hallucination” issue in LLMs by incorporating verification, correction, and uncertainty quantification mechanisms for generated content. This direction can extend research such as Li et al. (2024b) and Li et al. (2024c).
- **Explore Causal and Counterfactual Explanations:** Integrate causal reasoning tools, such as those referenced in Briglia et al. (2025) and Meli et al. (2025), into communication analysis. This would facilitate explanations not only of “what was communicated” but also of “why this communication was necessary” and “what the outcome might have been under a specified counterfactual intervention”, thereby deepening the explanatory depth.

4.3.3 Interaction relation verification via causal inference

This category of methods approaches the explanation of collaboration mechanisms from an **intrinsic logic perspective**, aiming to answer the fundamental questions: Do genuine causal relationships exist between agent behaviors? What is the cause, and what is the effect? Its technical core involves employing tools such as causal discovery, intervention experiments, and counterfactual analysis to rigorously test and quantify causal effects between actions. It forms a complementary relationship with methods like information structure analysis and semantic interpretation, distinguishing between the **descriptive** and the **explanatory**: the former reveals *what is*, while the latter reveals *why it is*.

To rigorously verify causality between behaviors, this category constructs a hierarchical analytical framework: First, through **interaction structure learning based on causal discovery**, it infers potential causal graphs from data. Subsequently, by **quantifying causal effects via counterfactual reasoning**, it assesses the impact of specific actions or events. Going further, it achieves synergy between interpretability and policy learning by *deeply integrating causal models into the learning architecture*. This framework aims move beyond purely correlational analysis by examining potential causal influences in collaboration.

Interaction structure learning based on causal discovery

These methods automatically learn the causal structure among variables from observational data. Briglia et al. (2025) utilized the *PC algorithm* for online learning of a causal Bayesian network and employed the causal graph for intervention. The framework is algorithm-agnostic and can enhance safety, but its causal discovery relies on a random policy, resulting in low efficiency and high variance, with limited effectiveness in highly cooperative tasks. Chen and Zhang (2023) introduced a *context-aware Bayesian network*, capturing dynamic dependencies through differentiable Directed Acyclic Graph (DAG) learning, which improves cooperation efficiency and supports the CTDE paradigm. However, DAG

learning can be unstable under partial observability. Wu et al. (2022) introduced the *Convergent Cross Mapping (CCM)* method for task allocation, enabling quantitative assessment of coordination levels and identification of sub-team structures. Its drawbacks include high computational complexity and potential insensitivity to extremely sparse networks. Meli et al. (2025) employed the *PCMCi+ algorithm* to learn causal structures from time-series data, combining feature attribution with causal discovery. Limitations include the potential inability of integrated gradients to capture indirect effects of communication, and the sensitivity of causal discovery to the length and quality of time-series data.

Causal effect quantification based on counterfactual reasoning

These methods quantify causal influence by constructing “what-if” scenarios. Zhao and Xie (2025) proposed the *Multi-level Advantage Counterfactual Advantage (MACA)* function, which explicitly models multi-level credit assignment. It is suitable for complex collaboration scenarios and offers theoretical guarantees, but it relies on global state for training, and the attention computation increases complexity. Jin et al. (2025b) introduced a *Counterfactual Group Relative Policy Advantage (CGRPA)* module for counterfactual contribution measurement, providing an intuitive explanatory tool for credit assignment. However, the accuracy of its counterfactual estimation is sensitive to hyperparameters, and generalization capability may be limited in some complex environments. Chai et al. (2024) proposed a *Model-based Counterfactual Imagination method (MACD)*, enabling efficient counterfactual evaluation based on the latest policy. Its limitations include the complexity of world model training and the restriction of counterfactual evaluation accuracy by model prediction error. Gao et al. (2025) proposed a communication method based on *Counterfactual Message Value (CMV)*, yielding an explicitly interpretable communication pattern and effectively reducing redundancy. However, its effectiveness heavily depends on the accuracy of the centralized critic.

Counterfactual reasoning has also been widely applied in other works. Wenhan et al. (2024) achieved precise credit assignment through *counterfactual reward estimation*, effectively handling heterogeneous agents and mixed action spaces. This requires training an additional reward network, increasing computational complexity, and is sensitive to hyperparameters. Shi et al. (2025) performed counterfactual reasoning by learning a generative model, capable of capturing the causal mechanism of reward generation and effectively promoting cooperation. Its limitations include dependency on the accuracy of the generative model and potential fragility in the presence of non-cooperative or adversarial agents. Maliha and Hougen (2025) proposed a *mask-based perturbation framework* to identify key states, achieving interpretability for multi-agent coordination. However, it relies on the selection of a mask threshold and requires additional training of a mask network. Nagpal et al. (2025) leveraged an LLM for credit assignment, producing intuitive and comprehensible natural language explanations. Its drawbacks include heavy reliance on the quality and reasoning speed of the LLM, resulting in substantial computational overhead and high latency.

Architecture design based on causal model integration

These methods deeply integrate causal models into the learning architecture. Yao et al. (2024) constructed a *Structural Causal Model (SCM)* and utilized causal intervention to quantify influence, providing a causal explanation for black-box value decomposition. Its implementation relies on manual partitioning of the external state and the accuracy of the world model, while also increasing computational complexity. Du et al. (2024) quantified

context-dependent causal influence by constructing a *Multi-Agent Causal Graphical Model* (MACGM), using it as an intrinsic reward to guide collaborative strategy from the ground up. Limitations include high computational cost for conditional mutual information and the need for tuning the temperature parameter of the intrinsic reward. Selmonaj et al. (2025b) proposed an *Information-theoretic Shapley value-based Instrumental Cooperative Value* (ICV) method, which can provide multi-perspective behavioral explanations completely without a reward signal. However, computing the Shapley value for all permutations suffers from combinatorial explosion, and the sequential value Markov game framework is an approximation of the actual parallel environment.

In summary, **interaction relationship verification techniques based on causal reasoning** provide a causally oriented pathway to interpretability for answering *whether “behaviors are driven by genuine causality”* in collaboration, through discovering causal structures, quantifying counterfactual effects, and integrating causal models. To systematically delineate their characteristics, Table 15 presents a comparative analysis of representative studies, covering technical approaches, core mechanisms, advantages, limitations, and applicable scenarios.

Notwithstanding these advances, the field faces several notable challenges:

- **Computational Complexity:** Causal discovery algorithms (e.g., PCMCI+ as employed by Meli et al. (2025)) and counterfactual reasoning methods (Chai et al. 2024; Wenhan et al. 2024) are computationally intensive, limiting their scalability in large-scale MASs.
- **Estimation Instability:** The reliability of counterfactual estimates remains a concern; methods from Wenhan et al. (2024) and Shi et al. (2025) exhibit sensitivity to model specifications and hyperparameters, posing risks of estimation bias and variance.
- **Interpretability Deficit:** Resulting causal structures or influence measures, as explored in Briglia et al. (2025) and Meli et al. (2025), often remain abstract and lack intuitive interpretability for domain experts, leading to a persistent “explanation gap” between technical causal models and practical understanding.

Future research should prioritize addressing these limitations through the following directions:

- **Developing Efficient Causal Algorithms:** Investigate lightweight causal discovery and counterfactual reasoning methods specifically designed for MASs, targeting the computational bottlenecks identified in Meli et al. (2025) and Chai et al. (2024) through approximation techniques and distributed computing paradigms.
- **Improving Estimation Robustness:** Explore more stable causal effect estimation methods building upon the work of Wenhan et al. (2024) and Shi et al. (2025), incorporating robust statistical techniques and sensitivity analysis to reduce dependence on specific model assumptions.
- **Promoting Intuitive Explanations:** Create specialized visualization tools for causal structures learned by methods like Briglia et al. (2025) and Yao et al. (2024), and integrate causal explanations with natural language generation to facilitate the translation from technical “causal graphs” to comprehensible “causal narratives” for diverse stakeholders.

Through continued exploration along these directions, causal inference may become an increasingly important component for constructing next-generation trustworthy MASs. This evolution will provide a solid foundation for robust safety-relevant safety guarantees and decision transparency in complex collaborative environments, ultimately enabling more reliable and explainable autonomous systems.

4.3.4 Collaboration structure discovery via roles and behavioral patterns

This category of methods approaches the explanation of collaboration mechanisms from an **organizational structure perspective**, aiming to answer the question: What kinds of functional division of roles and stable collaborative patterns have emerged in the system? Its technical core involves using unsupervised learning or structured methods to automatically identify functional role categories from agents' behavioral data and to analyze the recurring or relatively stable collaborative patterns between different roles. It complements techniques like information structure analysis and causal verification by contrasting the *transient with the long-term*, and the *microscopic with the macroscopic*: the former focus on short-term interactions, while the latter focus on long-term, stable social structures.

To uncover the stable organizational structures emergent within a system, this category explores from multiple perspectives: **Dynamic Role Assignment and Evolution** focuses on the adaptive formation of functional divisions of labor; **Hierarchical Intention and Cognitive Alignment** endeavors to explain the high-level goals and mental models underlying policies; **Organizational Models and Structured Policies** introduce explicit frameworks to constrain and guide collaborative behavior; and **Collaborative Pattern Extraction and Abstraction** directly mines reusable interaction prototypes from behavioral data. These perspectives collectively characterize the macro-scale form and evolutionary patterns of collaboration.

Dynamic role assignment and evolution

These methods dynamically assign functional roles through online learning mechanisms. The *Role-Oriented Graph Convolutional Network (ROGC)* proposed by Liu et al. (2022b) exhibits strong adaptability in dynamic role assignment, and intra-group communication enhances collaboration efficiency. However, its clustering method is sensitive to sample quality, and the number of roles must be predefined. The *Role-Based Heterogeneous Action Representation (RBHAR)* method proposed by Li et al. (2024d) also offers strong adaptability in dynamic role assignment, and heterogeneous action representation increases policy diversity. Its limitations include the need to predefine and fix the number of roles, and the increased computational complexity from interactive action representation. Peng et al. (2023) adapt to dynamic environments through *Periodic Grouping (GCMA)*, reducing redundancy via intra-group communication. Drawbacks include the need to tune the grouping interval parameter and increased computational overhead from GNN-based communication. Lu et al. (2025) improve decision interpretability and safety through *rule-based preference modeling* and achieves experience sharing via federated learning. However, confidence modeling increases computational complexity, and the preference model depends on rule design. In an energy market scenario, Ma et al. (2025b) achieve dynamic role coordination through a *sequential update strategy*, providing theoretical guarantees for policy update monotonicity and intuitive market state reasoning. Limitations include potential increased

computational time from sequential updates and the assumption of perfect observation of preceding strategies, which may be restrictive in practice.

Hierarchical intention and cognitive alignment

These methods explain the formation of collaboration mechanisms at the level of intentions and cognition. The *Reinforced Meta-Thinking Agent (ReMA)* framework proposed by Wan et al. (2025b) significantly enhance the transparency and controllability of the reasoning process, but its training is computationally expensive, and multi-round interactions are prone to falling into an “echo trap”. Fu et al. (2025) generates interpretable multi-modal collaborative behaviors through a *consistency policy and an intention learner*, where intentions provide clear collaborative semantics. However, the intention codebook size must be predefined, and the overall framework is complex. The *Hierarchical Multi-agent Coordination Framework* (MIMA) proposed by Zhang and Xu (2024) features a clear hierarchy and well-defined information learning objectives, demonstrating powerful zero-shot generalization ability. Its model structure is complex and difficult to train, however. Ge et al. (2022) enhance collaborative interpretability through *Cognitive Difference and Consistent Representation* (CDCR), offering a novel difference-based communication perspective. A limitation is that the consistent representation strategy may not obtain necessary supervision signals during execution. The *Multi-Agent Goal Induction Consistency* (MAGIC) framework proposed by Chen et al. (2022) significantly improves exploration efficiency and collaborative performance in complex tasks. It relies on task-specific goal sequence partitioning, however, and its attention mechanism can be susceptible to observation noise.

Organizational models and structured policies

These methods constrain or guide behavior by introducing explicit organizational frameworks or structured policies. Soulé et al. (2025a) explicitly embeds the *MOISE+ organizational model* into training, providing high-level organizational semantic explanations and enhancing behavioral controllability. Its effectiveness depends on manually defined organizational norms, and role rigidity in dynamic environments may limit adaptability. Ibrahim and Fayad (2022) learn semantically meaningful latent variables through a *Hierarchical Latent Policy Model*, achieving efficient cooperation and policy interpretability in complex environments. This comes at the cost of high computational complexity, and the latent policy dimensions require careful design. Su et al. (2025) enhances collaboration mechanisms through *coalition identifiers and structural entropy*, improving algorithm performance and interpretability. Drawbacks include reliance on manual setting of coalition labels and high computational complexity for structural entropy calculation.

Collaborative pattern extraction and abstraction

These methods directly extract and abstract reusable collaborative prototypes from behavioral data. The *Interaction Pattern Disentanglement (OPT)* method proposed by Liu et al. (2024b) significantly enhances generalization capability and interpretability. Its computational complexity is high (quadratically related to the number of entities), however, and it relies on the design of sparse activation functions and contrastive loss. Domenech I Vila et al. (2024) extended the *discrete predicate-based policy graph* method to MAS, filling a gap in transferring single-agent methods to multi-agent settings. Limitations include the policy graph’s reliance on discrete predicate design (lacking generality) and difficulty in covering all potential states. Addressing multi-robot task allocation, Bezerra et al. (2025) explains dynamic coalition formation through *intention sharing*, achieving significant performance and strong scalability. It depends on communication range, however, and training

time is long. Ruan et al. (2024) utilizes a *Determinantal Point Process (DPP)* for dynamic agent grouping, with multi-level decision enhancement adapting to complex collaborative scenarios. Its pre-training computational cost is high, and graph search and grouping processes increase inference overhead. The *Hierarchical Cognitive Graph Autoencoder (HCGA)* proposed by He et al. (2025a) exhibits strong adaptability, and its hierarchical structure enhances cognitive efficiency and interpretability. Drawbacks include high computational complexity, and while interpretability is enhanced through structural design, it does not provide intuitive human-readable explanations. Soulé et al. (2025b) represents a paradigm of hybrid methods, featuring clear multi-agent division of labor, strong behavioral interpretability, and organizational constraints that improve learning efficiency. It relies on domain knowledge to define roles and missions, however, and incurs significant computational overhead.

In summary, **collaborative structure discovery techniques based on roles and behavioral patterns** provide a direct pathway to interpretability for understanding *how teams form stable divisions of labor and organizational structures* in collaboration, through dynamic role assignment, cognitive alignment, organizational modeling, and pattern extraction. To systematically outline their characteristics, Table 16 presents a comparative analysis of representative studies, covering technical approaches, core mechanisms, advantages, limitations, and applicable scenarios.

Despite significant advances in dynamic role assignment, cognitive alignment mechanisms, collaboration pattern extraction, and explicit structure analysis, several challenges persist:

- **Semantic Interpretability Gap:** Automatically discovered roles (Liu et al. 2022b; Li et al. 2024d) often lack explicit high-level semantic labels, making direct correspondence to domain-specific concepts (e.g., “assault”, “support”) challenging for practical interpretation.
- **Dynamic Adaptation Limitations:** The adaptability of roles and patterns to environmental changes remains constrained in most methods (e.g., Liu et al. (2022b); Ruan et al. (2024)), limiting their effectiveness in rapidly evolving scenarios.
- **Evaluation Metric Deficiency:** This field lacks systematic evaluation frameworks to quantitatively measure the quality, stability, and effectiveness of discovered roles and patterns across different applications.
- **Prior Knowledge Integration Challenge:** A fundamental tension exists between the explicit organizational models of Soulé et al. (2025a) and the data-driven discovery approaches of Liu et al. (2024b) and Soulé et al. (2025b) regarding optimal balance between incorporating domain knowledge and maintaining learning flexibility.

To address these limitations, future research should prioritize the following directions:

- **Semantic Role Annotation Enhancement:** Integrate advanced natural language technologies, particularly LLMs, to automatically generate semantic labels for roles discovered in existing methods. Approaches such as those by Liu et al. (2022b), Li et al. (2024d), and Soulé et al. (2025b) would benefit from such semantic enrichment to bridge the interpretability gap between computational discoveries and domain understanding.
- **Online Evolution Mechanism Development:** Enhance the adaptability of dynamic

grouping methods by creating real-time role assignment mechanisms capable of responding to environmental changes. Representative works by Ruan et al. (2024) and Zhang and Xu (2024) provide foundational concepts that could be extended with on-line learning capabilities and rapid adaptation algorithms for improved dynamic performance.

- **Comprehensive Evaluation Framework Construction:** Establish multi-dimensional evaluation frameworks incorporating structural stability, task relevance, semantic consistency, and adaptation speed metrics. Such frameworks would enable quantitative comparison of methodological strengths and weaknesses across diverse approaches, including those demonstrated by Liu et al. (2024b), Liu et al. (2022b), and Soulé et al. (2025b), while facilitating standardized benchmarking.

By systematically addressing these challenges, role and behavioral pattern discovery technology may contribute to the development of adaptive and interpretable MASs and support more transparent human–AI collaboration in complex applications. This advancement will lay the essential groundwork for truly effective human-AI collaborative teamwork in complex, real-world applications, ultimately enabling more transparent and trustworthy autonomous systems.

4.3.5 Collaboration logic explanation via value function decomposition

This category of methods approaches the explanation of collaboration mechanisms from the perspective of **incentive mechanisms**, aiming to answer the fundamental question: How are individual collaborative behaviors incentivized by decomposing the global team objective through value/credit allocation? Its technical core involves analyzing how a centralized team value function is decomposed into individual value functions, or quantifying individual or coalitional contributions based on theories such as game theory or information theory. It forms a complementary relationship with the aforementioned four categories, contrasting **incentive-driven** with **behavioral manifestation**: the former explains *why to collaborate*, while the latter explains *how to collaborate*.

To elucidate how team objectives drive individual collaboration, this category constructs a diverse toolkit for explanation: **Interpretable value function decomposition architectures** embed analyzability from the source of model design; **Contribution quantification based on game theory and information theory** provides formalized approaches and, in some cases, axiomatic properties or theoretical analyses for credit assignment; and **Semantic and logic enhancement methods** strive to transform abstract numerical rewards into human-understandable concepts and rules. Together, these methods aim to unveil the incentive origins and allocation logic underlying collaboration.

Value function decomposition architectures

These methods construct end-to-end trainable decomposition architectures and proactively analyze their internal mechanisms. Liu et al. (2025b) designed an interpretable architecture based on *Mixtures of Recurrent Soft Decision Trees* (MIXRTs). The decision process is traceable, and credit assignment is intuitive, but the expressive capacity of the tree model may have an upper limit. Liang et al. (2024) proposed the *Multiplicative Value Decomposition* (MVD) algorithm and introduced a *Virtual Synchronous Proxy* (VSP) mechanism. Multiplicative interactions enhance the model's ability to represent asynchronous dependencies,

but the introduction of a virtual proxy adds an abstraction layer and complexity. Liu et al. (2023c) constructed a decomposition architecture based on *Generalized Additive Models* (GAMs), providing intrinsically interpretable decomposition and visualization, with superior performance in difficult scenarios. However, high-order interactions reduce interpretability, and VAE training increases overhead.

In the direction of GNNs, the *QGNN* method proposed by Kortvelesy and Prorok (2022) naturally models dependencies using GNNs with adjustable communication ranges, but it is only suitable for discrete action spaces and does not consider practical communication issues. Wei et al. (2022) inherently provides dynamic, fine-grained interpretability through Graph Attention Networks (GATs). However, its interpretability is limited to relative importance within the team and does not provide action-level explanations for individual agents. Zhu et al. (2025) proposed the *Communication via Factored Value Function Policy Gradient* (CFPG) framework, which for the first time combined value decomposition with learnable communication in continuous action spaces. Its assumption of fully-connected broadcast communication is unrealistic, however, and explanations remain at the weight level.

In the direction of causal modeling and reward redistribution, Zhang et al. (2024a) proposed the *Causally-Aware Spatio-Temporal (CAST)* reward decomposition method, relaxing the linear reward assumption and providing clear interpretability through causal structure. It still assumes linear superposition in the temporal dimension and invertibility of the mixing function in the spatial dimension, however. Xiao et al. (2022) proposed the *Agent-Time Attention for Reward Redistribution (AREL)* method, where an attention mechanism flexibly models spatiotemporal dependencies without requiring a predefined reward structure. The computational cost of the attention mechanism grows with the number of agents, however, and it lacks theoretical guarantees at the causal level.

Game theory and contribution quantification

These methods employ mathematical concepts from cooperative game theory as the theoretical foundation for credit assignment. Heuillet et al. (2022) proposed a *post-hoc, model-agnostic method*. It is completely model-agnostic, easy to integrate, and provides intuitive explanations, but its computational cost is extremely high, and the explanations (static average contributions) cannot reflect dynamic decision-making. Qiu et al. (2023) proposed a method based on the *Shapley value*, providing quantitative contribution allocations based on Shapley-value principles. Computing the Shapley value is complex, however, and explanations are at a global average level, not refined to per-step decisions. Li et al. (2025d) proposed a method based on the *nucleolus concept*, capable of explaining dynamic cooperative organizational structures with theoretical guarantees of fairness and stability. Computing the nucleolus is extremely complex, however, and effectiveness relies on clearly defined subtasks. Han et al. (2022a) constructed a general credit assignment framework using *game-theoretic semi-values*, introducing model-based game-theoretic credit assignment to continuous control for the first time and significantly improving sample efficiency. Model learning may introduce errors, however, and computational cost grows exponentially with the number of agents. The *Core-Originated Reward Assignment (CORA)* method proposed by Ji et al. (2025) utilizes the core solution to provide theoretically-guaranteed fairness and stability, effectively mitigating the over-update problem. Its computational complexity grows exponentially with the number of agents, however, and the core solution may not be unique. The *Difference Rewards (DR)* mechanism proposed by Castellini et al. (2025) avoids the bootstrapping problem in learning Q-functions, resulting in more stable and efficient learn-

ing and providing an intuitive measure of individual contribution. In partially observable environments, however, the baseline for difference rewards may be biased.

Information theory and mutual information methods

These methods leverage information-theoretic principles to provide explanations at the information level for collaboration. Ding et al. (2023a) proposed a *graphical mutual information maximization* method, explicitly optimizing the retention of feature and topological information to improve communication quality. Estimating mutual information in high-dimensional spaces remains challenging, however. Yang et al. (2025a) proposed a coordination optimization framework based on *reward redistribution and experience reuse*, explicitly optimizing team collaborative behavior to enhance policy coordination. The mutual information threshold requires tuning, however, and experience pool stratification increases storage and sampling complexity. Ding et al. (2024) was the first to introduce the *Graph Information Bottleneck (GIB)* principle into multi-agent communication learning, significantly improving adversarial robustness and communication efficiency. Its interpretability output is rather implicit, however, not providing intuitive human-readable explanations.

Semantic and logic enhancement methods

These methods introduce semantic representations or formal logic to translate abstract reward signals into human-understandable concepts and rules. In the application of LLMs, Wang et al. (2025b) supports mixed-quality human feedback and resists low-quality feedback through weight decay, enhancing policy interpretability. It relies on LLM parsing accuracy, however, and human feedback is costly. The *LCA framework* proposed by Lin et al. (2025) uses an LLM to generate agent-specific dense rewards, effectively handling sparse rewards and supporting heterogeneous agents. It relies on the LLM's task comprehension ability, however, and ranking generation is computationally expensive. The *QLLM framework* proposed by Jiang et al. (2025) leverages an LLM to generate a *Training-Free Credit Assignment Function (TFCAF)*, which is training-free, highly interpretable, and adapts to high-dimensional states. It depends on the quality of LLM code generation, however, and the generation process requires multiple iterations. Zhang et al. (2025a) proposes a method that incorporates LLMs into RL to enhance interpretability. In this framework, the LLM generates human-readable explanations for decisions, while a Retrieval-Augmented Generation (RAG) mechanism is used to make the reward design process traceable. However, the slow inference speed of LLMs may hinder real-time performance, and the quality of explanations is contingent upon careful prompt engineering.

In the direction of formal logic and concept modeling, Liu et al. (2024a) introduced *Linear Temporal Logic (LTL)* to formalize task logic, using formal logic to enhance decision transparency and support flexible multi-task expression. Designing LTL formulas relies on prior knowledge, however, and logic conversion may incur additional overhead. The *Concept-based Multi-agent Q-learning (CMQ)* proposed by Ge et al. (2025) integrates a concept bottleneck model into value decomposition, significantly enhancing model interpretability through the concept bottleneck and supporting human-in-the-loop intervention. The number of concepts must be predefined, however, and it may increase computational overhead.

In summary, **collaborative logic explanation techniques based on team value function decomposition** provide a principled pathway to interpretability for understanding *how global goals incentivize and allocate individual credit* in collaboration, through architec-

tural design, game-theoretic quantification, information-theoretic methods, and semantic enhancement. To systematically delineate their characteristics, Table 17 presents a comparative analysis of representative studies, covering technical approaches, core mechanisms, advantages, limitations, and applicable scenarios.

Despite substantial progress, several fundamental challenges in multi-agent value decomposition remain only partially resolved:

- **Theoretical Guarantee Deficiencies:** Ensuring the theoretical rationality of value function decomposition—particularly establishing formal guarantees for decomposition reasonableness—requires fundamental advances. This is critical for preventing pathological behaviors such as the “lazy agent” problem, as highlighted in studies like Kortvelesy and Prorok (2022) and Ge et al. (2025).
- **Computational Complexity Limitations:** Game-theoretic approaches, such as those introduced by Heuillet et al. (2022), Li et al. (2025d), and Han et al. (2022a), provide strong theoretical fairness guarantees but suffer from prohibitive computational complexity in large-scale systems, limiting their practical applicability.
- **Semantic Interpretation Gaps:** Despite beneficial progress through semantic enhancement techniques, as seen in Ge et al. (2025) and Zhang et al. (2025a), a significant explanatory gap persists between learned value functions and high-level semantic concepts that are intuitively understandable to human operators.
- **Dynamic Adaptation Challenges:** Most existing methods display limited adaptability to rapidly shifting collaborative relationships in non-stationary environments, constraining their effectiveness in real-world applications with evolving conditions.

Looking ahead, future research is expected to advance along several promising trajectories:

- **Deepening Causal Inference Integration:** Incorporating causal inference frameworks, such as those explored by Briglia et al. (2025) and Meli et al. (2025), into value decomposition could help distinguish correlation from causation in agent contributions. This integration would fundamentally improve the rationality of contribution quantification in methods like Heuillet et al. (2022) and Li et al. (2025d), establishing more robust foundations for credit assignment.
- **Advancing Computational Efficiency:** Developing more efficient approximation techniques for computing solution concepts such as the Shapley value and Nucleolus is essential for addressing the computational bottlenecks identified in works including Heuillet et al. (2022) and Li et al. (2025d), enabling application to larger-scale MASs.
- **Constructing Hierarchical Explanation Frameworks:** Building on semantic enhancement approaches exemplified by Ge et al. (2025) and Zhang et al. (2025a), future work should aim to develop hierarchical explanation systems that systematically bridge low-level value signals with high-level, human-intelligible concepts, creating more accessible interpretability for diverse stakeholders.
- **Enhancing Open-World Adaptation:** Research should investigate adaptive value decomposition mechanisms suited for open and dynamic environments, improving the robustness and generalization of foundational methods such as that of Kortvelesy and Prorok (2022) and enabling effective performance in real-world applications with non-stationary conditions.

By addressing these challenges through integrated theoretical and practical advances, value function decomposition approaches will evolve into more powerful, interpretable, and adaptable frameworks for understanding and optimizing collaborative behaviors in complex MASs.

4.3.6 Summary and research directions for agent interaction and collaboration mechanisms

This section provides a comprehensive summary and comparative analysis of the five technical paradigms for the explainability of agent interaction and collaboration mechanisms. It then synthesizes their intrinsic connections into a cohesive, multi-perspective explanatory framework. Finally, it discusses current limitations and outlines promising future research directions.

Comprehensive summary of technical paradigms

This section has systematically examined on the five technical paradigms for agent interaction and collaboration explainability. A comprehensive comparison of these paradigms is presented in Table 5, which contrasts their respective technical categories, core explanatory questions, advantages, and limitations.

Information structure analysis based on attention mechanisms and relationship modeling 4.3.1 dynamically visualizes dependencies using tools like Graph Attention Networks (Qi et al. 2022; Sun et al. 2023), offering high intuitiveness but providing correlational rather than causal insights. **Interaction semantic explanation based on communication protocol analysis 4.3.2** enhances transparency via information-theoretic (Yang et al. 2024) or language model-based methods (Li et al. 2024b, c), but faces trade-offs between efficiency, semantic richness, and generalization. **Interaction relationship verification based on causal reasoning 4.3.3** ensures rigorous causal distinction using discovery (Briglia et al. 2025; Meli et al. 2025) and counterfactual reasoning (Zhao and Xie 2025; Gao et al. 2025), though it is computationally intensive and scales poorly. **Collaborative structure discovery based on roles and behavioral patterns 4.3.4** uncovers team structures via unsupervised learning (Liu et al. 2022b) or organizational models (Soulé et al. 2025a), but discovered roles often lack semantic labels and adaptability. **Collaborative logic explanation based on team value function decomposition 4.3.5** explains collaboration through value decomposition (Kortvelesy and Prorok 2022; Ge et al. 2025) or game-theoretic quantification (Heuillet et al. 2022; Li et al. 2025d), yet struggles with decomposition rationality and combinatorial complexity.

Intrinsic connections: a multi-perspective explanatory framework

The five paradigms, while each having distinct focuses in explanatory lens and methodological core, are not mutually exclusive. Instead, they constitute a set of complementary explanatory perspectives that progresses from informational foundations to incentive mechanisms, and from surface-level interaction to deep logic, collectively providing an organizing framework for studying for the core question: “How do agents achieve coordination through interaction?” The interrelations among these paradigms can be understood along three key dimensions:

- **Complementary Explanatory Perspectives:** Explanatory depth increases progressively. *Information structure analysis* describes “how information flows” (the foundation).

Interaction semantic explanation reveals “how to coordinate through dialogue” (the expression). *Interaction relationship verification* explains “whether there is genuine causality between behaviors” (the logic). *Collaborative structure discovery* elucidates “how the team organizes and divides labor” (the form). *Collaborative logic explanation* reveals “how the system drives collaboration through rewards” (the mechanism).

- **Taxonomic Coverage:** The five categories represent prominent analytical perspectives within the review taxonomy *f*: informational foundation (*What Information*), expressive modality (*How to Express*), intrinsic logic (*Why Causal*), organizational form (*How Organized*), and incentive origin (*For What Incentive*). Together, they form a complete taxonomic framework for this problem space.
- **Functional Complementarity:** They mutually supplement each other to approximate a complete understanding of collaborative behavior. For example, explaining a successful coordinated attack in *StarCraft* may involve: first, observing through attention heat-maps (*information structure analysis*) that multiple allied units simultaneously “focus on” a key enemy target, revealing the informational focal point of collaboration; next, finding via communication logs (*interaction semantic explanation*) that a unit issued a “focus fire” command, clarifying the expressive modality of coordination; then, using counterfactual reasoning (*interaction relationship verification*) to verify “if the command had not been received, would the focus fire have occurred?”, determining the causal necessity of that communication; subsequently, confirming through role analysis (*collaborative structure discovery*) that the attackers were the “damage dealer” role while the “controller” role provided support, understanding the organizational division of labor; and finally, using value decomposition (*collaborative logic explanation*) to obtain contribution scores for each unit, clarifying why the algorithm rewards this collaboration—because it correctly assessed the differential contributions of different roles to team victory, thereby revealing the incentive origin of the collaboration.

Limitations and future directions

Existing explanation techniques for interaction and collaboration often rely on isolated paradigms (e.g., analyzing only communication graphs or computing contribution values), making it difficult to provide end-to-end explanations of collaboration mechanisms. Their computational efficiency and robustness also face severe challenges in complex, open environments. Future research needs to achieve breakthroughs by focusing on the following specific technical pathways:

- **Unified Systems via Paradigm Fusion:** Design hybrid explanatory frameworks that organically integrate relation graphs (Zhu et al. 2024), causal reasoning (Briglia et al. 2025), and natural language generation (Li et al. 2024b) to realize a complete explanatory chain from “phenomenon description” to “mechanism exposition.”
- **Online Optimization Closed Loop:** Explore transforming explanatory signals—such as counterfactual contributions and collaboration bottlenecks—into intrinsic rewards or safety constraints, constructing a real-time enhancement loop of “explanation-diagnosis-intervention-optimization.”
- **Efficiency and Robustness Enhancement:** To meet the demands of large-scale deployment, develop lightweight Shapley value approximation algorithms (Li et al. 2025d), incremental causal discovery methods (Meli et al. 2025), and efficient dynamic graph

Table 5 Comprehensive comparison of explainability technique paradigms for agent interaction and collaboration

Technique category	Core explanatory question	Representative techniques	Explanatory focus	Core advantages	Inherent limitations & challenges
Information structure analysis via attention mechanisms & relational modeling	What are the dependencies and interaction topology between agents during decision-making?	Refined GAT attention (Sun et al. 2023); Qi et al. (2022); Incentive sharing modeling via weight matrices (Gemp et al. 2022; Radke et al. 2023); Dynamic graph learning (Liu et al. 2022a; Duan et al. 2024b); Hypergraph modeling (Zhu et al. 2024)	Interactional information structure and dependency strength	Highly intuitive, dynamically visualizes interaction preferences and topological evolution; currently the most widely applied technical path	Causality of explanations is questionable (correlation does not imply causation); low semantic abstraction level; computational efficiency bottlenecks in large-scale systems
Interaction semantics interpretation via communication protocol analysis	What explicit communication (content, target, timing) do agents use to coordinate?	Information bottleneck optimization (Yang et al. 2024; Ding et al. 2023b); Structured topology learning Meng and Tan (2024); Hu et al. (2024b); LLM-driven natural language communication (Li et al. 2024b; Godfrey et al. 2024); Game-theoretic model construction (Gemp et al. 2024; Daskalakis et al. 2024)	Explicit dialogues and micro-coordination mechanisms in collaboration	Provides the most direct semantic explanations for explicit communication systems; integration with LLMs significantly narrows the human-machine semantic gap; game-theoretic modeling offers formalized, computable interaction semantics (Gemp et al. 2024); LLM narrative parsing reveals potential for transforming text into rational behavioral motivations (Daskalakis et al. 2024)	Trade-offs between communication efficiency and semantic richness; poor generalization and robustness of learned protocols; LLM hallucination risks; high computational complexity in game-theoretic methods; LLM narrative parsing remains at conceptual validation stage
Interaction relation verification via causal inference	Do true causal relationships exist between agent behaviors? What is the causal chain?	Online causal discovery (Briglia et al. 2025; Meli et al. 2025); Counterfactual contribution quantification (Zhao and Xie 2025; Chai et al. 2024); Causal model integration (Yao et al. 2024; Selmonaj et al. 2025b)	Driving logic and causal chains in collaboration	Provides the most rigorous and reliable explanatory framework, distinguishing correlation from causation; crucial for safety-critical scenarios	Extremely high computational complexity, difficult to scale; counterfactual estimation sensitive to model assumptions; interpretability of learned causal structures needs improvement
Collaborative structure discovery via roles & behavioral patterns	What functional division of labor (roles) and stable collaboration patterns emerge?	Dynamic role clustering (Liu et al. 2022b; Li et al. 2024d); Hierarchical intent alignment (Fu et al. 2025; Zhang and Xu 2024); Interaction pattern decoupling (Liu et al. 2024b)	Macroscopic sociological structure and functional division of labor within the team	High-level, structured team perspective; aids understanding of long-term, stable collaboration paradigms and organizational forms	Automatically discovered roles lack high-level semantic labels; insufficient adaptability to role evolution in non-stationary environments; lacks unified evaluation systems

Table 5 (continued)

Technique category	Core explanatory question	Representative techniques	Explanatory focus	Core advantages	Inherent limitations & challenges
Collaboration logic explanation via team value function decomposition	How does the global team goal incentivize individual behavior through value/credit decomposition?	Interpretable decomposition architectures (Liu et al. 2025b, 2023c); Game-theoretic credit allocation (Heuillet et al. 2022; Li et al. 2025d); Semantic logic enhancement (Ge et al. 2025; Zhang et al. 2025a)	Incentive alignment and credit assignment logic in collaboration	Principled explanation from the system incentive source; tools like Shapley values have solid mathematical foundations	Theoretical challenges in ensuring decomposition rationality (e.g., monotonicity); combinatorial explosion in game-theoretic methods; explanatory gap to high-level semantics

visualization techniques (Duan et al. 2024b), while simultaneously strengthening the robustness of the explanatory models themselves.

- **Adaptation to Open Environments:** Extend the adaptability of explanation mechanisms to complex scenarios characterized by non-stationarity, partial observability, and human-agent hybrid teams.

In conclusion, the advancement of explainability for multi-agent interaction hinges on integrated progress across cross-paradigm integration, online learning, and open-environment adaptation. By addressing the fundamental limitations of current approaches through cohesive methodological development, these explanation frameworks are poised to form the cornerstone of next-generation MAS capable of sophisticated, transparent, and trustworthy coordination.

4.4 Explainability techniques for team strategy and emergent behavior

The intelligence of MASs stems not only from the capabilities of individual agents but also **emerges from** the complex interactions and collective behaviors arising from their coordination. To truly understand and trust such systems, it is essential to systematically explain the principles behind team strategy and emergent behavior. In this review, explanations of team strategy and emergent behavior are organized around five complementary questions (as shown in Fig. 6): (1) the **Semantic Description**—What is the team’s overall strategy and its associated high-level objective? (2) the **Dynamic Process**—How does the team’s strategy evolve over time and transition between phases? (3) the **Ontological Foundation**—whether the observed team strategy constitutes a genuine and irreducible functional whole? (4) the **Counterfactual assessment**—How do outcomes vary under different strategies, as shaped by the key decisions that determine success? and (5) the **Generative Mechanism**—What fundamental principles underlie collective behavior, and how can they be formally modeled?

These five questions provide an organizing framework for surveying team strategy and emergent behavior explainability, progressing from semantic characterization to underlying generative principles. To address them, research has converged on five complementary technical paradigms: **Team Behavior Abstraction and High-Level Goal Recognition** for semantic description, **Macroscopic State Modeling and Phase Analysis** for revealing dynamic processes, **Causal Emergence Theory** for validating ontological foundations, **Counterfactual Team-Level Analysis** for causal examination, and **Learning-Based World Models and Team Behavior Generation** for uncovering generative mechanisms. Together, these paradigms provide complementary analytical perspectives ranging from strategy description to generative modeling. The following sections detail how they collectively address these questions, providing a unified lens for explaining multi-agent team strategy and emergence.

4.4.1 Team behavior abstraction and high-level goal identification

This category of techniques aims to answer the question, “*What strategy is the team executing as a whole?*” The core of these techniques lies in compressing and mapping lengthy, low-level joint action trajectories into high-level, symbolic strategy concepts, task objec-

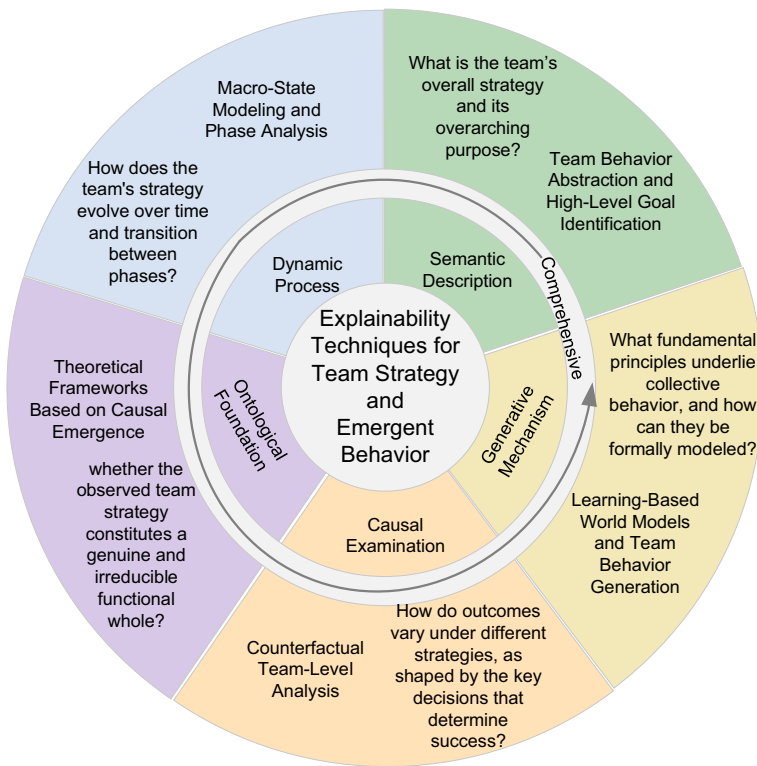


Fig. 6 Technical framework for explaining team strategies and emergent behaviors: core issues and five complementary technologies

tives, or team structures (e.g., grouping and division of labor, key landmarks, subgoal sequences). Within the interpretability framework, these methods reside at the **semantic bridge layer**, aiming to establish human-understandable semantic connections between microscopic individual interactions and macroscopic team tactics, thereby labeling and describing the “strategic intent” of emergent behaviors. They provide a static semantic foundation for subsequent analysis of the dynamic processes of team strategies (how they evolve) and define the objects to be explained for causal and generative mechanism analysis (why and how such strategies are formed).

Explicit grouping with behavioral consistency constraints

The core of this category of techniques is to actively shape a team architecture of “within-group collaboration and between-group division of labor” by predefining or learning grouping structures and designing explicit behavioral consistency loss functions. Representative methods have different emphases in their implementation paths. Yang et al. (2025c) proposes a **Dual-Level Behavioral Consistency (DLBC)** framework, which uses grouping and measures behavioral distributions using Wasserstein distance, introducing a dynamic scaling factor to balance within-group consistency and between-group diversity losses, thereby explicitly regulating behavioral diversity and prompting the team to emerge with clear functional divisions such as “interception groups” and “pursuit groups”. Its advantages include significantly improving within-group collaboration and between-group task alloca-

tion efficiency, along with strong generality. Another work Duan et al. (2024a) provides a more formal probabilistic perspective, introducing a **Group-Aware Coordination Graph (GACG)** that models within-group dependencies using Gaussian marginal distributions and directly optimizes policy distributions through group distance loss, causing within-group policies to converge and between-group policies to diverge. This method offers a highly structured, mathematically formalizable explanatory framework for team collaboration structures. The common limitations of these two types of methods are that their grouping strategies are often static, with limited adaptability to dynamic environmental changes, and the additional constraints or loss functions may increase training complexity.

Unsupervised representation learning for discrete pattern discovery

The core objective of this category of techniques is to learn semantically meaningful group consensus or strategy pattern representations from agents' historical interaction data in an unsupervised manner, thereby achieving automated identification and definition of team structures and strategy types. Representative works can be broadly divided into three categories: first, methods based on group consensus extraction, for example, Ruan et al. (2023) innovatively employs a *Vector Quantized Variational Autoencoder* to construct a *Group Clustering Graph*, extracting discrete "group consensus embeddings" from historical observations, then explicitly defining instantaneous team structures and transforming them into team-level strategies via hypernetworks to guide individual decisions, demonstrating excellent collaborative performance in complex tasks; second, methods focusing on temporal pattern recognition, for example, Na et al. (2025) focus on the dynamic temporal characteristics of strategies, by encoding and clustering state-action trajectories, unsupervisedly identifying different task or strategy patterns, and training classifiers to enable agents to recognize the collective strategy state of the team in real time, thereby supporting dynamic strategy switching and interpretation; third, hierarchical behavior deconstruction methods, for example, work Omidshafiei et al. (2022) proposes a Multi-agent Offline Hierarchical Behavior Analyzer (MOHBA), which uses variational inference to learn joint and individual latent variables from offline trajectories, achieving automatic clustering and hierarchical deconstruction of team overall strategy patterns and individual behavior patterns, with the advantage of requiring no reward signals and being able to discover high-level behavior concepts consistent with human expert cognition, but also relying on large amounts of trajectory data, and the interpretation process requires subsequent visualization analysis. Overall, the common advantages of these methods are their ability to adaptively discover strategy structures and reduce dependence on prior knowledge; the challenges they face include high model complexity, the need for careful tuning of hyperparameters such as the number of clusters, and sensitivity to data quality and distribution.

Symbolic planning and key state abstraction

The core of this category of techniques is to leverage formal planning languages or data-driven analysis to extract high-level temporal goal sequences or representative key states from task logic, thereby providing an understandable macro blueprint and visual anchors for team strategies. Representative method (Mu et al. 2025) uses a *Hierarchical Task Network planning language (HDDL)* to establish a temporal logic model for team tasks, generating interpretable subgoal sequences, and designs a bidirectional feedback mechanism to allow symbolic planning to co-evolve with the exploration experience of low-level agents, significantly improving team efficiency and interpretability in complex sparse reward environments. Another work (Pandya et al. 2024) adopts a data-driven approach, first clustering

team strategy vectors to obtain discrete macro strategy types, then extracting “*landmark states*” shared by all trajectories as visual summaries of the strategy, and providing them to users combined with visualization and natural language descriptions. The former provides rigorous logical explanations but relies on initial symbolic model design; the latter provides intuitive visual explanations but faces scalability challenges when strategy patterns are numerous.

Hierarchical visualization and memory abstraction

This category of techniques aims to compress complex team spatiotemporal behaviors into high-level representations that are easily understandable by humans through visualization and summarization means such as heatmaps, environmental abstraction, or memory encoding. Representative framework (Tabrez et al. 2024) uses METIS graph partitioning to hierarchically abstract the environment, combined with dynamic bimodal visual guidance (switching between regional heatmaps and local action arrows), synchronizing the cognition of human–machine teams, optimizing division of labor and scalability in large-scale spatial search tasks. Zhang et al. (2023) employs position heatmap technology, aggregating individual movement trajectories from successful experiences into spatial target symbols such as “gathering points”, combined with an asynchronous deadline mechanism to guide agent decisions, suitable for sparse reward environments. Wai et al. (2024) presents a method that abstracts low-level action sequences into concise, high-level strategic descriptions. This is achieved by a two-stage process of selecting key events and removing redundancies, producing text that is suitable for strategy review and instructional purposes. These methods greatly enhance the intuitive understandability of behaviors, but they are typically post-hoc explanations, and their effectiveness relies on the quality of empirical data or increases system complexity.

In summary, team behavior abstraction and high-level goal recognition techniques significantly enhance the understandability of collaborative strategies by establishing semantic bridges between microscopic actions and macroscopic tactics. To systematically organize the technical context of this category of methods, Table 18 provides a detailed comparison and summary of its main technical branches across multiple dimensions such as technical approach, core mechanism, representative literature, advantages, and limitations. Methods in this category significantly enhance team strategy interpretability by mapping low-level behavioral sequences to high-level strategic semantics. Substantial progress has been achieved in automated strategic pattern identification and semantic labeling, with discrete pattern discovery and symbolic planning demonstrating particularly strong explanatory potential. Future research should prioritize modeling dynamic team structure evolution to overcome limitations present in existing approaches—such as static grouping assumptions and hyperparameter sensitivity observed in Ruan et al. (2023) and Duan et al. (2024a). Furthermore, deeper integration of abstract strategy representations with domain knowledge—building upon bidirectional feedback mechanisms demonstrated in Mu et al. (2025)—proves crucial for generating more insightful and actionable high-level explanations.

4.4.2 Macro-state modeling and phase analysis

This category of techniques focuses on answering the question, “*How does a team strategy evolve over time and transition between phases?*” Its technical core lies in the dimensionality reduction, structured analysis, and visualization of the team’s joint state space to identify

key macroscopic states, temporal patterns, and critical transition points within the collaborative process. In the interpretability framework, these methods operate at the *dynamic process layer*, aiming to reveal the “dynamic architecture” and intrinsic rhythm of team behavior. This category represents a temporal deepening of “behavior abstraction” (which answers “what”), characterizing the lifecycle of strategies. It provides observable phenomena concerning system dynamics for causal emergence theory (which examines team-level emergence claims) and identifies the strategic phases requiring in-depth investigation for counterfactual analysis (which tests key decision points).

Dynamic relationship modeling and gradient analysis

This approach quantifies the instantaneous interaction strength or causal dependencies among agents to reveal the dynamic evolution of leadership and collaborative relationships within a team. A typical example (Chen et al. 2025a), in a cooperative box-pushing task, quantifies instantaneous dependencies by calculating the gradient of one agent’s action on another agent’s state. This clearly reveals how dependency hierarchies dynamically emerge and switch across task phases like “linear pushing” and “turning”, transforming abstract hierarchical relations into computable metrics for the first time. This method offers profound insight into the micro-mechanisms of small-scale, tightly-coupled systems but is difficult to scale directly to large-scale MASs.

Spatio-temporal graph modeling and strategy evolution analysis

This approach models and analyzes the long-term spatio-temporal evolution patterns of team strategies by constructing and continuously updating explicit agent relation graphs, or by integrating evolutionary algorithms. Peng et al. (2025b) innovatively combines Evolutionary Algorithms (EAs) with GNNs. Its graph relation module dynamically models agent interactions, while its graph decomposition module utilizes structural information for the value function, thereby providing intuitive insight into team spatio-temporal collaboration patterns while effectively balancing exploration and cooperation. This framework performs excellently in complex dynamic environments, but the construction of the graph model and maintenance of the evolutionary population incur significant computational and communication overhead.

Multimodal behavior recognition and quantitative analysis

This approach transforms agent trajectory data into features and employs machine learning methods such as dimensionality reduction, clustering, and classification to quantitatively identify, distinguish, and verify complex behavioral modalities that emerge from a team under specific strategies. For example, Sharma et al. (2022) provides a replicable framework for quantitative behavioral analysis in a pursuit-evasion game. This is achieved by converting evader position data into multi-granularity heatmap features, reducing dimensionality via PCA, and employing a classifier to successfully identify and validate emergent strategies like “bimodal orbiting”. Such methods can objectively discover non-intuitive behavioral patterns, but the validity of their conclusions heavily depends on the design of the simulation environment, and their generalizability to real-world complex scenarios requires further verification.

In summary, macroscopic state modeling and phase analysis techniques shift the explanatory focus from static structure to dynamic process, profoundly revealing the intrinsic rhythm and phased patterns of team strategy evolution. To clarify the characteristics and applicable boundaries of different technical paths, Table 19 systematically reviews and compares this category of methods from perspectives such as core mechanism, key tools, and

representative studies. While existing methods demonstrate maturity in capturing phased transitions and spatiotemporal patterns, several fundamental challenges remain unresolved. Future research directions should prioritize:

- Developing more precise quantitative indicators for phase transitions to enable automated prediction of strategic critical points and tipping points in team behavior.
- Constructing cross-scale causal dynamical models to elucidate the mechanistic links between micro-level interactions and emergent macro-patterns, extending principles such as the dependency quantification approaches introduced in Chen et al. (2025a).
- Investigating adaptive phase transition mechanisms suitable for non-stationary environments, thereby enhancing methodological applicability in open and dynamic scenarios through generalization of spatiotemporal modeling approaches such as those introduced in Peng et al. (2025b).

4.4.3 Theoretical frameworks based on causal emergence

This category of techniques aims to answer the question, “*Is a team strategy a real and irreducible functional whole?*” Its technical core is to employ mathematical tools from complex systems theory (e.g., nonlinear dynamics, causal emergence theory) to quantitatively argue that new, stronger causal laws exist at the team’s macroscopic level and that these laws may not be reduced to the level of individual behavior. In the interpretability framework, these methods reside at the **theoretical foundation layer**, striving to provide a principled explanation for the “ontological status” of team-level intelligence. It goes beyond the phenomenological descriptions of the previous two categories (what it is, how it changes), laying a philosophical and scientific foundation for the assertion of the “team as a causal whole”. It thereby provides fundamental theoretical support for all research aimed at understanding and optimizing team-level strategies and behaviors.

Dynamics modeling of collective learning: This approach employs theoretical tools from Complex Systems Science (CSS) to model the multi-agent collective learning process itself as a nonlinear dynamical system, aiming to qualitatively and quantitatively argue for the irreducibility of team-level macroscopic behavior from first principles. Its foundational work by Barfuss et al. (2025) propose a “Collective Reinforcement Learning Dynamics (CRLD)” framework, analyzing the formation, stability, and evolution of cooperation patterns by introducing concepts such as multi-stability, basins of attraction, critical transitions, and hysteresis. This framework provides a lightweight yet profound qualitative theoretical explanation for the “team as a functional whole” from a dynamical systems perspective, capable of dissecting complex phenomena like cooperation stability. However, the framework currently relies on low-dimensional, deterministic idealized assumptions, making it difficult to directly capture the full complexity of high-dimensional, highly stochastic real-world MARL systems. The practical application and computability of its theoretical models remain core challenges for the future.

Theoretical frameworks grounded in causal emergence provide rigorous, mathematically formalized explanations for the “wholeness” of MASSs, uniting philosophical depth with analytical precision. However, this research direction remains predominantly theoretical, as current models rely on low-dimensional and idealized assumptions that limit direct appli-

cability to high-dimensional, real-world MARL problems. Critical future research tasks include:

- Developing scalable approximation methods for analyzing high-dimensional CRLD to bridge the gap between theoretical elegance and practical applicability.
- Establishing quantifiable relationships between measures of macro-level causal informativeness and team task performance metrics, thereby connecting theoretical explanation with practical design principles for collaborative systems.
- Integrating these theoretical constructs with practical techniques—such as behavior abstraction and world models—to provide stronger foundation for interpreting observed strategic patterns and emergent behaviors.

4.4.4 Counterfactual team-level analysis

This category of techniques addresses the causal question: “*What are the key decision factors for team success? What would happen if the strategy were changed?*” Its technical core is to systematically construct and evaluate counterfactual scenarios at the team level to investigate the necessity, vulnerabilities, and alternative possibilities of strategic choices. In the interpretability framework, these methods operate at the **causal testing layer**, dedicated to evaluating the causal structure and robustness of team strategies. It serves as a hypothetical validation of the key phases or decision points identified by macroscopic state modeling, enabling the diagnosis of the underlying logic and dependencies of the strategies described by the first two categories. It provides a unique causal diagnostic perspective for strategy optimization and acts as a crucial bridge connecting phenomenological description with strategic intervention.

Causal querying and diagnosis of team-level strategies: This approach involves building query systems that support counterfactual reasoning for trained team policies, allowing researchers to directly conduct “what-if” causal inquiries into the team’s collective decision logic to diagnose policy robustness, key decision factors, and vulnerabilities. Its representative framework Boggess et al. (2022) abstract learned MARL policies into a multi-agent MDP to construct a natural language explanation interface supporting three types of queries: “When” (under what team state an action is taken), “Why Not” (why an alternative action was not taken), and “What” (what the team’s decision is). The “Why Not” query implements typical counterfactual reasoning. By using relevance filtering to automatically focus on key factors, this framework effectively enhances human understanding and trust in team strategies. Research in this direction remains scarce, and aspects such as the automation level of such system frameworks, their ability to support long-horizon sequential decisions, and deep integration with the learning process all warrant further exploration.

Counterfactual team-level analysis provides a unique perspective grounded in causal reasoning principles, serving as a powerful methodological tool for evaluating and optimizing team strategies. Nevertheless, this research direction remains notably under-explored and has yet to coalesce into a systematic body of work with established theoretical foundations. The future potential of this approach is substantial, with promising research directions including:

- Automating and scaling the computational efficiency of counterfactual reasoning pro-

cesses to minimize reliance on manually constructed queries and enable application to complex, large-scale MASs.

- Extending analytical scope from single decision points to complete temporal policy sequences to assess long-term, cascading counterfactual outcomes and strategic dependencies.
- Deeply integrating counterfactual reasoning with RL techniques such as curriculum learning and opponent modeling, enabling active generation of scenarios that expose policy weaknesses or explore novel strategy spaces, thereby supporting the identification of policy vulnerabilities and alternative strategy patterns.

4.4.5 Learning-based world models and team behavior generation

This category of techniques aims to answer the question, “*What fundamental laws or generative rules govern team behavior?*” Its technical core is to learn generative models (e.g., interpretable reward functions, symbolic rules, procedural knowledge) capable of simulating the team’s macroscopic dynamics, thereby reverse-engineering the underlying logic from the internal mechanisms that drive behavior. In the interpretability framework, these methods reside at the **generative mechanism layer**, extending explanation from observed behavioral patterns toward candidate generative representations. It attempts to provide a unified, mechanistic generative explanation for the semantic intent, dynamic patterns, holistic principles, and causal key points respectively described by the previous four categories. As a cutting-edge direction pursuing the deepest level of interpretability, it aims to open the “black box” of team collaborative behavior generation.

Semantic reward and observation generation

This approach leverages generative models with rich semantic knowledge, such as LLMs, to automatically synthesize readable, interpretable reward functions or environment observation augmentation functions, thereby making the “rules” that drive team behavior transparent and semantic. The representative framework by Wei et al. (2025) uses an LLM to generate code drafts for mixed reward functions and observation augmentation functions based on task descriptions, then iteratively optimizes these semantic functions via an evolutionary algorithm, forming a “generate-evaluate-feedback” loop. This method can automatically produce semantic driving signals that align with human intuition, reduces the burden of manual design, and has shown improved performance in partially observable cooperative tasks. Its limitations include dependence on the quality and stability of LLM generation, the high computational cost of the evolutionary process, and the need for further validation of its effectiveness in more complex environments.

Comprehensive interpretability toolbox construction

This approach targets safety-critical or logically complex domains, aiming to integrate and implement a comprehensive set of multi-perspective, actionable interpretability techniques for hierarchical or black-box multi-agent policy systems, providing full mechanistic explanations from micro-decisions to macro-strategy. In air combat simulation, Selmonaj et al. (2025a) adopted a systematic approach that integrates a suite of methods—including policy simplification, reward decomposition, Shapley value-based feature contribution analysis, structural causal models, and hierarchical policy analysis—to deliver multidimensional, data-driven explanations of tactical behaviors. The strength of this work lies in its close integration with domain requirements, offering a multi-faceted explanation covering

short-term actions to long-term strategy. Its drawbacks include the high complexity of some methods, making real-time application difficult, and the depth and accuracy of explanations relying to some extent on the design of domain knowledge.

Learning-based world model techniques pursue explainability at the level of generative principles, representing a cutting-edge frontier in MAS interpretability. Existing methods have made significant strides in areas such as reward decomposition, semantic reward generation, and multi-faceted mechanistic explanation. Promising future research avenues comprise:

- The deep integration of LLMs—inspired by semantic reward generation methods such as those in Wei et al. (2025)—to develop interactive, dialog-capable explanation systems that bridge the semantic gap between technical models and human understanding.
- The construction of fully interpretable symbolic or procedural world models to identify interpretable candidate rules for representing complex team behaviors, enabling more transparent and controllable MASs.
- The exploration of brain-inspired team cognition models that simulate human-like collective decision-making under uncertainty, thereby producing collaborative behaviors and explanations that are more naturally aligned with human intuition and reasoning patterns.

4.4.6 Summary and research directions for team strategy and emergent behavior

This section provides a comprehensive summary and comparative analysis of the five technical paradigms for the explainability of team-level strategies and emergent behaviors. It then synthesizes their intrinsic connections into a cohesive, multi-perspective interpretability framework. Finally, it discusses current limitations and outlines promising future research directions.

Comprehensive summary of technical paradigms

This section systematically elaborated on the five technical paradigms for team-level strategies and emergent behavior explainability. A comprehensive comparison of these paradigms is presented in Table 6, which contrasts their respective technical categories, core explanatory questions, advantages, and limitations.

This section has systematically examined explainability techniques for team strategy and emergent behavior through five distinct analytical perspectives: *Behavioral Abstraction* 4.4.1, *Macro-State Modeling* 4.4.2, *Causal Emergence* 4.4.3, *Counterfactual Analysis* 4.4.4, and *Generative World Models* 4.4.5. The five technical paradigms address the fundamental questions regarding a strategy's "nature", "dynamics", "efficacy", "components", and "formation". Collectively, these paradigms establish a comprehensive framework for understanding how collective intelligence manifests in MASs, spanning from phenomenological description to mechanistic explanation of team-level behaviors.

Team behavior abstraction 4.4.1 map low-level action sequences with high-level semantic labels via strategy clustering (Ruan et al. 2023) and symbolic planning (Mu et al. 2025), establishing a semantic bridge between micro and macro levels but often suffers from static grouping and scalability challenges. **Macro-State Modeling and Phase Analysis** 4.4.2 leverage dynamical systems theory (Chen et al. 2025a) and spatio-temporal graph modeling (Peng et al. 2025b) to reveal the dynamic architecture and phased evolution of

Table 6 Comprehensive comparison of explainability techniques for team strategy and emergent behavior

Technique category	Core explanatory question	Representative techniques	Explanatory focus	Core advantages	Inherent limitations & challenges
Team behavior abstraction & high-level goal identification	What strategy is the team executing as a whole?	Group clustering via VQ-VAE (Ruan et al. 2023); Hierarchical symbolic planning (Mu et al. 2025); Landmark state extraction (Pandya et al. 2024); Hierarchical behavior deconstruction (Omidshafiei et al. 2022)	Strategic intent, high-level semantic labels, hierarchical behavior concepts	Establishes micro-macro semantic bridges, enhances strategy understanding; Some methods (e.g., Omidshafiei et al. (2022)) support reward-free offline hierarchical analysis	Static grouping, parameter sensitivity, reliance on domain knowledge/data quality, poor scalability; Some methods (e.g., Omidshafiei et al. (2022)) require post-hoc visualization for interpretation
Macro-state modeling & phase analysis	How does the team strategy evolve and transition between phases?	Strategy gradient dependency quantification Chen et al. (2025a); Spatio-temporal graph evolutionary modeling (Peng et al. 2025b)	Dynamic architecture, phase transition, spatio-temporal patterns	Reveals intrinsic rhythm and critical transition points of strategies	High computational overhead, weak generalizability, lack of quantitative phase transition metrics
Theoretical frameworks based on causal emergence	Is team strategy a real, irreducible functional whole?	Collective Reinforcement Learning Dynamics (CRLD) (Barfuss et al. 2025)	Ontological status, causal independence	Provides philosophical and mathematical foundations for holism	Theoretical models are idealized and difficult to apply to high-dimensional complex systems
Counterfactual team-level analysis	What factors are key to success? What if strategies changed?	Team-level natural language counterfactual query framework (Boggess et al. 2022)	Causal structure and robustness, strategy vulnerabilities	Provides a unique causal reasoning perspective, guiding strategy diagnosis and optimization	Limited research, no established system, lacking automation and scalability
Learning-based world models & team behavior generation	What fundamental rules generate team behavior?	LLM-generated semantic reward functions (Wei et al. 2025); Multi-perspective tactical explanation toolbox (Selmonaj et al. 2025a)	Generative mechanism, fundamental principle	Pursues deepest level of interpretability from generative principles, representing the cutting-edge frontier	Insufficient real-time interpretability, complex models, reliance on generation quality or prior knowledge

team strategies, yet current methods lack maturity in quantitative phase transition metrics and cross-scale causal dynamics modeling. **Theoretical Frameworks based on Causal Emergence 4.4.3** provide rigorous mathematical foundations for treating team strategy as irreducible causal entities (Barfuss et al. 2025), though their practical application to high-dimensional real-world systems remains limited. **Counterfactual Team-Level Analysis 4.4.4** offers contrastive analyses of alternative strategy outcomes and robustness through systematic scenario evaluation (Boggess et al. 2022), but represents a nascent research

direction lacking systematic frameworks and automation. Finally, **Learning-Based World Models and Team Behavior Generation 4.4.5** pursues explainability from the depth of generative principles via methods such as symbolic knowledge fusion (Mu et al. 2025) and LLM-generated semantic rewards (Wei et al. 2025), representing the frontier of the field.

Intrinsic connections: a multi-perspective explanatory framework

The five paradigms are not an isolated list of literature, but rather constitute a functionally complementary, multi-perspective explanatory system that collectively provides an organizing framework for understanding team strategies and emergent behaviors. This systematic nature is manifested in three aspects:

- **Complementary Explanatory Perspectives:** The explanatory focus progresses from external behavioral manifestations to internal generative mechanisms. Behavior abstraction answers “*what* the team is doing,” providing the foundational semantic layer of understanding; macroscopic modeling reveals “*how* team behavior evolves,” characterizing its dynamic process; causal emergence argues “*why* the team can act as a whole,” establishing its ontological foundation; counterfactual analysis tests “*why* a team decision is critical,” dissecting its causal structure; and generative models provide “*how come* team behavior originates,” revealing its underlying driving rules. Each layer provides the context and foundation for the deeper analysis of the next.
- **Taxonomic Coverage:** The five categories systematically address the most fundamental questions about team strategies and emergent behaviors from five essential perspectives: *semantics (What-Strategy)*, *dynamics (How-Evolve)*, *ontology (Whether-Emergent)*, *causality (Why-If-Critical)*, and *mechanism (How-Come-Generated)*.
- **Functional Complementarity:** They mutually supplement each other to approach a complete understanding of team strategies and emergent behaviors. For instance, understanding the emergent “cooperative containment” strategy in a multi-agent pursuit task may require: first, using *team behavior abstraction* to identify that the overall strategy is “flanking and encircling” (semantics); then, employing *macroscopic state modeling* to analyze how the team formation transitions from a “dispersed” phase to an “encirclement” phase (dynamics); subsequently, relying on *causal emergence theory* to argue that the containment effect is a product of the team’s collective effort, not the sum of individual actions (ontology); using *counterfactual team-level analysis* to verify “whether the containment would fail if the flanking agent were absent” (causality); and finally, discovering through a *learning-based world model* that this behavior is driven by a team reward function for “minimizing the target’s escape area” (mechanism).

Limitations and future directions

The interpretation of team strategies and emergent behaviors currently remains predominantly oriented toward post-hoc or diagnostic analysis, lacking the ability to actively guide system optimization. The causal dynamic links between macroscopic phenomena and microscopic mechanisms are unclear, and evaluation heavily relies on subjective qualitative analysis. To this end, key future breakthroughs lie in:

- **Active Guidance Mechanisms:** Leveraging strategy patterns identified by team behavior abstraction (Ruan et al. 2023) or vulnerability analyses from counterfactual reasoning (Boggess et al. 2022) to construct curriculum learning or meta-learning signals,

thereby actively shaping more robust and efficient team collaborative behaviors.

- **Cross-Scale Causal Modeling:** Developing theoretical and computational models that connect macroscopic team state phase transitions with microscopic agent interaction causal graphs (Meli et al. 2025), quantitatively revealing the causal mechanisms through which microscopic interactions drive macroscopic strategic shifts.
- **Computable Causal Emergence:** Promoting the practical application of causal emergence theoretical frameworks (Barfuss et al. 2025) and developing approximation computation and analysis methods suitable for high-dimensional MARL systems.
- **Comprehensive Evaluation Paradigm:** Establishing integrated evaluation methodologies that combine high-fidelity simulation validation, formal metrics (e.g., policy stability), and systematic human experiments.

To advance beyond current limitations and establish the next generation of trustworthy multi-agent collaborative systems, future research must coordinate progress across four critical frontiers:

- **From Passive Analysis to Proactive Guidance:** Future work should transcend post-hoc explanation paradigms by leveraging interpretability techniques to actively guide the generation and optimization of team strategies. Strategy patterns identified through behavior abstraction (Ruan et al. 2023) and vulnerabilities exposed by counterfactual analysis (Boggess et al. 2022) could serve as signals for curriculum learning or meta-learning, actively shaping more robust and efficient collaborative behaviors while maintaining explanatory transparency throughout the optimization process.
- **Constructing Cross-Scale Causal Dynamical Models:** Addressing the explanatory gap in macro-state modeling requires developing formal frameworks that link macroscopic phase transitions with microscopic interaction mechanisms. Research should focus on building cross-scale causal dynamical models that quantitatively connect micro-level interaction causal graphs (Meli et al. 2025) with macro-level strategy shifts, enabling mechanistic understanding of how specific local interactions drive global behavioral patterns.
- **Advancing Practical Causal Emergence Frameworks:** A crucial objective involves transitioning causal emergence theories (Barfuss et al. 2025) from idealized models to practical applications through scalable approximation methods suitable for high-dimensional, partially observable MASs. Establishing quantitative relationships between causal emergence strength and team task performance would enable these frameworks to not only analyze macro-level causal organization in collective behavior but also guide the systematic design of efficient collaborative systems.
- **Establishing Integrated Evaluation Paradigms:** Mitigating the subjectivity inherent in current evaluations requires developing comprehensive assessment frameworks that integrate high-fidelity simulation verification, formal metrics (e.g., policy stability, adaptation speed), and systematic human-subject studies. Such integrated paradigms would provide systematic measurement of the practical utility and cognitive efficiency of team strategy explanations across diverse application domains.

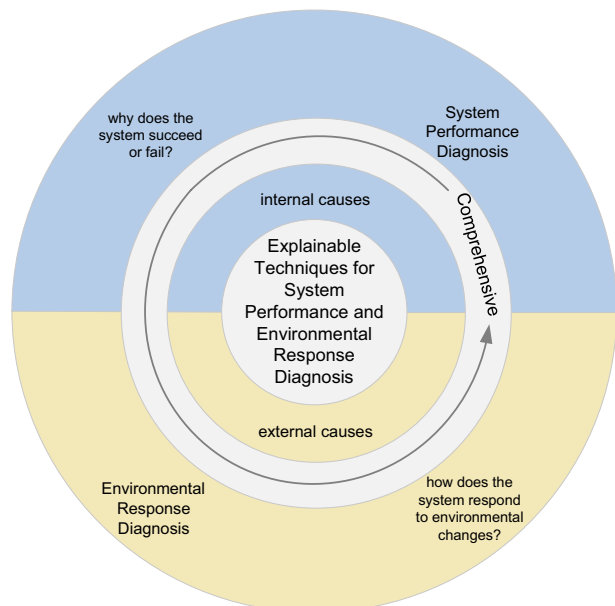
In conclusion, the advancement of explainability for team strategy and emergent behavior hinges on integrated progress across proactive guidance, cross-scale modeling, practical

theoretical frameworks, and rigorous evaluation. By addressing the fundamental limitations of current approaches through cohesive methodological development, these explanation frameworks are positioned to form the foundation for next-generation MASs capable of sophisticated, transparent, and trustworthy collective intelligence in complex operational environments.

4.5 Explainable techniques for system performance and environmental response diagnosis

Ensuring the reliable and efficient operation of a MAS in complex scenarios requires not only knowing its performance but, more critically, understanding the **causes of its performance and its performance limits**. Therefore, interpretability techniques focused on overall system behavior aim to establish a diagnostic framework that systematically addresses two fundamental questions (as shown in Fig. 7): First, **why does the system succeed or fail?** (Focusing on internal causes—the mechanisms of collaboration, learning, and resource allocation within the system). Second, **how does the system respond to environmental changes?** (Focusing on external causes—the system’s adaptability and robustness to external perturbations, non-stationarity, and boundary conditions). Together, these questions define the complete diagnostic problem space for assessing the “health” and “adaptive capacity” of an MAS. To answer these questions, current research has developed two complementary technical paradigms: **System Performance Diagnosis** and **Environmental Response Diagnosis**. The former acts as a **comprehensive system check-up**, utilizing performance attribution, bottleneck identification, and learning process analysis to reveal internal collaborative efficiency and decision logic. The latter functions as a *stress test*, employing sensitivity assessment, failure mode analysis, and boundary demarcation to examine the system’s survivability and failure boundaries in dynamic, open environments. These two paradigms are not isolated; rather, they provide complementary tools for system

Fig. 7 Technical framework for explaining system performance and environmental response diagnostics: core issues and two complementary technologies



behavior diagnosis, moving from **internal cause analysis** to **external cause evaluation**. The following sections elaborate on how these techniques individually focus on and collaboratively cover the aforementioned problem space.

4.5.1 System performance diagnosis

This category of methods is dedicated to answering the question, “*Why does the system succeed or fail?*” Their technical core lies in examining the modeled mechanisms associated with collaboration and performance generation within an MAS by quantifying individual contributions, analyzing learning dynamics, and identifying scalability constraints. Within the interpretability diagnostic framework, these methods focus on the **internal causes** of the system, aiming to open the black box of performance to understand the processes of credit assignment, policy coordination, and learning evolution among agents. The main technical approaches—**performance attribution**, **bottleneck identification**, and **learning process diagnosis**—together provide complementary diagnostic perspectives from static contribution analysis to dynamic training intervention. Performance attribution (e.g., difference rewards and causal discovery) provides a basis for understanding “*who is contributing*”; bottleneck identification attempts to locate the shortcomings limiting system scaling; and learning process diagnosis directly monitors and optimizes training dynamics. These three approaches can be used in combination, aiming to provide a direct, actionable internal perspective for algorithm improvement and system debugging.

System-level performance attribution analysis

Researchers focus on developing quantitative tools to reveal individual contributions to collective success. Early work emphasized efficient and practical contribution metrics. For example, Mahjoub et al. (2023) proposed an agent importance evaluation method based on difference rewards, which approximates individual contribution with linear time complexity by comparing the team reward difference between an agent’s actual participation and executing a “no-op” action. This method successfully revealed, for instance, that in the LBF environment, the MAA2C algorithm exhibits a more balanced distribution of agent contributions compared to MAPPO, thereby explaining its superior collaborative performance. However, such methods provide a correlational, rather than causal, attribution. To gain deeper insight into the causal mechanisms underlying performance, subsequent research shifted towards more explanatory causal discovery frameworks. Wang et al. (2025a) developed the MACCA method, which utilizes Dynamic Bayesian Networks to infer a structured causal graph of team reward generation from offline interaction data, clarifying the causal pathways from individual agent states and actions to the final outcome, thereby offering a modular and interpretable perspective on credit assignment. A complementary path focuses on deconstructing the underlying capability dimensions behind system performance. Tesic et al. (2025) constructed a hierarchical Bayesian model to attribute a system’s score in complex environments (e.g., Melting Pot) to a set of latent collaborative capabilities (e.g., reciprocity, convention-following ability) and a base performance parameter. Through counterfactual reasoning, this study suggested that, in certain evaluation tasks, a high system score may stem primarily from high base performance (e.g., excelling at tasks not requiring collaboration) rather than collaborative performance, raising important questions about the completeness of existing evaluation benchmarks. Furthermore, for specific application scenarios (e.g., wireless network resource management), Iturria-Rivera et al. (2024) adopted

reward decomposition techniques, breaking down the composite team reward function into interpretable sub-goals such as delay, packet loss rate, and data rate. Analysis of the decomposed value functions provided an interpretable account of how reward components were associated with system performance. These methods collectively form a complete spectrum from surface-level contribution quantification to deep causal and capability analysis, yet they face trade-offs between computational efficiency and explanatory depth, as well as between generality and domain dependency.

Bottleneck identification and scalability analysis

Although a mature sub-field has not yet fully formed, existing works have touched upon related issues within their performance attribution analyses. For instance, the severe credit assignment imbalance observed by Mahjoub et al. (2023) when quantifying contributions essentially reveals potential bottlenecks that constrain system performance scaling under specific algorithms or scenarios.

Learning process and convergence diagnosis

Research aims to improve learning efficiency by monitoring and intervening in training dynamics in real time. A representative work is the AELA method proposed by Yoo et al. (2025), which innovatively uses policy entropy or Q-value entropy as a proxy indicator for learning progress and dynamically adjusts the length of training episodes accordingly. This method can shorten episodes early in training to avoid ineffective exploration and extend episodes during policy convergence to learn long-term dependencies, thereby significantly accelerating convergence and improving final performance in complex environments like SMAC. This exemplifies a closed-loop approach that translates diagnostic insights directly into optimization interventions.

System Performance Diagnosis techniques aim to unveil the black box of internal collaboration and performance generation in MARL systems, with their methodologies showing an evolutionary trend from surface-level contribution quantification to deep causal mechanism mining. To systematically compare the strengths and weaknesses of existing methods, Table 20 summarizes the core characteristics of representative approaches, which can be broadly categorized into two paradigms. The first category, represented by efficient and practical techniques such as difference rewards (Mahjoub et al. 2023) and reward decomposition (Iturria-Rivera et al. 2024), focuses primarily on identifying *which* agents or actions influence system efficacy through computationally efficient mechanisms. In contrast, the second category employs more sophisticated analytical frameworks including causal discovery (Wang et al. 2025a) and latent capability modeling (Tescic et al. 2025), which aim to reveal underlying mechanisms of *why* and *how* these influences occur at substantially higher computational costs, representing a fundamental trade-off between explanatory depth and practical efficiency.

Current diagnostic methodologies face several significant challenges that constrain their practical application. A fundamental limitation concerns the inherent trade-off between computational efficiency and explanatory depth (Wang et al. 2025a; Tescic et al. 2025), which remains unresolved within existing technical frameworks. More critically, most current research remains confined to post-hoc diagnostic paradigms, lacking mechanisms for real-time feedback integration between diagnostic insights and learning processes. Consequently, transitioning from static diagnosis to dynamic optimization represents a crucial research direction requiring substantial methodological innovation and paradigm shifts in how diagnostic information is utilized within learning systems.

To address these limitations and advance the field toward more practical and impactful diagnostic frameworks, we identify three key research priorities that build upon current methodological foundations while addressing their core limitations:

- **Development of Lightweight Causal Diagnosis Frameworks:** While causal methods such as Wang et al. (2025a) provide high interpretability, their substantial computational overhead precludes real-time application in large-scale systems. Future investigations should prioritize development of approximate causal discovery algorithms and distributed inference architectures that retain explanatory depth while achieving computational efficiency required for practical deployment in complex, real-world MASs.
- **Towards Closed-Loop Diagnostic-Optimization Systems:** Future diagnostic paradigms must transcend passive explanation and evolve into active system components that directly influence learning and decision processes. Building upon dynamic intervention principles demonstrated in works like Yoo et al. (2025), research should explore utilization of diagnostic outputs—such as contribution assessments from Mahjoub et al. (2023) or capability metrics from Tesic et al. (2025)—as real-time feedback for adaptive credit assignment, structured curriculum learning, or dynamic agent role specialization, thereby establishing autonomous “diagnosis-decision-optimization” closed loops that continuously improve system performance.
- **Integration of Semantic and Quantitative Explanations:** To bridge the gap between quantitative metrics and human-understandable reasoning, future work should explore integration of LLMs and semantic technologies to translate numerical contributions, causal chains identified by methods like Wang et al. (2025a), and latent capability dimensions from Tesic et al. (2025) into natural language narratives enriched with domain knowledge, rendering explanations more insightful and actionable for human experts while maintaining mathematical rigor.

4.5.2 Environmental response diagnosis

This category of methods focuses on answering the question, “*How does the system respond to environmental changes?*” Its technical core lies in assessing the adaptability, robustness, and failure modes of a MAS when confronted with external environmental perturbations, non-stationarity, and boundary conditions. Complementing System Performance Diagnosis, this class of methods addresses the **external causes** that influence system behavior, aiming to examine the boundaries of reliable operation and assess potential failure mechanisms. It encompasses three progressively deeper analytical dimensions:

Environmental sensitivity analysis and robustness assessment

In the absence of a systematic evaluation framework, representative work such as Barnett et al. (2025) proposes a cooperative behavior measurement framework based on counterfactual boundary condition analysis. This approach distinguishes genuine from spurious cooperation, thereby revealing the sensitivity of such behavior to contextual factors and delineating its boundaries. However, this approach faces challenges in computing best responses and overlooks the dynamic adaptation of policies. Constructing a standard framework capable of measuring an MAS’s ability to maintain performance under varying perturbation intensities is a critical step bridging theoretical validation and practical deployment, requiring further investigation in future research.

Explanatory frameworks for dynamic adaptation mechanisms

A core challenge in environmental response diagnosis involves understanding and explaining how MARL systems adjust policies online in response to non-stationary environmental dynamics—a capability essential for real-world deployment. Currently, few studies directly address the interpretation of dynamic adaptation mechanisms at the policy level. Currently, there is a lack of real-time diagnostic tools akin to Yoo et al. (2025) for this purpose, with most methods still confined to post-hoc analysis. How to attribute and interpret the real-time adaptive behavior of a trained or actively running MARL system is a significant and underexplored field.

Analysis of failure modes and boundary conditions

This sub-area currently represents the most concentrated research focus within environmental response diagnosis, primarily involving diagnosis of root causes for specific system behaviors or failures through counterfactual reasoning and formal verification techniques that reveal system boundary conditions and limitations. The methodologies in this domain employ complementary approaches ranging from human-centric explanations to mathematically rigorous verification, each addressing different aspects of system reliability assessment.

Formal verification and temporal logic analysis emphasizes mathematical rigor and verification completeness for applications requiring certified reliability. Methods such as those in Boggess et al. (2023) employ probabilistic model checking and temporal logic formulae to perform formal root-cause analysis on behavioral deviations in MASs. This approach offers mathematical rigor and verification completeness but suffers from high computational complexity and potentially limited expressiveness of the logical specifications.

Counterfactual cooperation metrics As seen in Barnett et al. (2025), counterfactual boundary-condition analysis quantifies the degree of cooperation by computing the social welfare difference between an agent’s current policy and a selfish best-response policy, given the fixed policies of other agents. This method provides clear criteria and intuitive explanations but is sensitive to hyperparameters and struggles to compute best responses in complex environments.

Environmental Response Diagnosis focuses on evaluating the adaptability and robustness of MARL systems in non-stationary environments. Current research in this area is characterized by a progression from static post-hoc analysis towards more dynamic reasoning, with a strong reliance on techniques such as counterfactual inference and formal verification. The representative methods can be compared across several key dimensions, as summarized in Table 21. This landscape reveals diverse technical approaches, each with distinct methodological strengths tailored to specific diagnostic goals. For instance, the formal method adopted by Boggess et al. (2023) ensures diagnostic precision and completeness through mathematical rigor, using probabilistic model checking. Complementing this, Barnett et al. (2025) employs systematic counterfactual analysis to delineate the boundary conditions of collaborative behaviors, quantifying policy sensitivity to environmental parameters and yielding practical insights for robust system design.

Notwithstanding these contributions, the field remains nascent with current research predominantly characterized by focus on post-hoc root-cause analysis, as exemplified by Gyevnár et al. (2025) and Boggess et al. (2023), alongside offline evaluation frameworks such as proposed by Barnett et al. (2025). Critical gaps persist in investigating internal mechanisms governing system adaptation to non-stationary environmental conditions dur-

ing online operation, with corresponding explanatory tools for elucidating these dynamic processes remaining substantially underdeveloped. This represents a significant limitation given the fundamental importance of adaptation capabilities for real-world MAS deployment in dynamically changing environments.

To address these limitations and advance environmental response diagnostics toward practical utility in real-world applications, future research should prioritize the following interconnected directions that build upon current methodological foundations while addressing critical capability gaps:

- Developing Online Interpretability for Dynamic Adaptation:** Future work should prioritize creation of techniques that monitor and explain in real-time how systems adapt to sudden environmental shifts, including introduction of novel agents or changes in operational rules. This direction aims to fill the critical gap in “dynamic adaptation mechanism interpretation” for which dedicated diagnostic tools are currently absent, supporting proactive monitoring and earlier identification of potential failures in dynamic deployment scenarios.
- Enhancing the Scalability of Formal Methods:** While formal methods as exemplified by Boggess et al. (2023) provide mathematically rigorous diagnostic guarantees, their high computational complexity often renders them intractable in complex, high-dimensional environments. Research should overcome this bottleneck through development of more efficient model-checking algorithms, compositional verification techniques, or integration with approximate learning methods to broaden applicability to practical scenarios without sacrificing diagnostic reliability.
- Establishing a Hybrid Simulation-Real-World Evaluation Framework:** Diagnostic efficacy of methods such as Barnett et al. (2025) depends heavily on underlying simulation quality and fidelity. To address limitations of single-simulation reliance and enhance real-world applicability, research should establish hybrid evaluation paradigms integrating high-fidelity simulation, theoretical counterfactual reasoning (Barnett et al. 2025), and human subjective assessment to provide robust validation foundations, systematically validate LLM-generated explanations, and ultimately enhance the credibility and practical utility of diagnostic results across diverse application domains.

4.5.3 Summary and research directions for system performance and environmental response diagnosis

This section provides a comprehensive summary and comparative analysis of the two technical paradigms for system diagnosis. It then synthesizes their intrinsic connections into an integrated explanatory framework and concludes by discussing current limitations and outlining promising future research directions.

Comprehensive summary of technical paradigms

This section positions interpretability techniques as **diagnostic tools** for MARL systems, systematically elaborating on the two major diagnostic dimensions: System Performance and Environmental Response. Table 7 provides a comprehensive comparison of these two technical categories, showing that while they differ significantly in their core focus, perspective, and tools, they are methodologically complementary, together forming a complete framework for assessing the internal and external behaviors of an MARL system.

Table 7 Comprehensive comparison of system performance and environmental response diagnosis techniques

Technique category	Core explanatory question	Representative techniques	Explanatory focus	Core advantages	Inherent limitations & challenges
System performance diagnosis	Why does the system succeed or fail? What are its internal collaboration & decision mechanisms?	Performance attribution analysis (Difference Reward (Mahjoub et al. 2023), Causal Discovery (Wang et al. 2025a)), Learning process diagnosis (Entropy-driven adjustment (Yoo et al. 2025)), Reward Decomposition (Iturria-Rivera et al. 2024)	Individual contribution quantification, credit assignment, learning dynamics, system scalability	Reveals internal collaboration mechanisms, quantifies agent contributions, provides direct basis for algorithm optimization	Trade-off between computational efficiency and explanatory depth; mostly post-hoc diagnosis, lacking real-time feedback; Sensitivity to model assumptions and hyperparameters
Environmental response diagnosis	How does the system respond to environmental change? Under what conditions does it fail or adapt?	Counterfactual Reasoning (Cooperation Metrics (Barnett et al. 2025)), Formal Verification (Boggess et al. 2023)	System robustness, adaptive behavior, failure root causes, decision boundaries	Provides explanations with causal depth; High mathematical rigor; Clearly reveals boundary conditions of system behavior	Nascent research area with a lack of systematic evaluation frameworks; Insufficient tools for explaining online adaptation mechanisms; High computational complexity hinders application

A macroscopic synthesis reveals the complementary strengths and inherent limitations of each diagnostic paradigm. **System Performance Diagnosis 4.5.1** centers on the internal mechanisms of MARL systems through approaches such as performance attribution (Mahjoub et al. 2023; Tesic et al. 2025), bottleneck identification, and learning process analysis (Yoo et al. 2025), excelling in elucidating intrinsic processes governing inter-agent collaboration, credit assignment, and learning dynamics with methodologies progressively evolving from surface-level contribution quantification (Li et al. 2024f) toward uncovering profound causal mechanisms (Wang et al. 2025a), though its primary limitation lies in the predominant focus on internal system properties that often overlooks critical environmental interactions and adaptation requirements. **Environmental Response Diagnosis 4.5.2**, in contrast, concerns the interaction between the system and the **external environment**. Utilizing technologies like counterfactual reasoning (Barnett et al. 2025) and formal verification (Boggess et al. 2023), it aims to clarify the system's decision boundaries, robustness, and failure root causes. Research in this field remains in its early stages, with significant gaps, particularly in systematic sensitivity assessment and the interpretation of online adaptation mechanisms.

Intrinsic connections: a multi-perspective explanatory framework

The two diagnostic paradigms are intrinsically linked, forming a coherent explanatory framework from complementary perspectives.

- **Logical Progression:** The diagnostic focus expands progressively from the system's interior to its external environment. System Performance Diagnosis addresses “*how the system operates internally*” (mechanism), revealing collaborative logic and credit as-

signment among agents. Environmental Response Diagnosis focuses on “*how the system interacts with the external world*” (adaptation), evaluating its response and robustness to dynamic environments. Together, they complete a systemic examination from internal causes to external causes, and from mechanism to performance.

- **Taxonomic Coverage:** The two methods represent complementary theoretical perspectives for attributing and assessing overall system behavior from two dimensions: “**internal cause diagnosis**” and “**external cause evaluation**.” System Performance Diagnosis answers the “*source*” question of system performance (*Where does performance come from?*), while Environmental Response Diagnosis answers the “*boundary*” question of system behavior (*Where are the limits of behavior?*). They collectively constitute a complete problem space for assessing the *capabilities* and *limitations* of an MARL system, ensuring the comprehensiveness of the diagnostic framework.
- **Complementary Perspectives:** They focus on the **internal** and **external** causes that influence system behavior, respectively. System Performance Diagnosis probes inward, explaining the internal logic of performance generation (e.g., *Which agent’s contribution is critical?*). Environmental Response Diagnosis looks outward, assessing the system’s ability to cope with external challenges (e.g., *Under what kind of interference will the system fail?*). Just as medical diagnosis requires examining both organ function (internal cause) and a patient’s resistance to pathogens (external cause), the combination of both provides a holistic health assessment for an MAS.

Limitations and future directions

Current system diagnosis methods exhibit a **fragmentation** between internal (individual performance) and external (environmental response) analysis, making it difficult to locate the root causes of complex failures. Furthermore, the diagnostic process is often **decoupled** from the system optimization pipeline, and excessive computational complexity hinders practical application. To construct an autonomous diagnostic-optimization system, future research should focus on:

- **Unified Causal Diagnostic Theory:** Integrate individual contribution quantification (Mahjoub et al. 2023) and environmental boundary condition identification (Barnett et al. 2025) within a single causal framework to explore a linked diagnostic model.
- **Diagnostic-Driven Autonomous Loop:** Develop algorithms that can translate diagnostic insights (e.g., reward causal mechanisms (Wang et al. 2025a), cooperation boundaries (Barnett et al. 2025)) into real-time policy adjustment directives or safety constraints.
- **Lightweighting of High-Fidelity Methods:** Develop scalable approximation algorithms for distributed counterfactual reasoning and online causal learning to overcome the core computational bottlenecks of methods based on causal discovery (Wang et al. 2025a) and formal verification (Bogges et al. 2023).
- **Multi-Modal Evaluation Standards:** Establish comprehensive evaluation metrics that integrate diagnostic accuracy, causal fidelity, formal proof (Bogges et al. 2023), efficacy in improving expert decision-making.

Through continued exploration in the directions outlined above, systemic diagnostic technology has the potential to transform from a peripheral “analysis tool” for MARL systems

into an embedded “cognitive core” and “immune system”, ultimately contributing to methodological foundations for safe, reliable, and efficient MAS deployment in practice.

5 An evaluation framework and core challenges

Following a systematic review of techniques for explaining and interpreting MASs across all levels, the question of how to evaluate these methods in a scientifically rigorous and comprehensive manner has emerged as a central concern.

5.1 Towards a multi-dimensional evaluation framework for XMARL

The rapid development of techniques for enhancing transparency in MARL has not been matched by the establishment of standardised evaluation protocols. We therefore propose a multi-dimensional evaluation framework that moves beyond conventional performance metrics (e.g., cumulative reward). The framework, synthesised from existing literature and illustrated in Fig. 8, is organised along two orthogonal axes: **Core Transparency Quality Dimensions** and **Method Feasibility Constraints**. The former directly assesses the quality and utility of the rationales or transparent mechanisms produced, whereas the latter addresses practical deployability and engineering trade-offs. Together, these axes constitute a comprehensive and actionable evaluation system.

To make this framework immediately applicable, we explicitly link each evaluation dimension to the method categories and behavioral hierarchies introduced in Sect. 4. Table 8 provides a compact summary, specifying for each dimension: (i) the primary method classes

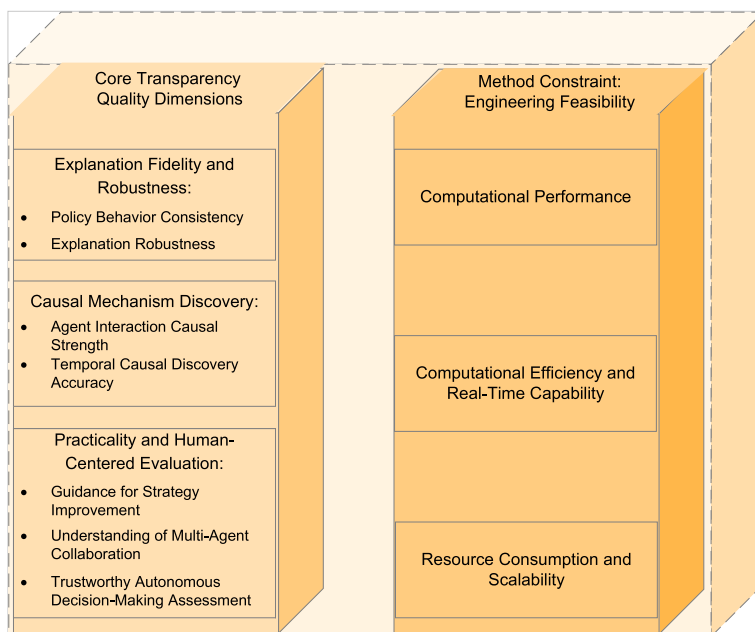


Fig. 8 A multi-dimensional evaluation framework for explainable multi-agent reinforcement learning

(organized by the four-level hierarchy of Sect. 4) to which it applies, (ii) the required data and core assumptions for measurement or approximation, and (iii) suitable baselines or reference points for comparative evaluation. Detailed definitions of each metric, along with practical proxies for scenarios where ideal assumptions are not met, are provided in the subsequent subsections.

Table 8 Mapping evaluation dimensions to XMARL method categories, required data/assumptions, and suitable baselines

Evaluation dimension	Primary method paradigms (cf. Section 4)	Required data/core assumptions	Suitable baseline
Fidelity & robustness	<i>Micro</i> : Policy abstraction & feature attribution	Access to original policy for sampling; perturbation model for input	Original opaque policy ; behavioral cloning without explanation; gradient-/perturbation-based saliency (e.g., LIME) etc
	<i>Meso</i> : Attention/relational modeling & communication analysis	Interaction logs; ability to mask/randomize communication	No-communication / unconstrained communication (for overhead comparison); shuffled attention weights; random interaction graph etc
	<i>Macro</i> : Behavior abstraction & learned world models	Trajectory dataset; proxy model for strategy comparison	Randomly assigned strategy labels; world model trained only on individual trajectories etc
	<i>System</i> : Performance diagnosis & environmental response analysis	Multiple evaluation runs; environmental perturbation suite	Heuristic performance monitoring (e.g., reward threshold alert); non-explainable anomaly detector etc
Causal mechanism discovery	<i>Micro</i> : Counterfactual reasoning	Intervention-capable simulator or identifiable causal assumptions	Random counterfactual action selection etc.
	<i>Meso</i> : Causal discovery & causal-model-informed value decomposition	Ground-truth causal edges (synthetic) or conditional independence tests	Correlation-based graph (e.g., mutual information, attention); random DAG etc
	<i>Macro</i> : Causal emergence & team-level counterfactuals	Macroscopic intervention model; aggregate causal metrics	Team-level correlation network; no-emergence null model (random team partition) etc
	<i>System</i> : Causal performance attribution	Causal graph over system-level variables	Ablation of single system component; Shapley value over agents (model agnostic) etc
Human-centered utility	<i>Micro</i> : Intrinsic motivation & goal analysis	Human-provided goal semantics; user study with task questionnaires	No explanation (control condition); randomly generated motivation labels etc
	<i>Meso</i> : Role/pattern discovery & LLM-based communication	Human evaluation of discovered roles; comprehension tests	Random role assignment; ablated communication (no content); rule-based fixed roles etc
	<i>Macro</i> : Strategy recognition & LLM-generated descriptions	Human assessment of strategy labels; trust surveys in human-AI teams	No strategy label provided; template-based textual description etc
	<i>System</i> : Operator-oriented diagnostics	Operator-in-the-loop study; task performance with/without diagnostic feedback	No diagnostic feedback (silent system); conventional performance dashboard etc
Engineering feasibility	<i>All levels</i> – applies universally	Profiling in controlled computational environment (wall-clock time, memory, communication)	Non-explainable MARL algorithm with equivalent policy architecture; vanilla MARL without transparency module etc

Baseline selection rationale.

The baselines listed in Table 8 are representative examples that provide informative reference points commonly adopted in XMARL and MARL evaluations. They are intended as illustrative, not exhaustive. For *Fidelity & Robustness*, comparisons are made against the original opaque policy (to quantify fidelity loss), along with simple attribution baselines (e.g., gradient-based saliency, behavioral cloning without explanation) to verify that the extracted model does not merely memorize. For *Causal Mechanism Discovery*, correlation-based graphs (built from mutual information or attention weights) serve as a standard non-causal baseline; random DAGs and random counterfactual actions further test whether the method captures genuine causal structure beyond chance. For *Human-Centered Utility*, the classic “no explanation” condition is the strongest baseline, supplemented by random/ablated explanations to control for placebo effects. For *Engineering Feasibility*, the non-explainable MARL counterpart with identical policy capacity is the natural reference, quantifying overhead without confounding by algorithmic differences. These baselines align with established practices in the reviewed literature (e.g., comparing against LIME in feature attribution, using shuffled attention in multi-agent communication, and employing no-explanation controls in user studies). Researchers may adopt additional baselines tailored to specific experimental settings, ensuring that reported metrics capture genuine transparency benefits.

5.1.1 Core Dimension 1: Explanation fidelity and robustness

Fidelity measures the extent to which post-hoc explanations accurately reflect the internal decision-making process of the underlying black-box MARL policy. Robustness, in turn, captures the stability of those explanations under environmental perturbations or observational noise. High-fidelity, robust explanations are essential for fostering user trust and enabling effective system debugging.

Policy behavior consistency

This metric quantifies the alignment between an **explainable proxy policy** and the original black-box policy when interacting with the same environment. In multi-agent settings, consistency can be assessed through the similarity of joint action distributions:

$$\text{Fidelity}_{\text{MA}} = \mathbb{E}_{\tau \sim \pi_{\text{joint}}} [\text{Sim}(\hat{\pi}_{\text{joint}}(\tau), \pi_{\text{joint}}(\tau))]$$

where τ is the joint trajectory, π_{joint} is the original (often opaque) joint policy, $\hat{\pi}_{\text{joint}}$ is the explainable proxy policy, and Sim is a trajectory similarity measurement function (e.g., joint action matching rate). Higher values indicate greater fidelity (Ribeiro et al. 2018).

Assumptions: The method requires sampling access to the original policy and an environment simulator to generate trajectories. *Practical proxy:* When full trajectory simulation is costly, one can measure the behavioural cloning accuracy of the proxy model on a fixed dataset of state-action pairs collected from the original policy.

Explanation robustness (stability)

Robustness can be quantified by comparing explanation feature sets under original and perturbed multi-agent states:

$$\text{Env-Robustness} = \frac{1}{N} \sum_{i=1}^N \frac{|E(s_i) \cap E(\tilde{s}_i)|}{|E(s_i) \cup E(\tilde{s}_i)|}$$

where s_i and $\tilde{s}_i$ are the original and perturbed multi-agent joint states, respectively, and $E(\cdot)$ is the generated set of explanation features. This metric measures the sensitivity of explanations to perturbations; a higher value indicates greater stability (Sundararajan et al. 2017). It quantifies the stability of explanations under environmental perturbations by comparing the similarity of explanation feature sets between original and perturbed states.

Assumptions: A perturbation model (e.g., Gaussian noise on observations) is defined. *Practical proxy:* When generating perturbed states is difficult, one can compute explanation consistency across multiple episodes with different random seeds, treating stochasticity as a natural perturbation.

5.1.2 Core Dimension 2: Causal mechanism discovery

This dimension evaluates a method's capacity to move beyond correlational patterns and uncover the causal mechanisms that underpin emergent agent behaviours. It focuses specifically on causal relationships such as inter-agent interactions and temporal dependencies.

Agent interaction causal strength

The causal effect of agent i 's action on agent j 's reward can be expressed as an Average Treatment Effect:

$$\text{Interaction-ATE}_{i \rightarrow j} = \mathbb{E}_{\tau}[\text{Reward}_j \mid \text{do}(a_i = a')] - \mathbb{E}_{\tau}[\text{Reward}_j \mid \text{do}(a_i = a'')]$$

where $\text{do}(\cdot)$ denotes an intervention, Reward_j is the reward obtained by agent j , and τ is the joint trajectory. The overall capacity of an explanatory method to capture collaborative causal effects can be summarised by the Mean Absolute Average Treatment Effect across all agent pairs:

$$\text{MA-AATE} = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j \neq i} |\text{Interaction-ATE}_{i \rightarrow j}|.$$

Assumptions: The environment must support explicit interventions ("do" operations) or provide a causal simulator. No unobserved confounders should affect both the cause and effect variables. *Practical proxy:* When interventions are infeasible, one can compute conditional mutual information $I(\text{Reward}_j; a_i \mid \text{context})$ as a correlational lower bound, or evaluate the stability of inferred causal edges under bootstrapped data resampling.

Temporal causal discovery accuracy

RL decisions exhibit pronounced temporal structure. The accuracy of an explanatory method in recovering the temporal causal graph over states, actions, and rewards is given by:

$$\text{Temporal-Causal-Accuracy} = \frac{\text{TP}_{\text{edge}} + \text{TN}_{\text{edge}}}{\text{TP}_{\text{edge}} + \text{FP}_{\text{edge}} + \text{TN}_{\text{edge}} + \text{FN}_{\text{edge}}}$$

where TP_{edge} , FP_{edge} , TN_{edge} , and FN_{edge} denote the numbers of true positives, false positives, true negatives, and false negatives in identifying temporal causal edges, respectively. This metric is particularly well suited for evaluating explanations in partially observable domains.

Assumptions: A ground-truth temporal causal graph is available (typically only in synthetic or simulated environments). *Practical proxy:* In the absence of ground truth, one can measure the predictive improvement of a downstream dynamics model when using the discovered causal graph as a structural prior (e.g., via Granger causality tests or forecast error reduction).

5.1.3 Core Dimension 3: Practicality and human-centered evaluation

The ultimate value of an explanation lies in its practical utility for human (e.g., operators, designers, regulators). This dimension evaluates the effectiveness of explanations in enhancing human comprehension, trust, and decision-making capabilities.

Guidance for strategy improvement

This metric quantifies the extent to which explanations enable users to improve multi-agent policy performance. It is defined as the relative performance gain achieved after human-guided policy adjustments:

$$\text{Guidance-Effectiveness} = \frac{J(\pi_{\text{human-adjusted}}) - J(\pi_{\text{original}})}{J(\pi_{\text{optimal}}) - J(\pi_{\text{original}})}$$

$J(\pi)$ denotes the performance metric of policy π (e.g., the expected cumulative reward), $\pi_{\text{human-adjusted}}$ is the policy adjusted by a human user after receiving the explanations, π_{original} is the original policy, and π_{optimal} represents a suitable optimal or reference policy. Values close to unity indicate strong guidance, directly linking explanation quality to the core objective of reinforcement learning—performance optimisation.

Assumptions: Human participants with domain knowledge are available; a reference policy exists. *Practical proxy:* Automate the adjustment process using a script that mimics human reasoning based on the explanation content (e.g., removing agents deemed “lazy” by the explanation) and measure the resulting performance change.

Understanding of multi-agent collaboration

This metric evaluates whether the provided insights help users understand the collaborative mechanisms among agents. It can be tested by the user’s ability to predict agent behaviors or team outcomes:

$$\text{Collab-Understanding} = \frac{1}{K} \sum_{k=1}^K \mathbb{I}(\text{Prediction}_{\text{human},k} = \text{Outcome}_{\text{actual},k})$$

where $\text{Prediction}_{\text{human},k}$ represents the user’s prediction in the k -th test scenario (e.g., which agent will take a key action, whether the team will succeed), $\text{Outcome}_{\text{actual},k}$ denotes the actual outcome, and K is the number of test scenarios. This metric quantifies the extent to which the provided transparency promotes understanding of multi-agent collaboration mechanisms through user prediction accuracy (Tjoa and Guan 2020).

Assumptions: Human participants must predict agent behaviours or team outcomes. *Practical proxy:* Train a separate neural network to predict the same outcomes using only the explanation outputs as input; its prediction accuracy serves as a proxy for the information content of the explanation.

Trustworthy autonomous decision-making assessment

In MARL deployment, a key goal of providing model transparency is to establish human trust in autonomous decision systems. This can be evaluated by measuring whether users are willing to relinquish control to the AI system at critical decision points:

$$\text{Trust-Score} = \frac{\text{Number of acceptances}}{\text{Number of critical decision points}}$$

where critical decision points refer to high-risk situations requiring human supervision or intervention. This metric directly reflects the acceptability of the RL in real-world deployment.

Assumptions: Critical decision points are predefined and a human-in-the-loop study is conducted. *Practical proxy:* Use the frequency with which the AI's suggested action is overridden by a rule-based safety monitor as an inverse proxy for trust (lower override rate implies higher trust in the transparent system).

5.1.4 Method constraint: engineering feasibility

This dimension assesses the practical overhead introduced by **transparency-enhancing mechanisms**—whether via interpretability-by-design or post-hoc explanation. The aim is to achieve high interpretability without incurring disproportionate degradation in system performance or resource efficiency.

Computational performance

This metric evaluates the change in the system's core performance (e.g., team cumulative reward, task success rate) after introducing the transparency mechanism:

$$\text{Performance-Ratio} = \frac{J(\pi_{\text{enhanced}})}{J(\pi_{\text{baseline}})}$$

where $J(\pi)$ denotes the performance metric of policy π , which can be defined in multi-agent environments as team cumulative reward ($\sum_{t=0}^T \gamma^t R_t^{\text{team}}$), task success rate, or policy convergence speed; π_{enhanced} represents the multi-agent policy integrated with the transparency-enhancing mechanism; and π_{baseline} refers to the original opaque multi-agent policy. A **Performance-Ratio** value close to or greater than 1 indicates that the transparency mechanism does not compromise system performance and may even enhance collaborative efficiency (Groce et al. 2014).

Assumptions: Both policies are evaluated under identical conditions. *Practical proxy:* For offline evaluation, use importance sampling to estimate the performance of π_{enhanced} using data collected from π_{baseline} , reducing the need for additional environment interactions.

Computational efficiency and real-time capability

The additional time required to generate transparency (e.g., computing explanations) is a critical factor for real-time applications:

$$\text{Time-Overhead} = \frac{T_{\text{enhanced}} - T_{\text{baseline}}}{T_{\text{baseline}}} \times 100\%$$

where T_{enhanced} denotes the time required for the system with transparency features to generate a decision (including any rationale), and T_{baseline} represents the time required for the original opaque system to generate a decision. This overhead should be as small as possible and should not grow exponentially with the number of agents N (Kuhnle et al. 2022).

Resource consumption and scalability

This assesses the feasibility of transparency mechanisms in resource-constrained environments (e.g., mobile robots), including memory, energy consumption, and communication overhead. For example, communication overhead can be expressed as:

$$\text{Communication-Overhead} = \frac{B_{\text{transparency}}}{B_{\text{control}}} \times 100\%$$

where $B_{\text{transparency}}$ denotes the bandwidth required for transmitting transparency-related information (e.g., explanations), and B_{control} represents the bandwidth required for transmitting basic control information. High bandwidth overhead may limit the application of such methods in communication-constrained environments (Guo et al. 2023).

Assumptions: Communication can be monitored and separated into control vs. transparency payloads. *Practical proxy:* Measure the scaling behaviour of total communication as a function of agent count N ; a transparency mechanism that scales linearly or sub-linearly with N is preferable to one with quadratic growth.

The core philosophy of engineering feasibility evaluation is to ensure that the transparency mechanism serves as an enhancement to the system, not a performance bottleneck. Evaluation should at least include: performance retention testing, efficiency benchmarking, and scalability verification under increasing system size. These metrics collectively help researchers make trade-offs between transparency quality and system performance, promoting the practical deployment and application of XMARL technology.

5.2 Core challenges and open problems

Despite notable progress, XMARL remains in its formative stages and confronts fundamental challenges that constrain both theoretical advancement and practical deployment. Figure 9 provides a systematic summary of these challenges, which derive from the limitations discussed across individual, interaction, team, and system levels in Chapter 4. They represent key obstacles on the path to truly transparent and trustworthy multi-agent systems and, simultaneously, define the problem space that guides the forward-looking research roadmap presented in Sect. 6.

5.2.1 The transparency trilemma

A persistent and foundational challenge in XMARL is the **Transparency Trilemma**, which captures the inherent tension among model expressiveness, explanation quality/interpretability, and computational efficiency. Inherently interpretable models (e.g., linear models, shallow trees) offer transparency but may have limited expressive capacity in complex tasks.

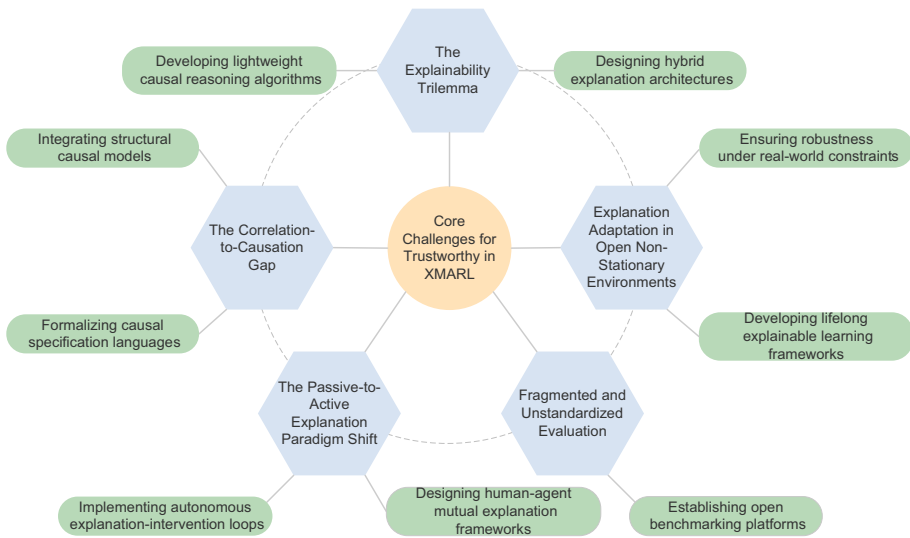


Fig. 9 Core challenges in explainable multi-agent reinforcement learning

Deep models provide expressiveness at the cost of opacity, necessitating post-hoc explanations that are often computationally demanding. This trilemma raises pressing open questions: How can hybrid neuro-symbolic architectures reconcile expressiveness with inherent interpretability? And how might lightweight causal reasoning be achieved for explanation in large-scale online settings?

5.2.2 The gap from correlation to causality

Mainstream XMARL explanation methods (e.g., attention and feature attribution) predominantly rely on revealing statistical correlations rather than causal mechanisms. This limitation becomes particularly pronounced in the context of complex multi-agent interactions and emergent behaviors, potentially limiting the reliability and actionability of explanations for safety-critical decisions. Therefore, key questions arise: How can we deeply embed causal inference frameworks, such as Structural Causal Models, into the core design of XMARL algorithms and their explanatory methods, rather than using them merely for post-hoc analysis? How can we develop computationally feasible causal formalization methods for multi-agent interactions?

5.2.3 Fragmentation and non-standardization of evaluation

The absence of established benchmarks, standardised evaluation protocols, and comprehensive quantitative metrics severely impedes the comparability of research findings and the systematic accumulation of knowledge. The field has yet to reach consensus on how to evaluate transparency quality across dimensions such as fidelity, causality, usability, and efficiency. A central open problem is the construction of an open, unified benchmarking platform that encompasses diverse task environments, multi-level evaluation metrics, and robust human-factor assessment methods that account for multiple stakeholder perspectives.

5.2.4 Paradigm shift from passive to active transparency

The current dominant paradigm in XMARL is “**post-hoc explanation**”, which separates the explanation process from learning itself, limiting its potential to proactively enhance system performance. A promising direction is to investigate an “**active transparency**” paradigm, integrating the generation and use of insights throughout the entire lifecycle of learning and deployment. This raises critical questions: How can we translate explanatory signals (e.g., counterfactual contribution) into intrinsic rewards or safety constraints to create a self-evolving loop of “explain-diagnose-intervene-optimize”? How can we design real-time interactive explanation frameworks that support bidirectional human-AI understanding and collaboration?

5.2.5 Adaptability in open non-stationary environments

Existing transparency-enhancing methods are predominantly validated in simplified, static, fully cooperative settings. Evidence regarding their performance remains limited in the face of open-world, mixed-motive scenarios with dynamically changing agent populations and non-stationary dynamics. This inadequacy represents a key barrier to real-world deployment. Pressing open problems include: How can explanation frameworks be designed to capture and convey long-term evolution in agent motivations and roles? And how can the robustness and effectiveness of transparency mechanisms be ensured under realistic constraints such as partial observability and limited communication?

This systematic analysis of core challenges and open problems serves both as a critical assessment of current limitations and as a clear articulation of the trajectory for future XMARL research. By identifying fundamental barriers and formulating specific research questions, this analysis provides a structured basis for discussing future development of explainable MASs and establishes the logical foundations for the conclusions and outlook presented in Sect. 6.

6 Conclusion and future research directions

Building upon a systematic review and critical analysis of the XMARL landscape, this section synthesizes the contributions of this work, distills the field’s core challenges, and proposes strategic directions for future research.

6.1 Research summary and contributions

This study provides a comprehensive systematization of the XMARL field, yielding the following primary contributions:

- First, we introduce a structured, multi-level classification framework. This framework not only effectively catalogs existing explainability techniques but, more importantly, organizes prominent explanatory perspectives from micro-level decisions to macro-level emergent behaviors. By decomposing the complex problem of multi-agent explainability into four interconnected yet distinct levels, it provides researchers with a clear

analytical pathway and structured research dimensions.

- **At the Micro-level of Individual Behavior**, we systematically analyzed core techniques—including policy abstraction and model extraction, attention and feature attribution, counterfactual reasoning and influence analysis, and intrinsic motivation and goal analysis. These approaches offer diverse explanatory perspectives for deciphering the internal decision-making logic of individual agents.
- **At the agent interaction and collaboration mechanisms level**, we explored methods for modeling and explaining agent interactions. These encompass relational modeling via attention mechanisms, communication protocol interpretation, causal interaction verification, collaborative structure discovery based on roles and patterns, and collaboration logic explanation through team value decomposition. These techniques are vital for understanding the underpinnings of multi-agent cooperation.
- **Regarding Team Strategy and Emergent Behavior**, we surveyed approaches for interpreting collective intelligence. These include team behavior abstraction and high-level goal identification, macro-state modeling and phase analysis, frameworks based on causal emergence theory, counterfactual team-level analysis, and team behavior generation using learned world models. These methods aim to uncover the principles governing system-level strategic dynamics and emergence.
- **Concerning System Performance and Environmental Response Diagnosis**, we examined techniques for performance diagnosis and environmental impact analysis. These methodologies provide critical tools for system-level diagnosis, robustness assessment, and trustworthy deployment.
- Furthermore, this study proposes a multi-dimensional evaluation system that moves beyond conventional performance-centric metrics. It establishes a structured set of evaluation dimensions assessing explanatory fidelity, causality, usability, and computational efficiency. This system offers a scientific basis for comparative analysis and guides the development of high-quality, practical explainability techniques.
- Finally, through a critical synthesis of the literature, we identify the fundamental tensions constraining the field's development, thereby providing a problem-oriented roadmap for future research breakthroughs. Collectively, these contributions provide a structured knowledge base that can help consolidate the theoretical underpinnings of XMARL and guide future inquiry.

6.2 Distillation of core problems

Our analysis condenses the central impediments in XMARL into three recurring challenges:

- **The Transparency Trilemma:** The inherent trade-off between model expressivity, explanatory transparency, and computational efficiency represents a recurring design trade-off. Achieving an optimal balance remains elusive and demands foundational innovation.
- **The Correlation-Causation Gap:** A pervasive methodological shortcoming is the reliance on correlational evidence rather than causal mechanisms. Bridging this gap is particularly challenging within the complex web of multi-agent interactions.
- **The Standardization Deficit:** The absence of unified benchmarks and evaluation protocols fragments the research landscape, hindering reproducible, comparable, and cu-

mulative progress.

Additional challenges include the difficulty of transitioning from passive explanation to proactive guidance, limited adaptability in open and dynamic environments, and questions regarding the practical deployability of existing methods.

6.3 Future research directions

The preceding analysis in this survey reveals that the development of XMARL is constrained by a series of interconnected challenges spanning from theoretical foundations to practical applications. This section synthesizes the technical limitations analyzed at various levels in Sect. 4 with the fundamental challenges summarized in Sect. 5.2 into a cohesive research roadmap (as shown in Fig. 9). This roadmap aims to systematically transform explainability from an external, add-on property into a core component of intelligent and trustworthy MASSs.

6.3.1 Tackling the “trilemma” and the “causal gap”—laying the theoretical and modeling foundation

The primary objective of this thrust is to overcome the fundamental tension among expressiveness, transparency, and efficiency (the “Explainability Trilemma”, Sect. 5.2.1) and to bridge the “correlation-to-causation gap” (Sect. 5.2.2). This motivates further development of intrinsically interpretable and causally informed models. Specifically, research should focus on:

- Developing hybrid architectures that integrate the automation and generalization strengths of deep learning with explicit, human-readable symbolic representations effective for interaction analysis (Sect. 4.3).
- Deeply embedding structural causal models and interventional discovery frameworks into the MARL learning loop, moving beyond post-hoc attribution. This involves creating lightweight, trainable causal modules (a need emphasized across Sects. 4.3 and 4.4) to provide real-time causal explanations for individual decisions and emergent team dynamics.

Progress in this thrust has the potential to provide the theoretical and modeling foundation needed for more trustworthy XMARL systems.

6.3.2 Advancing evaluation standardization and an active paradigm shift—building a closed-loop technological ecosystem

The second thrust addresses the critical need for robust evaluation (“Fragmented and Non-Standardized Evaluation,” Sect. 5.2.3) and catalyzes the “paradigm shift from passive to active explanation” (Sect. 5.2.4). The goal is to construct a closed-loop ecosystem where explanations both measure and enhance system performance. Key initiatives include:

- Constructing open, unified benchmark platforms that aggregate multifaceted evaluation

needs across all technical layers (Sects. 4.2–4.5). These platforms must offer diverse environments and a comprehensive suite of metrics covering fidelity, causality, human-AI usability, and computational cost.

- Designing algorithms capable of translating explanatory insights—such as counterfactual contributions (Sect. 4.3), identified policy vulnerabilities (Sect. 4.4), or root-cause diagnoses (Sect. 4.5)—into actionable feedback. This feedback should manifest as intrinsic rewards, curriculum learning guides, or safety constraints, thereby creating self-improving MARL systems.

This thrust aims to evolve explanation from a purely diagnostic tool toward a core driver of autonomous optimization, thereby closing the loop between understanding and improvement.

6.3.3 Ensuring robustness in open worlds—expanding towards a trustworthy application ecosystem

The final thrust directly tackles the challenge of “explanatory adaptability in open non-stationary environments” (Sect. 5.2.5), ensuring XMARL methods remain reliable and effective under realistic conditions. This involves scaling and rigorously testing explainability frameworks in complex, dynamic settings. Research must prioritize:

- Integrating techniques for open-world adaptation (Sect. 4.3), explaining non-stationary policies (Sect. 4.4), and robust diagnostic systems (Sect. 4.5) to create explanatory mechanisms that dynamically adapt to changing agent populations, objectives, and human-in-the-loop interactions.
- Focusing on safety-critical application domains (e.g., autonomous driving, embodied AI) to stress-test these adaptive frameworks. Leveraging domain-specific requirements will drive the development of resilient explanation models capable of operating under partial observability and communication constraints.

This thrust bridges the gap between laboratory validation and field deployment, paving the way for real-world impact.

These three thrusts are deeply interconnected and mutually reinforcing. Progress in the model foundation (Sect. 6.3.1) furnishes new components for the evaluation and optimization ecosystem (Sect. 6.3.2). The rigorous benchmarks and optimization loops of the second thrust, in turn, validate and refine the theories from the first. Together, they may support more robust evaluation and development for real-world application (Sect. 6.3.3), while challenges and data from applications provide continuous feedback to guide foundational research. This integrated roadmap highlights promising pathways for systematically advancing XMARL. Realizing the vision of transparent, trustworthy, and collaborative multi-agent intelligence that can be safely integrated into society will require sustained efforts across these interconnected directions, along with continuous empirical validation and community-wide collaboration.

6.4 Limitations of this survey

While this survey provides a systematic and critical analysis of current XMARL research, several limitations should be acknowledged to properly contextualize its contributions and conclusions.

- **Temporal and Thematic Scope.** The core systematic search is limited to January 2022–October 2025 and focuses exclusively on cooperative settings. Foundational pre-2022 works and studies in competitive or mixed-motive scenarios, though partially incorporated as background, were not subjected to the same systematic screening protocol. This may result in an incomplete portrayal of the broader XMARL evolutionary trajectory and the transferability of techniques across different MARL paradigms.
- **Search and Selection Constraints.** Despite a comprehensive search strategy across five databases and the use of backward/forward snowballing, relevant studies—particularly those in non-English languages, grey literature, or very recent preprints not yet indexed—may have been inadvertently omitted. Moreover, while explicit inclusion/exclusion criteria were employed, the qualitative judgment required in title/abstract screening and full-text assessment introduces a degree of subjectivity that cannot be entirely eliminated.
- **Categorization Ambiguity.** The proposed four-level taxonomy, although grounded in the structure of multi-agent decision-making, imposes a discrete classification on a continuous and often overlapping explanatory landscape. Despite the rigorous protocol for handling cross-level contributions (Sect. 2.5), borderline cases remain, and alternative organizational schemes could highlight different dimensions of the literature.
- **Evaluation of Preprint Quality.** Our criteria for including arXiv preprints (at least three of five indicators) represent a structured attempt to filter for quality in the absence of peer review. However, these indicators (e.g., page length, institutional affiliation) are proxy measures and may not perfectly correlate with methodological soundness or long-term impact.
- **Evolving Landscape.** XMARL is a rapidly advancing field. The analysis presented here reflects the state of the literature up to October 2025; newer publications emerging after this cutoff may address some of the challenges identified herein or introduce novel paradigms not captured in our synthesis.

These limitations do not invalidate the synthesis within its defined scope, but they delimit the scope of the conclusions drawn. Future iterations of this survey, or complementary reviews with different scopes and taxonomies, will be essential to maintain an up-to-date and broader map of the XMARL research frontier.

Appendix A: All tables

See Tables 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21.

Table 9 Comparative analysis of policy abstraction and model extraction methods

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Transparent model direct design	Employs tabular learning with discrete semantic states to directly obtain interpretable strategy matrices	Tabular RL (Expected SARSA), discrete state-action mapping	Luvistutto et al. (2025)	Visualization of strategy matrices from Q-values; animation of behavioral trajectories	Inherent model transparency without distillation; indirect analysis of collaborative behaviors feasible	Discretization dependency leads to poor scalability; strong reliance on manual parameter tuning; limited by assumptions	Small-scale, simple-state systems requiring high interpretability for demonstrations or specific industrial applications
	Uses interpretable symbolic programs as optimization targets, employing LLM-driven evolutionary search to discover high-performance transparent models directly from data	LLM-driven evolutionary search, program synthesis	Castro et al. (2025a, 2025b)	Generated Python source code with semantic variables and logical flow; computable complexity metrics	Fully automated, significantly reducing manual prior design; discovered models can surpass manually designed baselines; programs serve as testable scientific hypotheses with inherent interpretability	Extremely high computational cost; generated code readability and conciseness need improvement; search process sensitive to initial prompts and data quality	Sequential decision-making behavior analysis, cognitive modeling, scientific research requiring automated discovery of interpretable strategies from data
Rule distillation of black-box strategies	Distills neural network policies into symbolic models (e.g., decision trees) through imitation learning	Imitation learning, team Q-value resampling, LLM prompt engineering	Milani et al. (2022b); Xi-Jia et al. (2025); Chen et al. (2025c); Zhang et al. (2024b)	Structured decision trees; natural language descriptions; multi-level visualizations	Intuitive explanations with clear logic chains; enhances readability (Xi-Jia et al. 2025); systematic insights (Zhang et al. 2024b)	Poor adaptation to high-dimensional states; dependence on expert policy quality; distillation may introduce errors; lacks temporal explanations (Xi-Jia et al. 2025)	Tasks with high-quality expert policies requiring post-hoc auditing; scenarios needing explanations for non-experts; specific domain optimizations like traffic signal control

Table 9 (continued)

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Hierarchical & probabilistic strategy representation	Constructs behavior trees or strategy graphs to express temporal logic or perform probabilistic reasoning	Temporal-aware clustering, behavior tree formal modeling, probabilistic graphical models, teleological reasoning	Domènech I Vila et al. (2022); Gimenez-Abalos et al. (2025); Wan et al. (2025a)	Modular behavior trees; state-action transition graphs with probabilities; intent values with natural language explanations	Excels at temporal logic representation (Wan et al. 2025a); support counterfactual queries and deep motivational explanations (Domènech I Vila et al. 2022; Gimenez-Abalos et al. 2025)	requires manual clustering intervention (Wan et al. 2025a); depend on data quality and discretization (Domènech I Vila et al. 2022; Gimenez-Abalos et al. 2025); high construction complexity	Robotic task planning; analysis of individual decision processes and motivations in partially observable collaborative environments
Semantic mapping & strategy comparison	Maps policies to concept spaces or latent spaces for semantic interpretation or distance-based comparison	Neural logic reasoning, concept-based explanations, autoencoders, strategy distance metrics	Ji et al. (2023); Hu et al. (2024c); Reiestad (2025)	First-order logic rules; strategy distance matrices; concept saliency heatmaps and rules	rigorous logical rules (Ji et al. 2023); automatic concept discovery (Reiestad 2025); strong generalizability and quantitative comparison (Hu et al. 2024c)	constrained rule search space (Ji et al. 2023); potential information loss in concept estimation (Reiestad 2025); latent space as a black box with high computational cost (Hu et al. 2024c)	Symbolic observation environments; tasks requiring understanding of high-level semantic concepts; policy behavior analysis and diversity evaluation

Table 10 Comparative analysis of attention and feature attribution methods

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Condition/instruction-modulated attention	Generates saliency vectors from global conditions or high-level instructions to modulate the attention weights of a local encoder	Saliency vectors, Transformer encoder, modulated attention	Motokawa and Sugawara (2023b, 2023c)	Condition-dependent attention heatmaps; instruction-following object collection heatmaps	Explains dynamic policy adjustments; supports instruction control and causal exploration (Motokawa and Sugawara 2023c)	Requires predefined condition states; limited instruction scope; explanations often correlational	Human-computer interaction, task-following systems, applications requiring situational awareness
Multi-perspective/evolutionary analysis	Decomposes attention types or analyzes their temporal evolution to reveal attention preferences and strategy shifts	Attention masks (ablation), self-attention mechanisms, dynamic observation representation networks	Motokawa and Sugawara (2023a); Chen et al. (2024)	Multi-category attention heatmaps (e.g., location, class); evolution of attention weights over time	Enables fine-grained analysis of focus points (Motokawa and Sugawara 2023a); reveals macro-stage transitions in policy (Chen et al. 2024)	Explanations limited to perceptual level; lacks systematic quantitative tools (Chen et al. 2024)	Visual navigation, multi-object decision-making; analyzing long-term agent learning processes
Integration of prior knowledge	Incorporates human priors (e.g., fuzzy rules) into networks to guide attention allocation	Fuzzy logic systems, hypernetworks, contrastive variational bounds	Liu et al. (2025a); Yang et al. (2025b)	Prior-guided attention weight maps; attention heatmaps in semantic space	Guides policies towards human-intuitive behavior (Liu et al. 2025a); enhances semantic meaningfulness of attention (Yang et al. 2025b)	Priors may introduce bias (Liu et al. 2025a); explanations remain largely correlational (Yang et al. 2025b)	Imitation learning, educational robotics; complex visual tasks learning policies from pixel inputs
Faithful attribution via hybrid models	Employs mixture distributions as policy structure, enabling direct visualization of entity contribution weights	Mixture probability models, reparameterization trick	Schier et al. (2025)	Visualization of contribution weights from observed entities to the action distribution	Explanations are faithful to the decision process; strong generalizability; intuitive contribution display	Requires structured observation input; higher computational cost; sensitive to weight parameters	High-risk domains requiring transparent, real-time decisions
Hard attention for efficiency	Uses non-differentiable hard attention to select critical regions, training localization and behavior policies hierarchically	Non-differentiable attention, hierarchical MARL, policy gradient	Cheng et al. (2025)	Sequence of coordinates for attended image regions	Provides spatially explicit explanations; low computational cost; robust to noise	Complex and unstable training; less advantageous for simple tasks	Real-world tasks with high noise and high-dimensional visual inputs

Table 11 Comparative analysis of counterfactual reasoning and influence analysis methods

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Latent Variable Inference	Infers latent intents or states via probabilistic generative models and quantifies their influence	Variational Inference, EM Algorithm, Max-Entropy RL	Jiang et al. (2024); Taghavi and Farnoosh (2025a)	Intrinsic reward value; implicit explanation via policy	Provides a probabilistic framework for motivation; handles partial observability	Explanations are implicit and non-intuitive; computationally complex; relies on model assumptions	Discovering hidden decision factors from data; collaborative tasks under partial observability
Individual Incentive via Counterfactual Contribution Quantification	Measures the difference between factual and counterfactual global state transition distributions caused by an individual's action	CVAE, KL Divergence, Policy Disentanglement	Li et al. (2024f)	Individual intrinsic contribution score (exploration signal)	Provides direct, personalized causal signals; avoids credit assignment; filters noise	High computational cost; counterfactual comparisons are often limited to single-step	Cooperative exploration in sparse-reward environments
Automated Causal Explanation via Simulator Intervention & LLM Fusion	LLMs query simulators to generate counterfactual scenarios and synthesize causal explanations	Counterfactual Causal Theory, Predefined Query Templates, Macro-Action Abstraction, LLM Planning	Gyevnár et al. (2025)	Structured natural language description (includes causal relationships and counterfactual comparisons)	Strong causality and human readability; automated; model-agnostic	Dependent on simulator fidelity and LLM capability; multi-round interaction poses robustness risks	Post-hoc decision auditing in safety-critical domains; analysis of medium- and small-scale MASs

Table 12 Comparative analysis of intrinsic motivation and goal analysis methods

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Concept Bottleneck models	Embeds concept layers into policy networks for transparent decision pipelines	Concept Bottleneck Models, Auxiliary Loss, Counterfactual Analysis	Zabounidis et al. (2023); Grupen et al. (2022)	Concept vectors; intervention interface; performance metrics	Transparent decisions; allows runtime human intervention	Needs expert-defined concepts; may limit performance	Safety-critical domains (e.g., medical diagnosis)
Interpretable reward/personality	Designs reward functions with clear parameters or quantifiable personalities	Reward design, meta-RL, exploration theory	Miyake et al. (2024); Huang et al. (2025)	Reward parameters; personality labels	Direct motivation explanations; adaptable and robust	Requires manual design; computationally expensive	Social simulations; multi-agent coordination tasks
Semantic alignment via language	Uses structured language to align agent representations with human concepts	CLIP learning, captioning, NLP-MARL	Toquebiau et al. (2025)	Natural language communication; t-SNE visualizations	Intuitive explanations; improves generalization	Needs language systems and paired data	Human-agent interaction; cross-task transfer
State decomposition via physics	Decomposes states into physics concepts (e.g., forces) for model-based actions	Variational Autoencoders, physics models	Li et al. (2025b)	“Virtual force” vectors; probability density curves	Intuitive “white-box” decisions aligned with physics	Needs valid physical concepts; limited adaptability	Physics-based environments
Formal verification	Abstracts policies into formal models for logical verification	Temporal logic, model checking	Taghavi and Farnoosh (2025b)	Natural language explanations with failure conditions	Mathematically guaranteed explanations; precise	Computationally expensive; limited adaptability	Safety-critical domains

Table 13 Comparative analysis of interpretability methods via attention weights

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Attention mechanism design	Designs interpretable attention architectures for weight distributions	Graph Attention Networks, Hard/Soft Attention, Gating	Sun et al. (2023); Qi et al. (2022); Li et al. (2022); Chen et al. (2025b); Li et al. (2023); Shah et al. (2024)	Attention heatmaps, contribution scores	Inherent interpretability; intuitive collaboration display	High complexity; relies on simplified assumptions	Real-time interaction tasks (e.g., air combat)
Attention interpretability	Evaluates attention faithfulness via regularization	GraphMask, GNN-Explainer, Attention Regularization	Kailas et al. (2025); Itaya et al. (2023)	Attention distributions, binary masks, explanation graphs	Generalizable without performance loss	Binary masks limit diffuse influences	Post-hoc explanation for GNN-based MARL
Attention visualization	Converts attention weights to visual representations	Mean Field Theory, GAT, Clustering	Hao et al. (2023); Chen et al. (2025d); Lei et al. (2025); Du et al. (2022)	Heatmaps, evolution graphs, topology visualizations	Bridges model states with human cognition	Qualitative explanations; limited systematic evaluation	Human-in-the-loop debugging systems
Static graph construction	Constructs interaction graphs from domain knowledge	Graph Attention Networks, Predefined Structures	Han et al. (2022b)	Relation graphs, visualizations	Clear physical meaning; improves interpretability	Relies on predefined mechanisms	Domain-rich, fixed interaction scenarios
Dynamic graph learning	Dynamically constructs graphs for environmental changes	Temporal Networks, GNN, Evolutionary Algorithms	Fu et al. (2024); Peng et al. (2025a); Pesce and Montana (2023); Liu et al. (2022a); Duan et al. (2024b); Li et al. (2025a); Liu et al. (2023b)	Dynamic topology graphs, edge weight heatmaps	Adapts to dynamic environments; hierarchical interpretability	High computational cost; sensitive to hyperparameters	Dynamic interaction, large-scale agents
Complex graph structures	Models group-level collaboration via complex graphs	Hypergraphs, Factor Graphs, Hierarchical GAT	Zhu et al. (2024); Fan et al. (2025); Zhang et al. (2025b); Ren et al. (2023)	Hypergraph structures, communication graphs	Strong group-level modeling; transparent exploration	High cost; relies on prior knowledge	Group-level collaboration, heterogeneous agents
Weight matrix modeling	Embeds relation modeling into reward functions	Game Theory, Convex Optimization, Multi-objective	Gemp et al. (2022); Radke et al. (2023)	Mixing matrices, creed vectors, efficiency metrics	Quantifiable relation weights; principled equilibria	Requires strong assumptions; stationarity challenges	Mixed-motive games, social dilemmas

Table 14 Comparative analysis of interaction semantics interpretation via communication protocol analysis

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Information-theoretic optimization	Uses Information Bottleneck (IB), Graph IB, Variational Inference representations	Information Bottleneck (IB), Graph IB, Variational Inference	Yang et al. (2024); Ding et al. (2023b)	IB loss values; message visualizations (t-SNE)	Improves communication efficiency and robustness	Increases training complexity; lacks semantic transparency	Bandwidth-limited environments with noise
Dynamic weight analysis	Uses self-supervised aggregation for dynamic message weighting	Self-Supervised Aggregation, PAGNet	Guan et al. (2022); Zhang et al. (2025c)	Aggregation weights, generated messages	Enhances collaboration efficiency and scalability	Unstable in sparse-reward scenarios	Partially observable tasks with incomplete information
Intent-driven communication	Links communication actions to internal intent states	Intention-Correlation Network, Teammate Modeling	He et al. (2025b); Yuan et al. (2022)	Intent graphs, incentive messages, motivation visualizations	Clear communication motivation; targeted communication	Relies on pre-defined factors; high complexity	Complex collaborative tasks requiring coordination
Explicit communication matrices/graphs	Learns communication matrices or graph structures	Linear Complexity Modules, Bilevel Optimization	Meng and Tan (2024); Hu et al. (2024b); Shao et al. (2022)	Communication matrices, dynamic topology graphs	Transparent patterns; reduces cost; improves generalization	Increases training cost; static graphs may not adapt	Scenarios with stable communication patterns
Predefined protocol adherence	Follows human-designed protocols for coordination	Consensus Algorithms, CBPR, Distributed Collaboration	Sheng et al. (2022); Selimović et al. (2024)	Protocol logs, group structures, consensus decisions	Clear coordination logic; high efficiency and scalability	Protocols lack flexibility; assumes fully-connected communication	Scenarios with well-defined coordination patterns
LLM integration	Uses LLMs to generate/understand natural language messages	LLMs, Prompt Engineering	Li et al. (2024b, 2024c); Godfrey et al. (2024)	Dialogue logs, semantic alignment, plan texts	Reduces human-machine semantic gap; intuitive explanations	Depends on LLM data quality; high latency	Scenarios requiring semantic negotiation
Game-theoretic semantic modeling	Models interactions as incomplete-information games	Counterfactual Regret Minimization, PSRO, Extensive-Form Game	Gemp et al. (2024); Daskalakis et al. (2024)	Strategy instructions, equilibrium distributions, game trees	Formal frameworks for semantic communication; deep motivations	High complexity; depends on accurate game modeling	Scenarios requiring strategic negotiation

Table 15 Comparative analysis of interaction verification methods via causal reasoning

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Causal structure learning	Learns causal structures between variables from data	PC Algorithm, Bayesian Networks, DAG Learning	Briglia et al. (2025); Chen and Zhang (2023); Wu et al. (2022); Meli et al. (2025)	Causal network graphs, DAGs, coordination assessments	Provides causal structure explanations; framework-agnostic	Low efficiency; unstable under partial observability; high complexity	Tasks requiring causal discovery (e.g., safety-critical systems)
Causal effect quantification	Quantifies causal influences via counterfactual scenarios	Multi-Action Counterfactual Advantage, Counterfactual Message Value	Zhao and Xie (2025); Jin et al. (2025b); Chai et al. (2024); Gao et al. (2025)	Counterfactual contribution scores, value heatmaps	Rigorous causal framework; distinguishes correlation from causation	Requires global state; sensitive to hyperparameters	Safety-critical collaboration (e.g., medical, autonomous driving)
Causal model integration	Integrates structural causal models into learning architectures	Structural Causal Models, Multi-Agent Causal Graphical Models	Yao et al. (2024); Du et al. (2024); Selmonaj et al. (2025b)	Causal influence scores, intrinsic rewards, information graphs	Guides collaboration from source; provides causal explanations	High computational cost; manual state partitioning	Sparse-reward or reward-free collaborative environments
Counterfactual applications	Uses generative models or LLMs for counterfactual reasoning	Generative Models, Mask Networks, LLMs	Wenhan et al. (2024); Shi et al. (2025); Maliha and Hougen (2025); Nagpal et al. (2025)	Counterfactual reward estimates, mask diagrams, language explanations	Adapts to heterogeneous agents; captures causal mechanisms	High computational complexity; depends on model accuracy	Heterogeneous agent systems, hybrid action spaces

Table 16 Comparative analysis of collaboration structure discovery methods via roles and behavioral patterns

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Dynamic role assignment and evolution	Dynamically allocates functional roles through online learning mechanisms	K-means clustering, trajectory encoders, graph convolutional networks, periodic grouping, rule-based preference modeling, sequential update strategies	Liu et al. (2022b); Li et al. (2024d); Peng et al. (2023); Lu et al. (2025); Ma et al. (2025b)	Role assignment label evolution graphs, role feature visualizations, group structures, preference model parameters	Adaptively discovers roles reflecting team division dynamics; improves collaboration efficiency and strategy diversity; provides theoretical guarantees	Requires predefined number of roles; clustering sensitive to sample quality; grouping intervals require tuning; high computational complexity	Dynamic and open environments requiring adaptive role allocation within a team (e.g., UAV swarms, real-time strategy games)
Hierarchical intent and cognitive alignment	Explains collaboration mechanisms from intent and cognitive perspectives, learning shared intent or aligning cognitive representations	Meta-reasoning frameworks, consistent policy and intent learners, hierarchical coordination frameworks, cognitive difference and alignment representations, goal-consistency attention	Wan et al. (2025b); Fu et al. (2025); Zhang and Xu (2024); Ge et al. (2022); Chen et al. (2022)	Intent codebooks, cognitive similarity matrices, hierarchical instruction flows, multimodal collaborative behaviors, goal attention distributions	Provides high-level semantic explanations (intent, goals); enhances reasoning transparency and controllability; supports zero-shot generalization	Complex model structure with high training difficulty; intent codebook size requires presetting; prone to “echo chamber” effects; sensitive to observation noise	Complex collaborative tasks that demand high-level semantic interpretation ability and long-horizon planning (e.g., multi-stage coordination like in overcooked, multi-agent navigation)
Organizational models and structured policies	Introduces explicit organizational frameworks or structured policy constraints	MOISE+ organizational model, hierarchical latent policy model, coalition identifiers, structural entropy	Soulé et al. (2025a); Ibrahim and Fayad (2022); Su et al. (2025)	Organizational semantic graphs, latent variable semantic interpretations, coalition structure graphs, structural entropy values	Provides high-level organizational semantic explanations; enhances behavioral controllability; improves algorithm performance and interpretability	Relies on manually defined organizational norms; rigid roles may limit adaptability; high computational complexity; coalition labels require manual setting	Scenarios with well-defined organizational structures, rigid operational rules, or high safety requirements (e.g., industrial automation, traffic control systems)

Table 16 (continued)

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Collaboration pattern extraction and abstraction	Extracts reusable collaboration prototypes directly from behavioral data, such as interaction pattern disentanglement and discrete predicate policy graphs	Sparse activation functions, contrastive loss, discrete predicate logic, determinantal point processes (DPP), hierarchical cognitive graph autoencoders	Liu et al. (2024b); Domenech I Vila et al. (2024); Ruan et al. (2024); He et al. (2025a); Bezerra et al. (2025); Soulé et al. (2025b)	Interaction prototype sparse graphs, predicate-based policy decision paths, dynamic grouping graphs, hierarchical cognitive graphs	Automatically discovers cross-task shared collaboration logic, improving generalization; intuitive and verifiable explanations; high adaptability	High computational complexity; relies on discrete predicate design with limited universality; graph search and grouping increase inference overhead	Scenarios that require abstracting reusable patterns from experience for rapid adaptation to new, structurally similar tasks (e.g., multi-robot task scheduling, transfer learning for game AI)

Table 17 Comparison of collaboration logic explanation methods via team value function decomposition

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Value function decomposition architecture	Constructs trainable decomposition architectures and analyzes internal mechanisms	MIXRTs, MVD, GAM, GNN, GAT	Liu et al. (2025b); Liang et al. (2024); Liu et al. (2023c); Kortvelesy and Prorok (2022); Wei et al. (2022); Zhu et al. (2025)	Decision tree paths, multiplicative interaction weights, additive decomposition visualizations, graph attention weights, factorized value functions	Fully traceable decision process; intuitive credit assignment; inherent interpretability; naturally models agent dependencies	Limited expressive power of tree models; increased complexity with virtual agents; higher-order interactions can reduce interpretability; typically limited to discrete action spaces	Scenarios requiring end-to-end interpretable architectures and cooperative tasks with discrete or continuous action spaces
Causal modeling & reward redistribution	Combines causal heuristics or attention for return decomposition or reward redistribution	CAST, AREL	Zhang et al. (2024a); Xiao et al. (2022)	Causal structure graphs, spatio-temporal attention weights, redistributed reward values	Provides explicit causal interpretability; flexibly models spatio-temporal dependencies; no need for a pre-defined reward structure	Assumptions of temporal linear additivity and spatial reversibility required; attention computation cost scales with agent count; lacks guarantees from causal theory	Cooperative tasks with significant spatio-temporal dependencies requiring causal explanations (e.g., multi-robot sequential tasks)
Game-theoretic contribution quantification	Utilizes concepts from cooperative game theory (e.g., Shapley value, nucleolus) as the theoretical foundation for credit assignment	Shapley Value, Nucleolus, Semivalues, CORA, DR	Heuillet et al. (2022); Qiu et al. (2023); Li et al. (2025d); Han et al. (2022a); Ji et al. (2025); Castellini et al. (2025)	Shapley value contribution scores, nucleolus solutions, coalition structures, difference reward values	Provides quantitative, fair contribution explanations; grounded in solid mathematical theory; effectively mitigates over-updating issues; stable and efficient learning	Extremely high computational complexity (exponential growth); explanations are at a global average level; nucleolus may not be unique; difference reward baselines can be biased	Scenarios requiring fair credit assignment supported by cooperative game theory (e.g., resource allocation, team contribution evaluation)
Information-theoretic & mutual information methods	Leverages information-theoretic principles to provide explanations for cooperation	Graphical Mutual Information Maximization, Experience Replay, GIB, Rule-based Preference Models	Ding et al. (2023a); Yang et al. (2025a); Ding et al. (2024)	Mutual information estimates, hierarchical experience replay structure, GIB loss values, rule preference parameters	Explicitly optimizes team cooperative behavior; improves communication quality and adversarial robustness; combines rule-based and data-driven approaches, enhancing interpretability and safety	Mutual information estimation is challenging in high-dimensional spaces; thresholds require tuning; hierarchical replay increases storage and sampling complexity; rules rely on manual design	Cooperative scenarios requiring information-theoretic explanations, high adversarial robustness, or rule-based constraints (e.g., autonomous driving)

Table 17 (continued)

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Semantic & logic-enhanced methods	Introduces semantic representations or formal logic to translate abstract rewards into human-understandable concepts and rules	LLMs, LTL, Concept Bottleneck Models, Cascade Frameworks	Wang et al. (2025b); Lin et al. (2025); Jiang et al. (2025); Zhang et al. (2025a); Liu et al. (2024a); Ge et al. (2025)	Natural language feedback, dense reward functions, textual credit assignment, LTL formulae, concept bottleneck visualizations	Significantly enhances policy interpretability; supports human-in-the-loop intervention; adapts to high-dimensional states and sparse rewards; decisions are transparent and traceable	Relies on LLM parsing accuracy; human feedback is costly; LTL formula design depends on prior knowledge; concepts require pre-definition	Scenarios requiring human-readable explanations, formalized task logic, or human-in-the-loop intervention (e.g., complex planning, safety-critical decisions)

Table 18 Comparative analysis of team behavior abstraction and high-level goal identification methods

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Explicit grouping & behavior constraint	Uses grouping mechanisms and consistency losses to shape agent behavior for intra-group cooperation and inter-group specialization	Grouping strategies, Wasserstein distance, consistency loss	Yang et al. (2025c); Duan et al. (2024a)	Group structures, behavior distributed strategies (e.g., interceptors vs. chasers)	Clear role specialization; efficient task allocation; generalized and easily integrated	Often uses static grouping; performance depends on environment; may increase training complexity	Tasks requiring defined roles (e.g., pursuit-evasion, coordinated exploration)
Unsupervised representation & pattern discovery	Learns team consensus representations from historical data to automatically identify strategy patterns or team structures	VQ-VAE, clustering algorithms, trajectory encoding, hierarchical variational inference	Ruan et al. (2023); Na et al. (2025); Omidshafiei et al. (2022)	Team embeddings, strategy clusters, trajectory patterns, hierarchical behavior concepts, cluster visualizations	Adaptive pattern discovery; identifies dynamic/hierarchical structures; minimal prior knowledge needed; reward-agnostic (e.g., Omidshafiei et al. (2022))	Complex models with sensitive hyperparameters; requires preset cluster numbers; needs large offline data; interpretation is post-hoc	Multi-task scenarios with varied strategies; complex games; complex collaboration; offline behavior analysis from historical data
Symbolic planning & key-state abstraction	Extracts high-level task logic and key states to generate interpretable sub-goals or visual anchors	Hierarchical Task Networks, planning languages, key state extraction, visualization	Mu et al. (2025); Pandya et al. (2024)	Symbolic subgoals, landmark states (images/videos), natural language summaries	Human-understandable temporal logic; intuitive and user-validated	Relies on symbolic models or manual design; scalability limited for complex patterns	Hierarchical tasks; scenarios requiring intuitive explanations for humans
Hierarchical visualization & memory abstraction	Compresses team behavior into high-level goals via environment abstraction, heatmaps, or memory encoding	Graph partitioning, heatmaps, memory networks, Adaptive Resonance Theory	Tabrez et al. (2024); Zhang et al. (2023); Wai et al. (2024)	Hierarchical heatmaps, target points, deadline curves, abstract action texts	Scalable for large spaces; multi-granularity summaries for review and instruction	Heatmap quality depends on data; system complexity; abstraction loses details (post-hoc)	Large continuous spaces; sparse-reward environments; team strategy review and analysis

Table 19 Comparative analysis of macro-state modeling and phase analysis methods

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Dynamic relation modeling & gradient analysis	Quantifies instantaneous inter-agent dependencies (e.g., via gradients) to reveal dynamic evolution of leadership and dependency hierarchies	Policy gradients, dependency quantification, GNNs	Chen et al. (2025a)	Dependency strength graphs, dominant agent switching processes, hierarchical evolution visualizations	Clearly reveals micro-level driving mechanisms of leadership and dependencies across task phases	Typically limited to small-scale agents; analysis dimensions are relatively singular	Fine-grained behavior analysis in small-scale tightly-coupled systems (e.g., sequential dependency tasks)
Spatio-temporal graph modeling & evolutionary analysis	Constructs and updates explicit relation graphs among agents to model spatio-temporal evolution of strategies and analyze long-term collaboration patterns	GNNs, evolutionary algorithms, attention mechanisms	Peng et al. (2025b)	Dynamic relation graph topologies, strategy population evolution trajectories, spatio-temporal collaboration pattern visualizations	Effectively balances exploration and collaboration, provides long-term, global insights into strategy evolution	High computational and communication overhead; scalability is limited	Long-term strategy evolution analysis in complex dynamic environments (e.g., real-time strategy games)
Multi-modal behavior recognition & quantitative analysis	Transforms trajectory data into features, utilizing dimensionality reduction and classification techniques to quantitatively identify and validate complex emergent behavior modalities	PCA dimensionality reduction, machine learning classification, heatmap analysis	Sharma et al. (2022)	Behavior modality classification results, feature space distributions, quantitative behavior differentiation metrics	Provides an objective, quantitative framework for behavior differentiation; capable of identifying complex, non-intuitive behavior modalities	Environments are typically simplified; generalization to real-world complex scenarios needs further validation	Behavior pattern analysis in adversarial environments (e.g., pursuit-evasion games)

Table 20 Comparative analysis of system performance diagnosis methods

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Difference reward contribution quantification	Quantifies an agent's marginal contribution by computing the difference in team reward between its actual participation and a counterfactual scenario where it takes a no-op action	Difference reward, Shapley value approximation	Mahjoub et al. (2023)	Agent importance scores	Computationally efficient (linear complexity), strongly correlated with true contribution, algorithm-agnostic	Requires environment duplication; the no-op action may disrupt environmental dynamics	Cooperative algorithm comparison, analysis of parameter sharing effectiveness, credit assignment diagnosis in heterogeneous agents
Dynamic causal discovery	Learns causal graph models between team rewards and agent states/actions from historical data to uncover causal relationships	Dynamic Bayesian Networks, causal inference	Wang et al. (2025a)	Causal structure graphs, causal strength between variables	High explanatory depth; modular; capable of revealing nonlinear and temporal causal relationships	Computationally intensive; sensitive to hyperparameters and data structure	Offline credit assignment, causal analysis of system reward generation mechanisms
Latent capability modeling	Uses Bayesian models to infer and explain system performance as a combination of latent capability dimensions	Hierarchical Bayesian Networks, counterfactual reasoning	Tesic et al. (2025)	Posterior distributions of latent capabilities, fundamental performance parameters	Provides fine-grained capability profiles; capable of revealing evaluation biases; predictive	Heavily reliant on expert-defined meta-features of tasks; computationally expensive	In-depth evaluation of multi-agent benchmark algorithms, capability diagnosis and comparison
Reward decomposition analysis	Decomposes the global reward function into semantically explicit subcomponents and traces their respective impacts	Value function decomposition, multi-head neural networks	Iturria-Rivera et al. (2024)	Q-values corresponding to each sub-reward component, contribution ranking	Tightly integrated with the decision-making process; directly informs reward function design	Increases model complexity; decomposition scheme requires domain expertise	Analysis and debugging of tasks with composite rewards, parameter optimization in wireless network management
Entropy-driven learning diagnosis	Utilizes policy uncertainty (entropy) to dynamically adjust training mechanisms (e.g., episode length)	Information entropy, adaptive control	Yoo et al. (2025)	Learning progress indicators, dynamic adjustment policies	Lightweight, highly generalizable, significantly improves convergence efficiency	Sensitive to hyperparameters (e.g., window size); relies on meaningful environmental reward signals	Learning process optimization, training acceleration for long-horizon or sparse-reward tasks

Table 21 Comparative analysis of environmental response diagnosis methods

Technical route	Core mechanism	Key theory	References	Explanation output	Main advantages	Limitations	Scenarios
Formal temporal verification	Verify system behavior properties with temporal logic formulas	Probabilistic Model Checking, Temporal Logic	Bogges et al. (2023)	Logical judgment results, violation trajectories	High mathematical rigor, complete diagnostic conclusions	High computational complexity, limited logical expressiveness	Exact root cause diagnosis of system temporal behavior deviations
Counterfactual cooperation measurement	Calculate cooperation degree by comparing current strategies with selfish best responses	Stochastic Games, MDP Solving, Counterfactual Reasoning	Barnett et al. (2025)	Scalar cooperation score	Clear judgment criteria, intuitive explanation, context-sensitive	Difficult to solve best responses, ignores dynamic strategy adaptation	Cooperative behavior analysis, security threat detection, offline evaluation

Acknowledgements The authors sincerely thank Professor Zhao Zhuang for his invaluable assistance in locating and obtaining some difficult-to-obtain literature. Additionally, we are grateful to the anonymous reviewers for their constructive comments and suggestions, which significantly enhanced the quality of this manuscript.

Author contributions L.S.: Conceptualization, Methodology, Investigation, Writing – Original Draft. X.C.: Methodology, Validation, Visualization, Writing – Review & Editing. Z.P.: Investigation, Validation, Writing – Review & Editing. F.C.: Visualization, Writing – Review & Editing. H.C.: Funding acquisition, Project administration, Supervision, Writing – Review & Editing (as the Corresponding Author).

Funding This review received no external funding.

Data availability No datasets were generated or analysed during the current study.

Materials availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval and consent to participate Not applicable, as this is a review article and does not involve any new studies with human or animal subjects.

Consent for publication Not applicable, as this article contains no personally identifiable information.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Acero F, Li Z (2024) Distilling reinforcement learning policies for interpretable robot locomotion: gradient boosting machines and symbolic regression. In: 2024 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 6840–6847
- Acharya K, Raza W, Dourado C, Velasquez A, Song HH (2023) Neurosymbolic reinforcement learning and planning: a survey. *IEEE Trans Artif Intell* 5(5):1939–1953. <https://doi.org/10.1109/tai.2023.3311428>
- Alahmari D (2025) Agentic control towers: multi agent llm rl. In: Accelerating logistics through generative AI, digital twins, and autonomous operations. p 161
- Amour M, Brik B, Ksentini A (2024) Leveraging llms to explain drl decisions for transparent 6g network slicing. In: 2024 IEEE 10th international conference on network softwarization (NetSoft). IEEE, pp 204–212
- Amitai Y, Septon Y, Amir O (2024) Explaining reinforcement learning agents through counterfactual action outcomes. In: Proceedings of the AAAI conference on artificial intelligence. pp 10003–10011
- Annasamy RM, Sycara K (2019) Towards better interpretability in deep q-networks. In: Proceedings of the AAAI conference on artificial intelligence. pp 4561–4569. <https://doi.org/10.1609/aaai.v33i01.33014561>

- Barfuss W, Flack J, Gokhale CS, Hammond L, Hilbe C, Hughes E, Leibo JZ, Lenaerts T, Leonard N, Levin S et al (2025) Collective cooperative intelligence. *Proc Natl Acad Sci USA* 122(25):e2319948121
- Barnett SA, Wantlin K, Adams RP (2025) Measuring cooperation with counterfactual planning. In: International conference on game theory and AI for security. pp 23–41. https://doi.org/10.1007/978-3-032-08064-6_2
- Barros e Sá GC, Madeira CAG, (2025) Deep reinforcement learning in real-time strategy games: a systematic literature review. *Appl Intell* 55(4):243
- Bezerra LC, Dos Santos AM, Park S (2025) Learning policies for dynamic coalition formation in multi-robot task allocation. *IEEE Robot Autom Lett*. <https://doi.org/10.1109/lra.2025.3592080/mm1>
- Boggess K, Kraus S, Feng L (2022) Toward policy explanations for multi-agent reinforcement learning. <https://doi.org/arXiv:2204.12568> [arXiv:2204.12568](https://arxiv.org/abs/2204.12568)
- Boggess K, Kraus S, Feng L (2023) Explainable multi-agent reinforcement learning for temporal queries. <https://doi.org/arXiv:2305.10378> [arXiv:2305.10378](https://arxiv.org/abs/2305.10378)
- Briglia G, Mariani S, Zambonelli F (2025) A roadmap towards improving multi-agent reinforcement learning with causal discovery and inference. *arXiv preprint* [arXiv:2503.17803](https://arxiv.org/abs/2503.17803)
- Cao HY, Liu X, Dong SK et al (2024) Review of explainability research for reinforcement learning. *Chin J Comput* 47(8):1853–1882 ((in Chinese))
- Castellini J, Devlin S, Oliehoek FA, Savani R (2025) Difference rewards policy gradients. *Neural Comput Appl* 37(19):13163–13186. <https://doi.org/10.1007/s00521-022-07960-5>
- Castro PS, Tomasev N, Anand A, Sharma N, Mohanta R, Dev A, Perlin K, Jain S, Levin K, Dabney N, Novikov W, Turner A, Eckstein GC, Daw MK, Miller ND, Stachenfeld KJ, KL (2025a) Automated discovery of interpretable cognitive programs underlying reward-guided behavior. In: Computational and systems neuroscience meeting. <https://doi.org/COSYNE2025> [arXiv:2502.05636](https://arxiv.org/abs/2502.05636)
- Castro PS, Tomasev N, Anand A, Sharma N, Mohanta R, Dev A, Perlin K, Jain S, Levin K, Dabney N, Novikov W, Turner A, Eckstein GC, Daw MK, Miller ND, Stachenfeld KJ, KL (2025b) Discovering symbolic cognitive models from human and animal behavior. In: Proceedings of the 42nd international conference on machine learning, proceedings of machine learning research. pp 6849–6890 (PMLR)
- Chai J, Fu Y, Zhao D, Zhu Y (2024) Aligning credit for multi-agent cooperation via model-based counterfactual imagination. In: Proceedings of the 23rd international conference on autonomous agents and multiagent systems. pp 281–289
- Chen D, Zhang Q (2023) Context-aware Bayesian network actor-critic methods for cooperative multi-agent reinforcement learning. In: International conference on machine learning. pp 5327–5350 (PMLR)
- Chen X, Liu X, Zhang S, Ding B, Li K (2022) Goal consistency: an effective multi-agent cooperative method for multistage tasks. In: *IJCAI*, pp 172–178
- Chen B, Cao Z, Bai Q (2024) Sattf: A scalable attentive transfer framework for efficient multiagent reinforcement learning. *IEEE Trans Neural Netw Learn Syst* 36(4):6627–6641. <https://doi.org/10.1109/tnnls.2024.3387397>
- Chen S, Chen X, Bao Q, Zhang H, Wong CUI (2025a) Adaptive multi-agent reinforcement learning with graph neural networks for dynamic optimization in sports buildings. *Buildings* 15(14):2554. <https://doi.org/10.3390/buildings15142554>
- Chen L, Li Z, Xiong Y (2025b) Boosting multi-agent collaboration with dual attention actor-critic. In: 2025 5th International conference on machine learning and intelligent systems engineering (MLISE). IEEE, pp 222–229. <https://doi.org/10.1109/mlise66443.2025.11100230>
- Chen R, Milani S, Zhang Z, Sadeh N, Fang F (2025c) Making teams and influencing agents: efficiently coordinating decision trees for interpretable multi-agent reinforcement learning. <https://doi.org/arXiv:2505.19316> [arXiv:2505.19316](https://arxiv.org/abs/2505.19316)
- Chen G, Wang G, Van Beek A, Yan MZ, Y (2025d) Emergence of hierarchies in multi-agent self-organizing systems pursuing a joint objective. [arXiv:2508.09541](https://arxiv.org/abs/2508.09541)
- Cheng Z, Wu X, Yu J, Sun W, Guo W, Xing X (2023) Statemask: Explaining deep reinforcement learning through state mask. *Adv Neural Inf Process Syst* 36:62457–62487
- Cheng J, Li J, Shi H, Zhang T (2025) Localizing state space for visual reinforcement learning in noisy environments. *Eng Appl Artif Intell* 156:110998. <https://doi.org/10.1016/j.engappai.2025.110998>
- Crespi M, Ferigo A, Custode LL, Iacca G (2023) A population-based approach for multi-agent interpretable reinforcement learning. *Appl Soft Comput* 147:110758. <https://doi.org/10.2139/ssrn.4467882>
- Dai Y, Chen Q, Zhang J, Wang X, Chen Y, Gao T, Xu P, Chen S, Liao S, Jiang H et al (2022) Enhanced oblique decision tree enabled policy extraction for deep reinforcement learning in power system emergency control. *Electr Power Syst Res* 209:107932. <https://doi.org/10.1016/j.epsr.2022.107932>
- Daskalakis C, Golowich N, Zhang K (2023) The complexity of Markov equilibrium in stochastic games. In: The thirty sixth annual conference on learning theory, PMLR. pp 4180–4234
- Daskalakis C, Gemp I, Jiang Y, Leme RP, Papadimitriou C, Piliouras G (2024) Charting the shapes of stories with game theory. In: *NeurIPS 2024 workshop on creative AI*

- Dhebar Y, Deb K, Nagesh Rao S, Zhu L, Filev D (2022) Toward interpretable-ai policies using evolutionary nonlinear decision trees for discrete-action systems. *IEEE Trans Cybern* 54(1):50–62. <https://doi.org/10.1109/tycb.2022.318066>
- Ding S, Du W, Ding L, Zhang J, Guo L, An B (2023a) Multiagent reinforcement learning with graphical mutual information maximization. *IEEE Trans Neural Netw Learn Syst*. <https://doi.org/10.1109/tnnls.2023.3243557>
- Ding S, Du W, Ding L, Zhang J, Guo L, An B (2023b) Robust multi-agent communication with graph information bottleneck optimization. *IEEE Trans Pattern Anal Mach Intell* 46(5):3096–3107. <https://doi.org/10.1109/tpami.2023.3337534>
- Ding S, Du W, Ding L, Guo L, Zhang J (2024) Learning efficient and robust multi-agent communication via graph information bottleneck. In: *Proceedings of the AAAI conference on artificial intelligence*. pp 17346–17353
- Domènech I Vila M, Gnatyshak D, Tormos A, Alvarez-Napagao S (2022) Testing reinforcement learning explainability methods in a multi-agent cooperative environment. In: *Artificial intelligence research and development*. IOS Press, pp 355–364. <https://doi.org/10.3233/faia220358>
- Domenech I Vila M, Gnatyshak D, Tormos A, Gimenez-Abalos V, Alvarez-Napagao S (2024) Explaining the behaviour of reinforcement learning agents in a multi-agent cooperative environment using policy graphs. *Electronics* 13(3): <https://doi.org/10.20944/preprints202401.1421.v1>
- Du W, Ding S, Zhang C, Shi Z (2022) Multiagent reinforcement learning with heterogeneous graph attention network. *IEEE Trans Neural Netw Learn Syst* 34(10):6851–6860. <https://doi.org/10.1109/tnnls.2022.3215774>
- Du X, Ye Y, Zhang P, Yang Y, Chen M, Wang T (2024) Situation-dependent causal influence-based cooperative multi-agent reinforcement learning. In: *Proceedings of the AAAI conference on artificial intelligence*. pp 17362–17370. <https://doi.org/10.1609/aaai.v38i16.29684>
- Duan W, Lu J, Xuan J (2024a) Inferring latent temporal sparse coordination graph for multiagent reinforcement learning. *IEEE Trans Neural Netw Learn Syst*. <https://doi.org/10.1109/tnnls.2024.3513402>
- Duan W, Lu J, Xuan J (2024b) Group-aware coordination graph for multi-agent reinforcement learning. In: *IJCAI international joint conference on artificial intelligence*. pp 3926–3934. <https://doi.org/10.24963/ijcai.2024.434>
- Duttagupta A, Jabbari M, Fiandrino C, Fiore M, Widmer J (2025) Symbxrl: symbolic explainable deep reinforcement learning for mobile networks. In: *IEEE INFOCOM 2025-IEEE conference on computer communications*. IEEE, <https://doi.org/10.1109/infocom55648.2025.11044492>
- Fan W, Yu Z, Ma C, Li C, Yang Y, Zhang X (2025) Towards efficient collaboration via graph modeling in reinforcement learning. In: *Proceedings of the AAAI conference on artificial intelligence*. pp 16505–16513. <https://doi.org/10.1609/aaai.v39i16.33813>
- Fu Q, Qiu T, Yi J, Pu Z, Ai X (2024) Self-clustering hierarchical multi-agent reinforcement learning with extensible cooperation graph. *IEEE Trans Emerg Topics Comput Intell*. <https://doi.org/10.1109/tetci.2024.3449873>
- Fu Y, Zhu Y, Li H, Zhao Z, Chai J, Zhao D (2025) Cpig: Leveraging consistency policy with intention guidance for multi-agent exploration. *IEEE Trans Cogn Dev Syst*. <https://doi.org/10.1109/tcds.2025.3578001>
- Gajcin J, Dusparic I (2024) Redefining counterfactual explanations for reinforcement learning: overview, challenges and opportunities. *ACM Comput Surv* 56(9):1–33
- Gao Z, Qu C, Hao Y, Ke L (2025) Communication in multiagent reinforcement learning via counterfactual message value. *IEEE Trans Syst Man Cybern Syst*. <https://doi.org/10.1109/tsmc.2025.3604230>
- Ge H, Ge Z, Sun L, Wang Y (2022) Enhancing cooperation by cognition differences and consistent representation in multi-agent reinforcement learning. *Appl Intell* 52(9):9701–9716. <https://doi.org/10.1007/s10489-021-02873-7>
- Gemp I, McKee KR, Everett R, Guzmán D, Bachrach E, Balduzzi Y, Tacchetti D, A (2022) D3c: Reducing the price of anarchy in multi-agent learning. In: *Proceedings of the 21st international conference on autonomous agents and multiagent systems*. international foundation for autonomous agents and multiagent systems. Richland, SC, AAMAS, pp 498–506. <https://doi.org/10.5555/3535850.3535907>
- Gemp I, Patel R, Bachrach Y, Lancot M, Dasagi V, Marris L, Piliouras G, Liu S, Tuyls K (2024) Steering language models with game-theoretic solvers. *arXiv preprint arXiv:2402.01704*
- Ge Z, Zhu Y, Chen C (2025) Concept learning for cooperative multi-agent reinforcement learning. In: *2025 IEEE 26th China conference on system simulation technology and its applications (CCSSTA)*. IEEE, pp 625–631
- Gimenez-Abalos V, Alvarez-Napagao S, Tormos A, Cortés U, Vázquez-Salceda J (2025) Policy graphs and intention: answering “why” and “how” from a telic perspective. In: *Proceedings of the 24th international conference on autonomous agents and multiagent systems*. International Foundation for Autonomous Agents and Multiagent Systems. Richland, SC, AAMAS, pp 904–913

- Godfrey T, Hunt W, Soorati MD (2024) Marlin: Multi-agent reinforcement learning guided by language-based inter-robot negotiation. <https://doi.org/arXiv:2410.14383> 10.17760/d20467227 arXiv preprint
- Groce A, Kulesza T, Zhang C, Shamasunder S, Burnett M, Wong WK, Stumpf S, Das S, Shinsell A, Bice F, McIntosh K (2014) You are the only possible oracle: effective test selection for end users of interactive machine learning systems. *IEEE Trans Softw Eng* 40(3):307–323. <https://doi.org/10.1109/TSE.2013.59>
- Gruppen N, Jaques N, Kim B, Omidshafiei S (2022) Concept-based understanding of emergent multi-agent behavior. In: Deep reinforcement learning workshop NeurIPS
- Guan C, Chen F, Yuan L, Wang C, Yin H, Zhang Z, Yu Y (2022) Efficient multi-agent communication via self-supervised information aggregation. *Adv Neural Inf Process Syst* 35:1020–1033. <https://doi.org/10.1109/tnnls.2024.3420791>
- Guo YS, Campbell J, Stepputtis S, Li R (2023) Explainable action advising for multi-agent reinforcement learning. In: Proceedings of the 2023 IEEE international conference on robotics and automation (ICRA). pp 5515–5521. <https://doi.org/10.1109/ICRA48891.2023.10160557>
- Guo H, Wang X, Ma N, Zhang Y (2024) Robust reinforcement learning for robotic manipulation using lite vit and saliency maps. *Int J Front Eng Technol*. <https://doi.org/10.25236/ijfet.2024.060510>
- Gurses Y, Buyukdemirci K, Yildiz Y (2024) Developing driving strategies efficiently: a skill-based hierarchical reinforcement learning approach. *IEEE Control Syst Lett* 8:121–126. <https://doi.org/10.1109/lcsys.2024.3349511>
- Gyevnár B, Lucas CG, Albrecht SV, Cohen SB (2025) Integrating counterfactual simulations with language models for explaining multi-agent behaviour. <https://doi.org/arXiv:2505.17801> 10.1109/wsc63780.2024.10838923 arXiv preprint
- Han D, Lu CX, Michalak T, Wooldridge M (2022a) Multiagent model-based credit assignment for continuous control. In: Proceedings of the 21st international conference on autonomous agents and multiagent systems. pp 571–579
- Han Y, Piao H, Hou Y, Sun Y, Sun Z, Zhou D, Yang S, Peng X, Fan S (2022b) Deep relationship graph reinforcement learning for multi-aircraft air combat. In: 2022 International joint conference on neural networks (IJCNN). IEEE, <https://doi.org/10.1109/ijcnn55064.2022.9892208>
- Hao Q, Huang W, Feng T, Yuan J, Li Y (2023) Gat-mf: Graph attention mean field for very large scale multi-agent reinforcement learning. In: Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. pp 685–697. <https://doi.org/10.1145/3580305.3599359>
- He X, Ge H, Hao Z, Hou Y (2025) Maric:an efficient multi-agent real-time intention-based communication model for team cooperation. *Neural Comput Appl* 37(5):3925–3942. <https://doi.org/10.1007/s00521-024-10655-8>
- Helmboldt R (2025) Evaluating saliency map methods in reinforcement learning. Hochschule für Angewandte Wissenschaften Hamburg. <https://doi.org/10.1088/1742-6596/1757/1/012075> (PhD thesis)
- Heuillet A, Couthouis F, Diaz-Rodríguez N (2022) Collective explainable ai: explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. *IEEE Comput Intell Mag* 17(1):59–71. <https://doi.org/10.1109/mci.2021.3129959>
- He P, Wang W, Sun C, Wang Y (2025) Hierarchical cognitive graph autoencoder for multi-agent reinforcement learning. In: Proceedings of the annual meeting of the cognitive science society
- Hu D, Li Z, Ye Z, Peng Y, Xi W, Cai T (2024a) Multi-agent graph reinforcement learning for decentralized volt-var control in power distribution systems. *Int J Electr Power Energy Syst* 155:109531. <https://doi.org/10.1109/spec62217.2024.10893249>
- Hu T, Pu Z, Qiu AX, Yi T, J (2024b) Measuring policy distance for multi-agent reinforcement learning. In: Proceedings of the international joint conference on autonomous agents and multiagent systems. pp 834–844
- Hu S, Shen L, Zhang Y, Tao D (2024c) Learning multi-agent communication from graph modeling perspective. In: 12th International conference on learning representations, ICLR 2024
- Hu Z, Zhang Z, Li H, Chen C, Ding H, Wang Z (2024d) Attention-guided contrastive role representations for multi-agent reinforcement learning. In: 12th International conference on learning representations (ICLR)
- Huang S, Sun C, Wang RQ, Pompili D (2025) Toward adaptive and coordinated transportation systems: a multi-personality multi-agent meta-reinforcement learning framework. *IEEE Trans Intell Transp Syst*. <https://doi.org/10.1109/tits.2025.3560227>
- Huber T, Limmer B, André E (2022) Benchmarking perturbation-based saliency maps for explaining atari agents. *Front Artif Intell* 5:903875. <https://doi.org/10.3389/frai.2022.903875>
- Ibrahim M, Fayad A (2022) Hierarchical strategies for cooperative multi-agent reinforcement learning. <https://doi.org/arXiv:2212.07397> 10.2139/ssrn.4888294 arXiv preprint
- Itaya H, Sagawa T, Hirakawa T, Yamashita T, Fujiyoshi H (2023) Visual explanation for cooperative behavior in multi-agent reinforcement learning. In: 2023 International joint conference on neural networks (IJCNN). IEEE, <https://doi.org/10.1109/iro58592.2024.10801292>

- Iturria-Rivera PE, Gaigalas R, Elsayed M, Bavand M, Ozcan Y, Erol-Kantarci M (2024) Explainable multi-agent reinforcement learning for extended reality codec adaptation. arXiv preprint <https://doi.org/10.2411.14264>
- Jiang K, Liu W, Wang Y, Dong L, Sun C (2024) Discovering latent variables for the tasks with confounders in multi-agent reinforcement learning. *IEEE/CAA J Automatica Sinica* 11(7):1591–1604. <https://doi.org/10.1109/jas.2024.124281>
- Jiang Z, Zhang B, Wei A, Xu Z (2025) Qllm: Do we really need a mixing network for credit assignment in multi-agent reinforcement learning? arXiv preprint [arXiv:2504.12961](https://arxiv.org/abs/2504.12961)
- Ji B, Li G, Xiao G (2023) Enhancing the interpretability of deep multi-agent reinforcement learning via neural logic reasoning. In: *International conference on artificial neural networks*. Springer, pp 199–210. https://doi.org/10.1007/978-3-031-44204-9_17
- Ji M, Xu G, Wang L (2025) Cora: Coalitional rational advantage decomposition for multi-agent policy gradients. <https://doi.org/10.2506.0426510.48550/arXiv.2506.04265> arXiv preprint
- Jin J, Sun G, Chen S, Li Y, Liao Y, Mao W, Shen L (2025a) A multi-agent robust deep reinforcement learning approach for coordination of power distribution networks and microgrids with limited information exchange. *J Modern Power Syst Clean Energy* 1–12. <https://doi.org/10.1109/TPWRS.2023.3276827>
- Jin W, Du H, Liu G, Kim DI (2025b) Curriculum learning with counterfactual group relative policy advantage for multi-agent reinforcement learning. <https://doi.org/10.2506.07548> arXiv preprint Preprint at <https://doi.org/10.2506.07548>
- Jin W, Du H, Zhao B, Tian X, Shi B, Yang G (2025c) A comprehensive survey on multi-agent cooperative decision-making: scenarios, approaches, challenges and perspectives. <https://doi.org/10.2506.13415> arXiv Preprint at
- Kailas S, Jain S, Ravichandar H (2025) Evaluating and improving graph-based explanation methods for multi-agent coordination. <https://doi.org/10.2506.0988910.1109/tg.2023.3263013> arXiv preprint
- Kazhdan D, Shams Z, Lio P (2020) MARLeME: A multi-agent reinforcement learning model extraction library. In: *2020 International joint conference on neural networks (IJCNN)*. Glasgow, UK, pp 1–8
- Kim WK, Lee Y, Woo H (2022) Mean-variance based risk-sensitive reinforcement learning with interpretable attention. In: *Proceedings of the 2022 5th international conference on machine vision and applications*. pp 104–109. <https://doi.org/10.1145/3523111.3523127>
- Kim GS, Lee S, Woo T, Park S (2024a) Cooperative reinforcement learning for military drones over large-scale battlefields. *IEEE Trans Intell*. <https://doi.org/10.1109/tiv.2024.3472213>
- Kim J, Park S, Jung S, Cordeiro C (2024b) Cooperative multi-uav positioning for aerial internet service management: a multi-agent deep reinforcement learning approach. *IEEE Trans Netw Serv Manage* 21(4):3797–3812. <https://doi.org/10.1109/tnsm.2024.3392393>
- Kortvelesy R, Prorok A (2022) Qgnn: Value function factorisation with graph neural networks. arXiv preprint [arXiv:2205.13005](https://arxiv.org/abs/2205.13005)
- Kuhnle A, May MC, Schäfer L, Lanza G (2022) Explainable reinforcement learning in production control of job shop manufacturing system. *Int J Prod Res* 60(19):5812–5834. <https://doi.org/10.1080/00207554.2021.1972179>
- Lan Y, Gao H, Xu X, Fang Q, Zeng Y (2025) Smart: Sequential multiagent reinforcement learning with role assignment using transformer. *IEEE Trans Cogn Dev Syst* 17(3):615–630. <https://doi.org/10.1109/TCDS.2024.3504256>
- Lei L, Wu C, Chen H (2025) Dynamic decoupling-driven cooperative pursuit for multi-uav systems: a multi-agent reinforcement learning policy optimization approach. *Comput Mater*
- Li W, Wang X, Jin B, Luo D, Zha H (2022) Structured cooperative reinforcement learning with time-varying composite action space. *IEEE Trans Pattern Anal Mach Intell* 44(11):8618–8634. <https://doi.org/10.1109/tpami.2021.3102140>
- Li W, Liu W, Shao S, Huang S, Song A (2023) Attention-based intrinsic reward mixing network for credit assignment in multiagent reinforcement learning. *IEEE Trans Games* 16(2):270–281. <https://doi.org/10.1109/tg.2023.3263013>
- Li H, Mahjoub N, Chalaki H, Tadiparthi B, Lee V, K, Pari M, Lewis E, Sycara C, K (2024a) Language grounded multi-agent reinforcement learning with human-interpretable communication. *Adv Neural Inf Process Syst* 37:87908–87933. <https://doi.org/10.5555/3737916.3740706>
- Li L, Donato E, Lomonaco V, Falotico E (2024b) Continual policy distillation of reinforcement learning-based controllers for soft robotic in-hand manipulation. In: *2024 IEEE 7th international conference on soft robotics (RoboSoft)*. IEEE, pp 1026–1033
- Li D, Dong H, Wang L, Qiao B, Qin S, Lin Q, Zhang D, Zhang Q, Xu Z, Zhang B (2024c) Verco: Learning coordinated verbal communication for multi-agent reinforcement learning. [arXiv:2404.17780](https://arxiv.org/abs/2404.17780) arXiv preprint

- Li X, Liu Z, Chen S, Zhang J (2024d) Individual contributions as intrinsic exploration scaffolds for multi-agent reinforcement learning. In: Proceedings of machine learning research, pp 28387–28402. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203844543&partnerID=40&md5=4935fd0bdb41650fdbaf39e18b1d287>
- Li J, Liu Z, Su Z, Zhou Z (2024e) Rbhar: Role-based heterogeneous action representation in multi-agent reinforcement learning. In: International conference on neural information processing. Springer, pp 182–196. https://doi.org/10.1007/978-981-96-6579-2_13
- Li C, Zhang Y, Wang J, Hu Y, Dong S, Li W, Lv T, Fan C, Gao Y (2024f) Optimistic value instructors for cooperative multi-agent reinforcement learning. In: Proceedings of the AAAI conference on artificial intelligence. pp 17453–17460. <https://doi.org/10.1609/aaai.v38i16.29694>
- Li X, Li Z, Yang X, Yu X, Qiu J (2025a) Multi-agent target defense differential game: a hierarchical recognition-allocation-execution learning approach. *Int J Robust Nonlinear Control*. <https://doi.org/10.1002/rnc.70029>
- Li P, Fei Q, Chen Z (2025b) Interpretable multi-agent reinforcement learning via multi-head variational autoencoders. *Appl Intell* 55(7):577. <https://doi.org/10.1007/s10489-025-06473-7>
- Li Y, Cao Z, Qiao J, Hu S (2025c) Nucleolus credit assignment for effective coalitions in multi-agent reinforcement learning. In: Proceedings of the international joint conference on autonomous agents and multiagent systems, AAMAS, pp 1318–1326. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-105009820202&partnerID=40&md5=9f5719eb89f638ab1b900e254d8470b0>
- Li M, Wang Q, Xu Y (2025d) Gtde: Grouped training with decentralized execution for multi-agent actor-critic. In: Proceedings of the AAAI conference on artificial intelligence. pp 18368–18376. <https://doi.org/10.1609/aaai.v39i17.34021>
- Li Z, Agyeman BT, Mitrai I, Daoutidis P (2026) Learning to control inexact benders decomposition via reinforcement learning. *Comput Chem Eng* 205:109461. <https://doi.org/10.1016/j.compchemeng.2025.109461> (<https://www.sciencedirect.com/science/article/pii/S0098135425004648>)
- Liang Y, Wu H, Wang H, Cai H (2024) Asynchronous credit assignment for multi-agent reinforcement learning. <https://doi.org/arXiv:2408.03692> <https://arxiv.org/abs/2408.03692> <https://arxiv.org/abs/2408.03692>
- Lin M, Shi S, Guo Y, Tadiparthi V, Chalaki B, Pari EM, Stepputtis S, Kim W, Campbell J, Sycara K (2025) Speaking the language of teamwork: Llm-guided credit assignment in multi-agent reinforcement learning. *arXiv preprint* <https://doi.org/arXiv:2502.03723>
- Liu Y, Dou Y, Li Y, Xu X, Liu D (2022a) Temporal dynamic weighted graph convolution for multi-agent reinforcement learning. In: Proceedings of the annual meeting of the cognitive science society, openreview. <https://openreview.net/forum?id=X6FjJM3EYC>
- Liu Y, Li Y, Xu X, Liu D, Dou Y (2022b) Rogc: Role-oriented graph convolution based multi-agent reinforcement learning. In: 2022 IEEE international conference on multimedia and expo (ICME). IEEE, <https://doi.org/10.1109/icme52920.2022.9859945>
- Liu X, Liu SY, Zhuang YK, Gao Y (2023a) Explainable reinforcement learning: basic problems exploration and method survey. *J Softw* 34(5):2300–2316 ((in Chinese))
- Liu Z, Wan L, Sui X, Chen Z, Sun K, Lan X (2023b) Deep hierarchical communication graph in multi-agent reinforcement learning. *IJCAI* 208–216. <https://doi.org/10.24963/ijcai.2023/24>
- Liu Z, Zhu Y, Chen C (2023c) Na2q: Neural attention additive model for interpretable multi-agent q-learning. *arXiv pp*. pp 22539–22558
- Liu S, Song J, Zhou Y, Yu N, Chen K, Feng Z, Song M (2024a) Interaction pattern disentangling for multi-agent reinforcement learning. *IEEE Trans Pattern Anal Mach Intell* 46(12):8157–8172. <https://doi.org/10.1109/tpami.2024.3399936/mml>
- Liu C, Cong J, Chen B, Jin Y, Zhu E (2024b) Guiding multi-agent multi-task reinforcement learning by a hierarchical framework with logical reward shaping. *arXiv preprint* <https://doi.org/arXiv:2411.01184>
- Liu D, Ren F, Yan J, Su G, Kato S, Gu W, Zhang M (2025a) Human attention guided multiagent hierarchical reinforcement learning for heterogeneous agents. *Knowl-Based Syst* 316:113377. <https://doi.org/10.1016/j.knosys.2025.113377>
- Liu Z, Zhu Y, Wang Z, Gao Y, Chen C (2025b) Mixrts: Toward interpretable multi-agent reinforcement learning via mixing recurrent soft decision trees. *IEEE Trans Pattern Anal Mach Intell*. <https://doi.org/10.1109/tpami.2025.3540467>
- Lu S, Cai Y, Liu Z, Lian Y, Chen L, Wang H (2025) A preference-based multi-agent federated reinforcement learning algorithm framework for trustworthy interactive urban autonomous driving. *IEEE Trans Intell Transp Syst*. <https://doi.org/10.1109/tits.2025.3543810>
- Luvisutto A, Celani A, Renda F, Stefanini C, De Masi G (2025) Enhancing collaboration in uncertain environment: multi-agent reinforcement learning for underwater monitoring. *Expert Syst Appl* 277:127256. <https://doi.org/10.1016/j.eswa.2025.127256>
- Ma T, Peng K, Rong H, Qian Y (2023) Hierarchical coordination multi-agent reinforcement learning with Spatio-temporal abstraction. *IEEE Trans Emerg Topics Comput Intell* 8(1):533–547

- Ma J, Li C, Feng Z, Xiao L (2025a) Dynamic visual attention-based neuron awakening and shifting in deep reinforcement learning. *Eng Appl Artif Intell* 158:111486. <https://doi.org/10.1016/j.engappai.2025.111486>
- Ma Q, Liu Z, Liu X (2025b) Theoretically interpretable multi-agent reinforcement learning for continuous double auction-based p2p energy trading. *CSEE J Power Energy Syst*. <https://doi.org/10.2139/ssrn.489072>
- Mahjoub O, De Kock R, Singh S, Khelifi W, Vall A, Ka T, Pretorius A (2023) Efficiently quantifying individual agent importance in cooperative marl <https://doi.org/arXiv:2312.08466> arXiv preprint
- Mahmud S, Saisubramanian S, Zilberstein S (2023) Explanation-guided reward alignment. *IJCAI*. pp 473–482
- Maliha M, Hougen D (2025) Magic-mask: Multi-agent guided inter-agent collaboration with mask-based explainability for reinforcement learning. <https://doi.org/arXiv:2510.0027410.48550/arXiv.2510.00274> arXiv preprint
- Mehmeti A, Gigante G, Venticinque S (2025) Explainable reinforcement learning for assisting air traffic controllers. In: *International conference on advanced information networking and applications*. Springer, pp 148–157
- Meli D, Morasso C, Castellini A, Farinelli A (2025) What are you saying? explaining communication in multi-agent reinforcement learning. In: *11th Italian workshop on artificial intelligence and robotics*, vol 2024. (short6.pdf)
- Meng X, Tan Y (2024) Pmac: Personalized multi-agent communication. In: *Proceedings of the AAAI conference on artificial intelligence*. pp 17505–17513. <https://doi.org/10.1609/aaai.v38i16.29700>
- Milani S, Topin N, Veloso M, Fang F (2022a) A survey of explainable reinforcement learning <https://doi.org/arxiv.org/abs/2202.08434> Preprint at
- Milani S, Zhang Z, Topin N, Shi ZR, Kamhoua C, Papalexakis EE, Fang F (2022b) Maviper: Learning decision tree policies for interpretable multi-agent reinforcement learning. In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer, pp 251–266. https://doi.org/10.1007/978-3-031-26412-2_16
- Miyake K, Ito H, Faloutsos C, Matsumoto H, Morishima A (2024) Netevolve: Social network forecasting using multi-agent reinforcement learning with interpretable features. In: *Proceedings of the ACM web conference*. pp 2542–2551. <https://doi.org/10.1145/3589334.3647982>
- Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533
- Motokawa Y, Sugawara T (2023a) eda3-x: Distributed attentional actor architecture for interpretability of coordinated behaviors in multi-agent systems. *Appl Sci* 13(14):8454. <https://doi.org/10.3390/app13148454>
- Motokawa Y, Sugawara T (2023b) Strategy-following multi-agent deep reinforcement learning through external high-level instruction. *Procedia Comput Sci* 225:2798–2807. <https://doi.org/10.1016/j.procs.2023.10.272>
- Motokawa Y, Sugawara T (2023c) Interpretability for conditional coordinated behavior in multi-agent reinforcement learning. In: *2023 International joint conference on neural networks (IJCNN)*. IEEE, <https://doi.org/10.1109/ijcnn54540.2023.10191825>
- Mu X, Zhuo HH, Chen C, Zhang K, Yu C, Hao J (2025) Hierarchical task network-enhanced multi-agent reinforcement learning: toward efficient cooperative strategies. *Neural Netw* 186:107254. <https://doi.org/10.2139/ssrn.4888294>
- Nagpal K, Dong D, Bouvier JB, Mehr N (2025) Leveraging large language models for effective and explainable multi-agent credit assignment. In: *Proceedings of the international joint conference on autonomous agents and multiagent systems*. pp 1501–1510
- Na H, Lee K, Lee S, Moon IC (2025) Trajectory-class-aware multi-agent reinforcement learning. In: *The thirteenth international conference on learning representations*
- Ning Z, Xie L (2024) A survey on multi-agent reinforcement learning and its application. *J Autom Intell* 3(2):73–91. <https://doi.org/10.1016/j.jai.2024.02.003>
- Oliehoek FA, Amato C et al (2016) A concise introduction to decentralized POMDPs, vol 1. Springer. https://doi.org/10.1007/978-3-319-28929-8_3
- Omidshafiei S, Kapishnikov A, Assogba Y, Dixon L, Kim B (2022) Beyond rewards: a hierarchical perspective on offline multiagent behavioral analysis. In: *Advances in neural information processing systems*. pp 3444–3460
- Otto F, Celik O, Zhou H, Ziesche H, Ngo VA, Neumann G (2023) Proceedings of the 6th conference on robot learning. pp 1244–1265

- Pandya R, Zhao M, Liu C, Simmons R, Admoni H (2024) Multi-agent strategy explanations for human-robot collaboration. In: 2024 IEEE international conference on robotics and automation (ICRA). IEEE, pp 17351–17357. <https://doi.org/10.1109/icra57147.2024.10610720>
- Papoudakis G, Christianos F, Rahman A, Albrecht SV (2019) Dealing with non-stationarity in multi-agent deep reinforcement learning. arXiv preprint <https://doi.org/arXiv:1906.04737>
- Peng K, Ma T, Yu X, Rong H, Qian Y, Al-Nabhan N (2023) Gema: An adaptive multiagent reinforcement learning framework with group communication for complex and similar tasks coordination. *IEEE Trans Games* 16(3):670–682. <https://doi.org/10.1109/tg.2023.3346394>
- Peng K, Li P, Hao J (2025a) Enhancing graph-based coordination with evolutionary algorithms for episodic multi-agent reinforcement learning. In: Proceedings of the 24th international conference on autonomous agents and multiagent systems. pp 1623–1631
- Peng K, Zhu S, Ma T (2025b) Stpe-marl: Spatio-temporal multi-agent population evolution reinforcement learning. *ACM Trans Intell Syst Technol*
- Pesce E, Montana G (2023) Learning multi-agent coordination through connectivity-driven communication. *Mach Learn* 112(2):483–514. <https://doi.org/10.1007/s10994-022-06286-6>
- Qi S, Huang X, Peng P, Huang X, Zhang J, Wang X (2022) Cascaded attention: adaptive and gated graph attention network for multiagent reinforcement learning. *IEEE Trans Neural Netw Learn Syst* 35(3):3769–3779. <https://doi.org/10.1109/tnnls.2022.3197918>
- Qiu D, Dong Z, Zhang X et al (2022) Safe reinforcement learning for real-time automatic control in a smart energy-hub. *Appl Energy* 309:118403
- Qiu D, Wang Y, Wang J, Zhang N, Strbac G, Kang C (2023) Resilience-oriented coordination of networked microgrids: a Shapley q-value learning approach. *IEEE Trans Power Syst* 39(2):3401–3416
- Radke D, Larson K, Brecht T (2023) The importance of credo in multiagent learning. In: Proceedings of the 2023 international conference on autonomous agents and multiagent systems. International Foundation for Autonomous Agents and Multiagent Systems. Richland, SC, AAMAS, pp 1–9. <https://doi.org/10.5555/3545946.3598902>
- Reiestad E (2025) Better together? Explaining emergent behavior in multi-agent reinforcement learning. Master's thesis, NTNU
- Ren Y, Ye Z, Chen Y, Jiang X, Song G (2023) Soft-hgrns: soft hierarchical graph recurrent networks for multi-agent partially observable environments. *Front Inf Technol Electron Eng* 24(1):117–130. <https://doi.org/10.1631/fitee.22e0073>
- Ren F, Dong W, Zhao X, Zhang F, Kong Y, Yang Q (2024a) Two-layer coordinated reinforcement learning for traffic signal control in traffic network. *Expert Syst Appl* 235:121111. <https://doi.org/10.1016/j.eswa.2023.121111> (<https://www.sciencedirect.com/science/article/pii/S0957417423016135>)
- Ren Y, Zhang H, Du L, Zhang Z, Zhang J, Li H (2024b) Stealthy black-box attack with dynamic threshold against marl-based traffic signal control system. *IEEE Trans Ind Inf* 20(10):12021–12031. <https://doi.org/10.1109/tii.2024.3413356>
- Ribeiro MT, Singh S, Guestrin C (2018) Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence. p 32
- Roghani S, Koyuncu E (2025) An explainable and fair hierarchical rl-based alternative trajectory proposal framework for autonomous u-plan approval in u-space. In: 2025 Integrated communications, navigation and surveillance conference (ICNS). IEEE, pp 1–13
- Romdhana A, Merlo A, Ceccato M, Tonella P (2022) Deep reinforcement learning for black-box testing of android apps. *ACM Trans Softw Eng Methodol (TOSEM)* 31(4):1–29
- Ruan J, Hao X, Li D, Mao H (2023) Learning to collaborate by grouping: a consensus-oriented strategy for multi-agent reinforcement learning. <https://doi.org/arXiv:2307.1553010.3233/faia230493> arXiv preprint
- Ruan J, Xie R, Xiong X, Xu S, Xu B (2024) Made: Multi-scale decision enhancement for multi-agent reinforcement learning. In: ICASSP 2024–2024 IEEE international conference on acoustics. IEEE, Speech and Signal Processing (ICASSP), pp 31–35
- Schier M, Schubert F, Rosenhahn B (2025) Explainable reinforcement learning via dynamic mixture policies. In: 2025 IEEE international conference on robotics and automation (ICRA), pp. To be published. <https://doi.org/10.1109/icra55743.2025.11127782>
- Selimović T, Peti M, Bogdan S (2024) Multi-agent active perception based on reinforcement learning and pomdp. *IEEE Access* 12:48004–48016. <https://doi.org/10.2139/ssrn.4506843>
- Selmonaj A, Antonucci A, Schneider A, Rügsegger M, Sommer M (2025a) Explaining strategic decisions in multi-agent reinforcement learning for aerial combat tactics. arXiv preprint [arXiv:2505.11311](https://arxiv.org/abs/2505.11311)
- Selmonaj A, Strupl M, Szehr O, Antonucci A (2025b) Understanding action effects through instrumental empowerment in multi-agent reinforcement learning. <https://doi.org/arXiv:2508.1565210.3233/faia250947> arXiv preprint

- Shah AB, Wen Y, Chen J, Wu X, Fu X (2024) Safe offline-to-online multi-agent decision transformer: a safety conscious sequence modeling approach. In: 2024 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 12400–12407. <https://doi.org/10.1109/iros58592.2024.10801292>
- Shao J, Lou Z, Zhang H, Jiang Y, He S, Ji X (2022) Self-organized group for cooperative multi-agent reinforcement learning. *Adv Neural Inf Process Syst* 35:5711–5723
- Sharma PK, Zaroukian E, Asher DE, Howell B (2022) Emergent behaviors in multi-agent target acquisition. In: *Artificial intelligence and machine learning for multi-domain operations applications IV*, SPIE. pp 361–371
- Sheng J, Wang X, Jin B, Yan J, Li W, Chang TH, Wang J, Zha H (2022) Learning structured communication for multi-agent reinforcement learning. *Auton Agent Multi-Agent Syst* 36(2):50. <https://doi.org/10.1007/s10458-022-09580-8>
- Shi S, Zhang Y, Leibo JZ, Du Y (2025) Resolving complex social dilemmas by aligning preferences with counterfactual regret. openreview.net/forum?id=CgkGFeSpo0, openreview.net/pdf?id=CgkGFeSpo0. (iCLR 2025 OpenReview)
- Soulé J, Jamont JP, Ocelllo M, Traonouez LM, Théron P (2025a) An organizationally-oriented approach to enhancing explainability and control in multi-agent reinforcement learning. In: *Proceedings of the international joint conference on autonomous agents and multiagent systems*. pp 1968–1976
- Soulé J, Jamont JP, Ocelllo M, Traonouez LM, Théron P (2025b) Streamlining resilient Kubernetes autoscaling with multi-agent systems via an automated online design framework. <https://doi.org/arXiv:2505.21559>. [10.1109/cloud67622.2025.00015](https://doi.org/10.1109/cloud67622.2025.00015) arXiv preprint
- Sun Z, Wu H, Shi Y, Yu X, Gao Y, Pei W, Yang Z, Piao H, Hou Y (2023) Multi-agent air combat with two-stage graph-attention communication. *Neural Comput Appl* 35(27):19765–19781. <https://doi.org/10.1007/s00521-023-08784-7>
- Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: *Proceedings of the international conference on machine learning*. pp 3319–3328
- Sunehag P, Lever G, Gruslys A, Czarniecki WM, Zambaldi V, Jaderberg M, Lanctot M, Sonnerat N, Leibo JZ, Tuyls K (2017) Value-decomposition networks for cooperative multi-agent learning. <https://doi.org/arXiv:1706.05296>. [10.1016/j.neunet.2025.107692](https://doi.org/10.1016/j.neunet.2025.107692) arXiv preprint
- Su D, Peng H, Zeng G, Li P, Li A, Pan Y (2025) Emergence of cooperation in multi-agent reinforcement learning via coalition labeling and structural entropy. In: *Proceedings of the 2025 SIAM international conference on data mining (SDM)*, SIAM. pp 507–515. <https://doi.org/10.1137/1.9781611978520.55>
- Tabrez A, Luebbers MB, Ruvane K, Rabin AH, King KW, Gerichs W, Hayes B (2024) Hierarchical multi-agent reinforcement learning with explainable decision support for human-robot teams. In: *Proceedings of the ACM/IEEE international conference on human-robot interaction*. Boulder, CO, USA, (hRI '24)
- Taghavi M, Farnoosh R (2025a) Quantum computing and neuromorphic computing for safe, reliable, and explainable multi-agent reinforcement learning: Optimal control in autonomous robotics. *Iran J Comput Sci* 1–17. <https://doi.org/10.1007/s42044-025-00306-z>
- Taghavi M, Farnoosh R (2025b) Latent variable modeling in multi-agent reinforcement learning via expectation-maximization for uav-based wildlife protection. <https://doi.org/arXiv:2509.02579> arXiv preprint Preprint at <https://doi.org/arxiv.org/abs/2509.02579>
- Takagi Y, Tabalba R, Kirshenbaum N, Leigh J (2024) Abstracted trajectory visualization for explainability in reinforcement learning. In: 2024 IEEE conference on artificial intelligence (CAI). IEEE, pp 75–82
- Tesic M, Zhao Y, Leibo JZ, Trivedi RS, Hernandez-Orallo J (2025) Beyond the high score. In: *Prosocial ability profiles of multi-agent populations*. arXiv preprint <https://doi.org/arXiv:2509.14485>
- Thakre B, Yadav U, Bondre SV (2025) Deep reinforcement learning in healthcare and biomedical application. *Innov Biomed Eng*. Academic Press, pp 241–299
- Tjoa E, Guan C (2020) A survey on explainable artificial intelligence (xai): toward medical xai. *IEEE Trans Neural Netw Learn Syst* 32(11):4793–4813
- Toquebiau M, Jun JY, Benamar F, Bredeche N (2025) Towards language-augmented multi-agent deep reinforcement learning. <https://doi.org/arXiv:2506.05236> arXiv preprint Preprint at <https://doi.org/arxiv.org/abs/2506.05236>
- Towers M, Du Y, Freeman C, Norman TJ (2024) Explaining an agent's future beliefs through temporally decomposing future reward estimators. arXiv preprint [arXiv:2408.08230](https://arxiv.org/abs/2408.08230)
- Vos D, Verwer S (2023) Optimal decision tree policies for Markov decision processes. <https://doi.org/arXiv:2301.13185>. [10.24963/ijcai.2023/606](https://doi.org/10.24963/ijcai.2023/606) arXiv preprint
- Wai KP, Geng M, Pateria S, Subagdja B, Tan AH (2024) Explaining sequences of actions in multi-agent deep reinforcement learning models. In: *Proceedings of the 23rd international conference on autonomous agents and multiagent systems*. pp 2537–2539
- Wang C, Aouf N (2024) Explainable deep adversarial reinforcement learning approach for robust autonomous driving. In: *IEEE transactions on intelligent vehicles*

- Wang J, Xu W, Gu Y, Song W, Green TC (2021) Multi-agent reinforcement learning for active voltage control on power distribution networks. *Adv Neural Inf Process Syst* 34:3271–3284. <https://doi.org/10.2139/ssrn.5371728>
- Wang D, Gao N, Liu D, Li J, Lewis FL (2023) Recent progress in reinforcement learning and adaptive dynamic programming for advanced control applications. *IEEE/CAA J Automatica Sinica* 11(1):18–36
- Wang D, Antoniadis A, Luong KD, Zhang E, Kosan M, Li J, Singh A, Wang WY, Li L (2024) Global human-guided counterfactual explanations for molecular properties via reinforcement learning. In: *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. pp 2991–3000
- Wang Z, Du Y, Zhang Y, Fang M, Huang B (2025a) Macca: Offline multi-agent reinforcement learning with causal credit assignment. *Trans Mach Learn Res*
- Wang Z, Zhang Z, Fang F, Du Y (2025b) M3HF: Multi-agent reinforcement learning from multi-phase human feedback of mixed quality. In: *Proceedings of machine learning research*. pp 65429–65448
- Wan K, Liu Y, Liu H, Xu X (2025a) A framework for modeling cognitive processes in intelligent agents using behavior trees. In: *Proceedings of the 2025 5th international conference on internet of things and machine learning*. pp 267–271. <https://doi.org/10.1145/3749566.3749619>
- Wan Z, Li Y, Wen X, Song Y, Wang H, Yang L, Schmidt M, Wang J, Zhang W, Hu S (2025b) Rema: Learning to meta-think for llms with multi-agent reinforcement learning. <https://doi.org/10.48550/arXiv.2503.09501> arXiv preprint
- Wei Q, Li Y, Zhang J, Wang FY (2022) Vgn: Value decomposition with graph attention networks for multi-agent reinforcement learning. *IEEE Trans Neural Netw Learn Syst* 35(1):182–195. <https://doi.org/10.1109/tnnls.2022.3172572>
- Wei Y, Shan X, Miao R, Li J (2025) Lero: Llm-driven evolutionary framework with hybrid rewards and enhanced observation for multi-agent reinforcement learning. In: *International conference on intelligent computing*. Springer, pp 15–26. https://doi.org/10.1007/978-981-96-9894-3_2
- Wenhan Y, Qian L, Chua TJ, Zhao J (2024) Counterfactual reward estimation for credit assignment in multi-agent deep reinforcement learning over wireless video transmission. In: *2024 IEEE 44th international conference on distributed computing systems (ICDCS)*. IEEE, pp 1177–1189. <https://doi.org/10.1109/icdcs60910.2024.00112>
- Wu H, Ghadami A, Bayrak AE, Smereka JM, Epureanu BI (2022) Evaluating emergent coordination in multi-agent task allocation through causal inference and sub-team identification. *IEEE Robot Automat Lett* 8(2):728–735
- Xiao B, Ramasubramanian B, Poovendran R (2022) Agent-temporal attention for reward redistribution in episodic multi-agent reinforcement learning. In: *Proceedings of the international joint conference on autonomous agents and multiagent systems*. pp 1391–1399
- Xie T, Gao Z, Ren Q, Luo H, Hong Y, Dai B, Zhou J, Qiu K, Wu Z, Luo C (2025) Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. <https://doi.org/10.48550/arXiv.2502.14768> arXiv preprint
- Xi-Jia Z, Guo Y, Chen S, Stepputtis S, Gombolay M, Sycara K, Campbell J (2025) Model-agnostic policy explanations with large language models. <https://doi.org/10.48550/arXiv.2504.05625> arXiv preprint
- Xiong Y, Hu Z, Huang Y, Wu R, Guan K, Fang X, Jiang J, Zhou T, Hu Y, Liu H (2024) Xrl-bench: a benchmark for evaluating and comparing explainable reinforcement learning techniques. In: *proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. pp 6073–6082. <https://doi.org/10.1145/3637528.3671595>
- Xu Z, Bai Y, Zhang B, Li D, Fan G (2023) Haven: Hierarchical cooperative multi-agent reinforcement learning with dual coordination mechanism. In: *Proceedings of the AAAI conference on artificial intelligence*. pp 11735–11743
- Xu X, Wu F, Bilal M, Xia X, Dou W, Yao L, Zhong W (2024) Xrl-shap-cache: an explainable reinforcement learning approach for intelligent edge service caching in content delivery networks. *Sci China Inf Sci* 67(7):170303. <https://doi.org/10.1007/s11432-023-3987-y>
- Yang M, Zhao K, Wang Y, Dong R, Du Y, Liu F, Zhou M, U LH, (2024) Team-wise effective communication in multi-agent reinforcement learning. *Auton Agent Multi-Agent Syst* 38(2):36. <https://doi.org/10.1007/s10458-024-09665-6>
- Yang B, Gao L, Zhou F, Yao H, Fu Y, Sun Z, Tian F, Ren H (2025a) A coordination optimization framework for multi-agent reinforcement learning based on reward redistribution and experience reutilization. *Electronics* 14(12):2361. <https://doi.org/10.3390/electronics14122361>
- Yang S, Hu H, Shi E, Zhang S (2025b) Dual-level behavioral consistency for inter-group and intra-group coordination in multi-agent systems. arXiv preprint [arXiv:2506.18651](https://arxiv.org/abs/2506.18651)
- Yang H, Shi D, Xie JS, Chen G, Qiu Y, Yang C, S (2025c) Acting beyond learning: Imagination-assisted decision-making in the visual-based multi-agent cooperative scenarios. In: *Proceedings of the AAAI conference on artificial intelligence*. pp 21947–21955. <https://doi.org/10.1609/aaai.v39i20.35502>

- Yao D, Sun C, Li K, Zhou K, Li H, Wang R, Liu L (2024) Cmix: Causal value decomposition for cooperative multi-agent reinforcement learning. In: 2024 IEEE international conference on systems, man, and cybernetics. IEEE, pp 4105–4112. <https://doi.org/10.1109/smc54092.2024.10831507>
- Yoo B, Shin Y, Kim H, Yang CE, J (2025) Adaptive episode length adjustment for multi-agent reinforcement learning. In: Proceedings of the international joint conference on autonomous agents and multiagent systems. pp 2253–2261
- Yuan L, Wang J, Zhang F, Wang C, Zhang Z, Yu Y, Zhang C (2022) Multi-agent incentive communication via decentralized teammate modeling. In: Proceedings of the AAAI conference on artificial intelligence. pp 9466–9474. <https://doi.org/10.1609/aaai.v36i9.21179>
- Zabounidis R, Campbell J, Stepputtis S, Hughes D, Sycara KP (2023) Concept learning for interpretable multi-agent reinforcement learning. In: Conference on robot learning, PMLR. pp 1828–1837
- Zhang Q, Xu B (2024) Towards zero-shot generalization: mutual information-guided hierarchical multi-agent coordination. In: 2024 International joint conference on neural networks (IJCNN). IEEE, pp 1–9
- Zhang H, Li G, Liu CH, Wang G, Tang J (2023) Himacmic: hierarchical multi-agent deep reinforcement learning with dynamic asynchronous macro strategy. In: Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. pp 3239–3248. <https://doi.org/10.1145/3580305.3599379>
- Zhang Y, Zheng G, Liu Z, Li Q, Zeng H (2024a) Marlens: understanding multi-agent reinforcement learning for traffic signal control via visual analytics. IEEE Trans Visual Comput Graph. <https://doi.org/10.1109/tvcg.2024.3392587>
- Zhang Y, Du Y, Huang B, Fang M, Pechenizkiy M (2024b) A causality-inspired spatial-temporal return decomposition approach for multi-agent reinforcement learning. In: NeurIPS 2024 causal representation learning workshop, openreview.net/pdf?id=tNkThyt18
- Zhang Q, Wang P, Ni C, Rong W, Ren C, Liu X, Shan H (2025a) Comp-ig: coordinate-aware multi-agent pathfinding via learnable communication graphs. J Supercomput 81(8):881. <https://doi.org/10.1007/s11227-025-07377-6>
- Zhang M, Fang Z, Wang T, Lu S, Wang X, Shi T (2025b) Ccma: A framework for cascading cooperative multi-agent in autonomous driving merging using large language models. Expert Syst Appl. <https://doi.org/10.1016/j.eswa.2025.127717>
- Zhang Z, Cheng B, Wang Z, Zhou Y, Li G, Lu P, He B, Chen J (2025c) Pagnet: Pluggable adaptive generative networks for information completion in multi-agent communication. arXiv preprint [arXiv:2502.03845](https://arxiv.org/abs/2502.03845) Preprint at <https://arxiv.org/abs/2502.03845>
- Zhao X, Xie Y (2025) Multi-level advantage credit assignment for cooperative multi-agent reinforcement learning. In: The 28th international conference on artificial intelligence and statistics. <https://doi.org/10.3390/app12146938>
- Zhu T, Shi X, Xu X, Gui J, Cao J (2024) Hypercomm: Hypergraph-based communication in multi-agent reinforcement learning. Neural Netw 178:106432. <https://doi.org/10.1016/j.neunet.2024.106432>
- Zhu C, Dastani M, Wang S (2025) Communication with factorized policy gradients in multi-agent deep reinforcement learning. Neural Comput Appl. <https://doi.org/10.1007/s00521-025-11272-9>
- Zhuang Z, Lu CY, Chang Y, Wang YK, Do T, Lin CT (2025) Aegis: Human attention-based explainable guidance for intelligent vehicle systems. In: Proceedings of the 2025 CHI conference on human factors in computing systems. pp 1–19. <https://doi.org/10.1145/3706598.3713779>