vehicles

MDPI

Systematic Review

Multi-Objective Multi-Agent Reinforcement Learning for Urban Public Transport Fleets: A Systematic Review

Phan Gia Bao Le

Faculty of Information Technology, Swinburne Vietnam, FPT University, Ha Noi 11314, Vietnam; baolpg2@fe.edu.vn

Abstract

Urban public transport fleets face challenges in balancing operational efficiency and passenger-related service quality. This systematic review, reported with reference to PRISMA 2020, synthesises 41 studies identified through five academic databases and search platforms between January 2018 and May 2026. The corpus encompasses direct public-transport applications, general Multi-Agent Reinforcement Learning (MARL) and Multi-Objective Reinforcement Learning (MORL) research, and supporting literature from adjacent domains. The review examines MARL architectures, multi-objective reward formulations, coordination mechanisms, and operational outcomes while distinguishing application-specific findings from methodological insights. A conceptual Multi-Objective Partially Observable Stochastic Game (MO-POSG) formulation illustrates interactions between fleet-level objectives and passenger-related considerations. The synthesis discusses Centralised Training and Decentralised Execution (CTDE), value-based and policy-based learning, and scalarisation, constraint-based, and Pareto-based optimisation without assuming universal algorithmic superiority. Evidence from direct public-transport applications indicates potential benefits for coordinated vehicle control and service regularity, although reported passenger waiting-time, travel-time, and delay outcomes remain difficult to compare across studies. Offline MARL and conservative policy rectification are considered as potential approaches for limiting unsupported policy updates, without guaranteeing operational safety. Key deployment challenges include sensing uncertainty, adversarial vulnerabilities, and driver acceptance. Overall, the findings highlight limited application-specific validation and the need for transparent, comparable evaluation before broader real-world deployment.

Keywords: Multi-Agent Reinforcement Learning (MARL); multi-objective optimisation; public transport fleets

Check for updates

Received: 18 September 2026
Revised: 12 November 2026
Accepted: 14 November 2026
Published: 18 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Urban public transport networks face persistent challenges involving congestion, operational efficiency, service reliability, and passenger satisfaction. Addressing these challenges requires coordinated spatial and temporal decisions concerning vehicle routing, scheduling, fleet allocation, and service control. Conventional transport optimisation has addressed these problems through mathematical programming and network-based formulations, providing established mechanisms for representing operational constraints [1,2]. However, computational requirements and repeated optimisation may become challenging in large-scale, stochastic environments with changing demand and interdependent vehicle decisions, depending on the formulation and solution method.

Established spatio-temporal approaches, including mixed-integer optimisation, time–space networks, and rolling-horizon frameworks, can represent spatial connectivity, temporal dependencies, vehicle availability, and service constraints. In tourist transport, additional considerations include attraction time windows, accessibility, route feasibility, and heterogeneous passenger preferences. Le et al. [3] reviewed optimisation models and data-integration approaches for urban tourist routing, highlighting alternative representations of these constraints. Although conventional methods offer structured and interpretable formulations, their applicability under dynamic conditions depends on computational resources, available information, and problem complexity rather than any inherent inability to address uncertainty [1,2].

Alongside these approaches, Deep Reinforcement Learning (DRL) has been investigated for sequential transport decisions, including traffic signal control and routing [4,5]. Single-agent Reinforcement Learning (RL) may face coordination challenges when multiple vehicles interact through shared infrastructure [6,7]. Multi-Agent Reinforcement Learning (MARL) provides a framework for learning interdependent policies through agent interactions, potentially supporting decentralised fleet decisions under dynamic and partially observable conditions [8–10]. The network-level implications of routing decisions are illustrated by Le et al. [11], who examined the macroscopic effects of alternative tourist-bus routing strategies in simulation. However, these findings do not establish general superiority over established spatio-temporal optimisation methods.

Existing reviews cover DRL and MARL across Intelligent Transportation Systems, logistics, and commercial mobility applications [12,13]. Public and tourist transport differ in service schedules, passenger demand, capacity, route restrictions, and service-quality obligations [14,15]. Although multi-objective optimisation is well established in transportation and Multi-Objective Reinforcement Learning (MORL) has been extensively studied [2,16], a focused synthesis is needed to examine how MARL-based coordination represents operational and passenger-related objectives in these settings. The gap concerns application-specific evidence and methodological transferability rather than the absence of established optimisation methods. Table 1 summarises related reviews and their scope relative to the present study.

Table 1. Summary of related review articles and identified research gaps.

Reference	Primary Focus	Scope and Limitations Relative to This Review
Haydari & Yılmaz [4]	Broad review of DRL applications in ITS, including traffic signal control and routing	Does not specifically synthesise MARL architectures, fleet coordination, and passenger-related objectives in public and tourist transport
Zong et al. [13] Qin et al. [12]	Learning-based methods across transportation and related fleet-management domains	Different application emphasis; findings from commercial mobility or logistics require contextual interpretation before transfer to public and tourist transport
Hayes et al. [16]	Methodological foundations and applications of multi-objective reinforcement learning	Broad MORL perspective rather than a dedicated synthesis of public-transport fleet operations and their deployment constraints
This Review	Structured synthesis of MARL methods relevant to urban public and tourist transport	Examines fleet coordination, multi-objective reward formulations, application-specific evidence, and methodological and socio-technical deployment limitations

This review systematically synthesises MARL research relevant to urban public and tourist transport, distinguishing direct application evidence from supporting methodologi-

cal and adjacent-domain literature. It also presents a conceptual mathematical formulation to examine multi-objective interactions without implying experimental validation or superiority over alternative optimisation approaches. Its four main contributions are:

1. A structured synthesis of MARL methods relevant to public and tourist transport, highlighting differences from commercial mobility and logistics applications.
2. An examination of multi-objective reward formulations and trade-off strategies involving operational efficiency, passenger satisfaction, comfort, and scenic quality.
3. A critical discussion of offline MARL and conservative policy rectification as potential approaches to limiting unsupported policy updates, without assuming operational safety guarantees.
4. An assessment of algorithmic and socio-technical deployment challenges, including sensing uncertainty, adversarial vulnerabilities, driver acceptance, and simulation-to-reality transfer.

These contributions address four research questions:

RQ1: What MARL architectures and coordination mechanisms are represented in research relevant to urban public and tourist transport fleets?
RQ2: How are multi-objective reward functions and trade-off resolution strategies formulated within MARL-based transport coordination?
RQ3: To what extent does the reviewed evidence address trade-offs between operational efficiency and passenger-related service quality, including travel time, comfort, and tourist experience?
RQ4: What algorithmic, security-related, and socio-technical challenges affect the potential transfer of MARL approaches from simulation to real-world public and tourist transport operations?

The remainder of the paper is organised as follows. Section 2 describes the review methodology and evidence mapping. Section 3 introduces the synthesised Multi-Objective Partially Observable Stochastic Game (MO-POSG) formulation and objective-resolution strategies. Section 4 examines MARL architectures, coordination mechanisms, and operational evidence. Section 5 discusses deployment challenges, while Section 6 outlines research directions and review limitations. Section 7 concludes the review.

2. Review Methodology and Descriptive Analysis

This systematic review was reported with reference to the PRISMA 2020 statement [17]. The completed checklist is provided in Supplementary File S1, while study-level characteristics and search documentation are presented in Supplementary Files S2 and S3, respectively. The review was not prospectively registered, and no formal protocol was prepared before the search. PRISMA was used as a reporting framework for documenting identification, screening, eligibility assessment, and inclusion. This section describes the search strategy, selection criteria, data extraction, and descriptive synthesis of the 41 mapped studies.

2.1. Search Strategy

The literature search was structured around three conceptual clusters: (1) algorithmic architecture, (2) operational domain, and (3) optimisation focus. A general Boolean expression combined terms relating to Multi-Agent Reinforcement Learning (MARL), public and tourist transport operations, and multi-objective or constrained decision-making:

("MARL" OR "Multi-Agent Reinforcement Learning") AND ("Public transport" OR "Bus" OR "Tourist Fleet" OR "fleet management" OR "vehicle routing" OR "demand-responsive transport") AND ("Multi-Objective" OR "Routing" OR "optimisation" OR "optimization").

Broader routing and optimisation terms were retained to identify potentially relevant studies without explicitly labelled multi-objective formulations. Final eligibility was determined through screening and full-text assessment rather than keyword matching alone (Sections 2.2 and 2.3).

The original literature search was conducted in May 2026 across Google Scholar, ScienceDirect, IEEE Xplore, Scopus, and Web of Science, covering publications from January 2018 to May 2026. The Boolean expression provided a common conceptual framework adapted to each platform's search syntax, indexed fields, and available filtering capabilities. Supplementary File S3 documents the source-specific retrieval counts, search implementation, and reproducibility considerations. Although the original search framework and identification counts were retained, the complete executable queries, exact execution dates, and filtering configurations could not be fully reconstructed for all databases, limiting independent reproduction of the historical searches.

For IEEE Xplore, the original PRISMA identification process reported 410 records. During the first review round, an independent reviewer reported retrieving 510 records, comprising 217 journal articles and 293 conference papers, using a similar Boolean search framework. This observation was documented in the review report dated 8 October 2026. A subsequent author-initiated search conducted on 11 November 2026, using the Boolean expression presented above within the IEEE Xplore All Metadata search scope, returned 516 records, comprising 218 journal articles, 296 conference papers, one book, and one magazine. The updated search covered publications from January 2018 to November 2026, extending beyond the original review period ending in May 2026. The corresponding retrieval information and methodological distinctions are presented in Supplementary File S3, Section S3.4 and Table S3.3.

The three retrieval counts represent distinct search events rather than directly comparable identification datasets. Differences in executable search configurations, searchable fields, publication coverage, document-type settings, or database indexing may have contributed to the variation in retrieved records. However, the available documentation does not establish the relative contribution of these factors. In particular, the updated search returned 514 journal and conference records compared with the 510 records reported by the reviewer, but this difference does not demonstrate that four additional eligible studies were identified. The original count of 410 IEEE Xplore records was therefore retained in the PRISMA identification dataset to preserve consistency with the documented screening and eligibility process. Neither the reviewer-reported retrieval nor the updated author search was retrospectively incorporated into the original identification total of 1856 records.

These differences introduce uncertainty regarding the reproducibility and potential completeness of the original literature identification process. Because a documented record-level comparison and eligibility assessment across the retrieval datasets were not established, the numerical discrepancies cannot be interpreted as evidence of a specific number of omitted eligible studies. Consequently, the 41 studies retained for synthesis represent the corpus selected through the documented original PRISMA process rather than a demonstrably exhaustive collection of all potentially relevant publications. This limitation may affect the completeness of the evidence mapping and the interpretation of methodological distributions, although its magnitude cannot be determined from the available documentation.

The 2018 starting point was selected to focus the structured synthesis on more recent deep MARL and multi-agent coordination research. Earlier foundational studies were cited where relevant to the theoretical discussion, including reinforcement learning, value factorisation, and multi-agent methods [4,7,9,18]. These references were distinguished from studies identified through the predefined search and were not automatically counted

among the 41 mapped studies. The selection results and evidential roles are detailed in Sections 2.4 and 2.6 and Supplementary File S2.

2.2. Inclusion and Exclusion Criteria

The review distinguished studies directly evaluating Multi-Agent Reinforcement Learning (MARL) in public or tourist transport from supporting methodological and contextual literature. This distinction reflects differences in evidential relevance, as methodological contributions from adjacent domains do not independently establish application-specific effectiveness.

The following criteria defined direct public-transport application evidence:

- ***Publication Standards:*** English-language peer-reviewed journal articles, conference proceedings, and relevant academic preprints published between January 2018 and May 2026 were considered. Preprints required sufficient methodological or evaluation information and were distinguished from peer-reviewed publications.
- ***Methodological Focus:*** Studies were required to employ MARL or reinforcement-learning-based multi-agent coordination as a substantive methodological component. Relevant approaches included independent multi-agent learning, Centralised Training and Decentralised Execution (CTDE), value factorisation, and multi-agent actor–critic methods. General MARL/MORL research outside public transport was treated as methodological support.
- ***Application Domain:*** Direct evidence concerned urban public or tourist passenger transport, including fixed-route buses, tourist fleets, demand-responsive services, and comparable operations. Relevant characteristics included schedules, passenger demand, vehicle capacity, route restrictions, and inter-vehicle interactions. Adjacent applications were distinguished from direct public-transport evaluations.
- ***Problem Formulation:*** Studies addressed routing, scheduling, dispatching, holding control, or fleet coordination involving operational objectives or explicit constraints. Relevant measures included travel time, passenger waiting time, schedule adherence, costs, reliability, and service quality. Multi-objective and constrained optimisation were treated as related but distinct formulations.
- ***Evaluation Evidence:*** Direct application evidence required simulation-based or empirical evaluation relevant to the reported operational or methodological findings. Conceptual formulations without public-transport evaluation were not treated as direct evidence of MARL effectiveness. Reporting completeness was considered separately from study eligibility.

Studies outside these criteria could inform the broader methodological synthesis but were not automatically treated as direct public-transport evidence. The following distinctions were applied:

- ***Non-Academic Materials:*** Extended abstracts, workshop summaries lacking substantive methodology, and non-peer-reviewed commercial or industry reports were excluded. Academic preprints were considered separately.
- ***Commercial Ride-Hailing Platforms:*** Findings from market-driven ride-hailing and ride-sourcing systems were treated as contextual rather than direct evidence of scheduled or service-constrained public transport [19,20].
- ***Freight Logistics and Last-Mile Delivery:*** Freight routing, truck scheduling, and parcel delivery studies could inform transferable optimisation or coordination concepts but not directly establish passenger-related service outcomes.
- ***Unmanned Aerial Vehicles (UAVs):*** UAV coordination research was considered methodologically relevant where appropriate, but differences in operating conditions,

movement constraints, and regulations prevented its findings from being treated as direct public-transport validation.

The 41 mapped studies were assigned six synthesis roles, as documented in Supplementary File S2: three direct public-transport applications, two transport-specific methodological or analytical support studies, 21 general MARL/MORL methodological contributions, eight adjacent transport applications, four deployment- or context-oriented studies, and three studies informing emerging methodological directions. These categories describe evidential roles rather than equivalent levels of application-specific validation. Study-level eligibility classifications and rationales are also provided in S2.

The distinction between direct and supporting evidence was maintained throughout the synthesis. Methodological studies informed discussions of architectures, reward formulations, and coordination mechanisms, while transport-support and contextual studies contributed operational and deployment perspectives without independently demonstrating MARL effectiveness in public transport. Foundational references, methodological surveys, and conventional spatio-temporal optimisation studies, including mathematical programming, time–space networks, and rolling-horizon methods, were cited where relevant but were not automatically included in the 41-study corpus unless they met the stated eligibility criteria. References published after the May 2026 search cut-off did not retrospectively alter the original search or selection counts.

Finally, the reporting-quality considerations discussed in Sections 2.3 and 2.6 were used to guide evidence interpretation rather than retrospectively determine study inclusion. No additional studies were included or excluded solely on the basis of these considerations.

2.3. Study Selection and Data Extraction

Following deduplication, records underwent two-stage screening: titles, abstracts, and keywords were examined for relevance, followed by full-text eligibility assessment using the criteria in Section 2.2. Studies were distinguished according to their direct public-transport evidence or supporting methodological, adjacent-domain, and deployment-related roles. Screening, eligibility assessment, and data extraction were conducted by a single author without independent verification. Automated procedures were limited to duplicate detection and preliminary document-type and metadata filtering. Study authors were not contacted for missing information, and potential selection and classification biases are acknowledged.

Data were extracted from the 41 mapped studies using variables aligned with the research questions, including publication characteristics, application domain, synthesis role, methodological approach, MARL architecture or coordination mechanism where applicable, optimisation objectives, identifiable objective-resolution strategies, evaluation setting, and reported contributions or findings. These characteristics were interpreted within each study's original application context. The study-level evidence mapping presented in Section 2.6 summarises the methodological contributions and research focus of the 41 studies, while Supplementary File S2 provides detailed characteristics, evidence-eligibility classifications, and corresponding rationales.

Evidence eligibility and synthesis role were treated as distinct attributes: the former identifies direct application-specific evidence, whereas the latter describes a study's methodological or contextual contribution. Classifications were based on reported information without assuming that all studies implemented comparable architectures, optimisation strategies, or coordination mechanisms. Unreported characteristics were not inferred, and no independent classification verification or inter-rater agreement assessment was undertaken.

The extracted information supported the descriptive analyses in Section 2.5 and thematic comparisons in Sections 3–5. Given substantial heterogeneity in application domains, algorithmic formulations, evaluation conditions, and performance metrics, statistical meta-

analysis was not undertaken. Reported methodological distributions and operational findings were interpreted descriptively rather than as pooled performance estimates or evidence of universal algorithmic superiority.

No formal risk-of-bias or methodological-quality appraisal was conducted. Instead, six reporting-quality dimensions were identified to guide the interpretation of methodological evidence: (i) **evaluation setting**, including benchmarks, simulation, operational data, or field deployment; (ii) **baseline adequacy**, concerning the availability and relevance of comparative methods; (iii) **statistical reporting**, including repeated runs, uncertainty estimates, or statistical tests; (iv) **robustness evaluation**, including ablation, sensitivity, or stress testing; (v) **reproducibility information**, including code, datasets, algorithmic parameters, and experimental settings; and (vi) **operational realism**, considering passenger demand, vehicle capacity, service schedules, and network constraints where applicable.

These six dimensions were used as qualitative interpretative considerations rather than a formal study-level quality appraisal, numerical scoring system, or retrospective eligibility criteria. They informed the interpretation of reported findings according to each study's evaluation context and the extent of methodological information available. Specifically, findings supported by limited or insufficiently described baselines were interpreted only in relation to the reported comparison methods, without inferring general algorithmic superiority. Where repeated runs, uncertainty estimates, or statistical tests were absent or insufficiently documented, reported performance improvements were treated as study-specific observations rather than evidence of statistical significance or consistent performance across repeated evaluations. Similarly, where robustness assessments such as ablation, sensitivity analysis, or stress testing were not reported, the stability of findings under alternative conditions was considered unestablished. Reproducibility information was used to contextualise the extent to which reported methods and experimental conditions could be independently examined, without treating reporting completeness as evidence of methodological validity.

Operational realism was considered particularly relevant when interpreting the three direct public-transport applications. Simulation-based findings were examined in relation to the reported passenger-demand assumptions, vehicle-capacity constraints, service schedules, and network conditions, where such information was available. Results obtained under simplified or incompletely described operational conditions were not interpreted as demonstrations of real-world deployment effectiveness or generalisability across transport systems. For general MARL/MORL methodological studies and adjacent-domain applications, these dimensions primarily informed the discussion of methodological relevance and potential transferability rather than direct public-transport effectiveness. No studies were excluded, reweighted, or assigned numerical quality rankings on the basis of these considerations, and no formal risk-of-bias or certainty-of-evidence conclusions were derived. The implications for study-level evidence mapping and operational findings are discussed in Sections 2.6 and 4.4, respectively, while the limitations of this interpretative approach are acknowledged in Section 6.4.

2.4. PRISMA Flow Diagram and Selection Results

The literature identification and selection process followed four reported phases: identification, screening, eligibility assessment, and inclusion (Figure 1). The process resulted in 41 studies retained for structured descriptive and comparative synthesis, encompassing direct public-transport applications and supporting methodological and contextual literature.

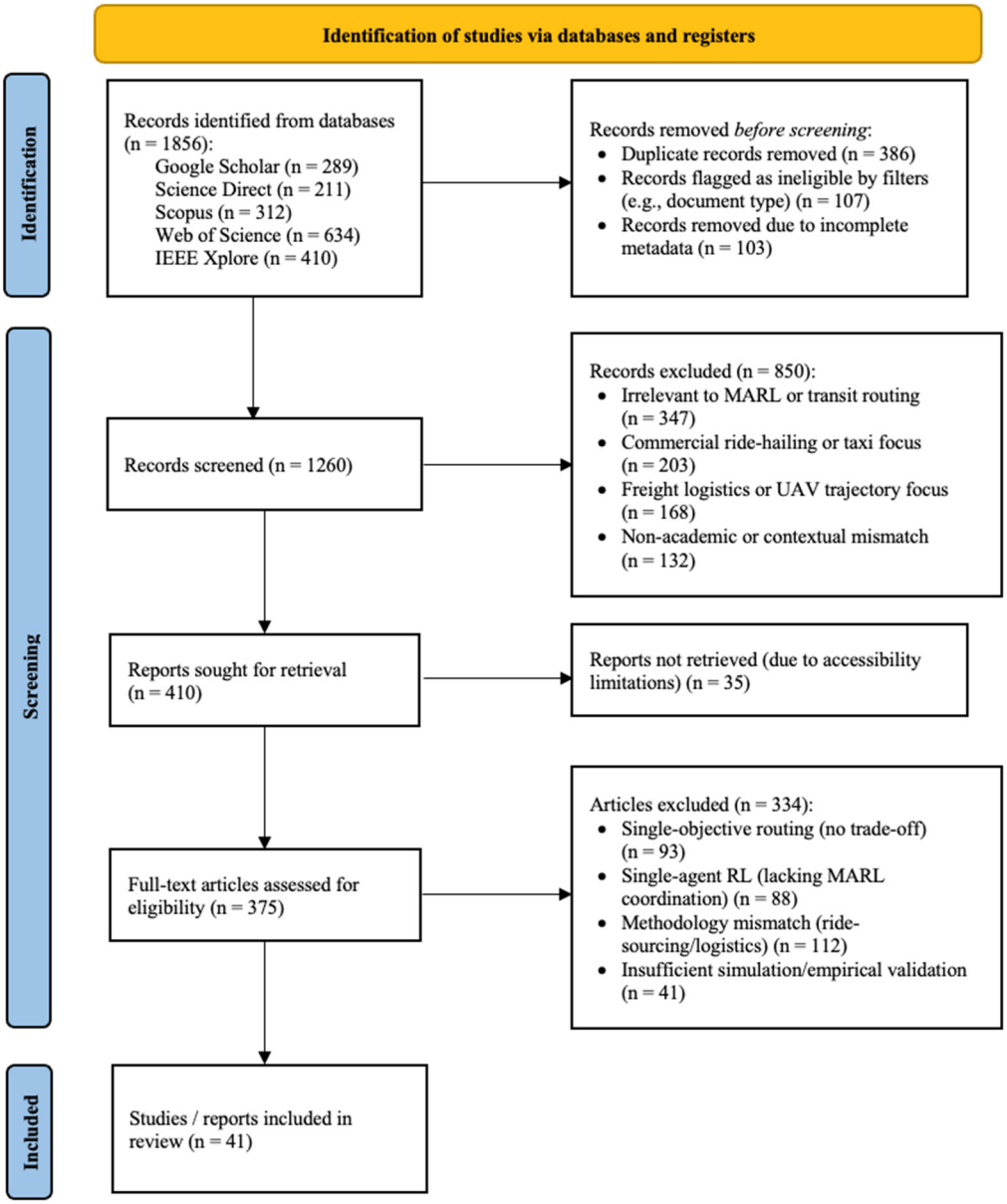


Figure 1. PRISMA flow diagram of the systematic literature review.

Across five academic databases and search platforms, 1856 records were identified. Before screening, 386 duplicates, 107 records flagged by preliminary document-type filters, and 103 records with incomplete metadata were removed, leaving 1260 records for title, abstract, and keyword screening. Of these, 850 were excluded: 347 for limited relevance to MARL or transport routing, 203 concerning commercial ride-hailing or taxi services, 168 focusing on freight logistics or UAV trajectories, and 132 for non-academic content or contextual mismatch.

Full-text retrieval was sought for 410 reports, of which 35 could not be retrieved, leaving 375 for eligibility assessment. A further 334 reports were excluded: 93 concerned single-objective routing without relevant trade-off analysis, 88 addressed single-agent reinforcement learning without MARL coordination, 112 had methodological or application-domain mismatches, and 41 lacked sufficient simulation-based or empirical evaluation. The remaining 41 studies were retained for synthesis.

As detailed in Supplementary File S2 and Section 2.6, the included studies comprise three direct public-transport applications, two transport-specific support studies, 21 gen-

eral MARL/MORL methodological contributions, eight adjacent transport applications, four deployment- or context-oriented studies, and three emerging-method studies. These categories describe different evidential roles rather than equivalent levels of public-transport-specific validation. The reporting-quality considerations discussed in Section 2.3 were used for evidence interpretation and did not retrospectively modify the reported selection counts.

2.5. Descriptive and Trend Analysis

A descriptive analysis of the 41 mapped studies examined publication trends, documented objective-resolution mechanisms, and Multi-Agent Reinforcement Learning (MARL) architectural characteristics. As detailed in Supplementary File S2, the corpus includes direct public-transport applications alongside general MARL/MORL methodological contributions and supporting studies from adjacent domains. The distributions describe the composition and methodological documentation of the selected literature rather than the prevalence of particular approaches across the wider research field. Study-level characteristics and eligibility rationales are presented in Supplementary File S2, with a summary provided in Section 2.6.

- ***Temporal distribution***

Figure 2 presents the reported publication distribution from 2018 to May 2026, distinguishing journal articles, conference proceedings, and academic preprints. The largest annual groups occur in 2023 (eight studies) and 2025 (nine studies), while 2018 and 2019 each contain three publications. The two studies recorded in 2026 reflect the May search cut-off and should not be interpreted as evidence of declining research activity. As the corpus includes methodological and adjacent-domain literature, these counts do not represent the annual prevalence of MARL applications directly evaluated in public transport.

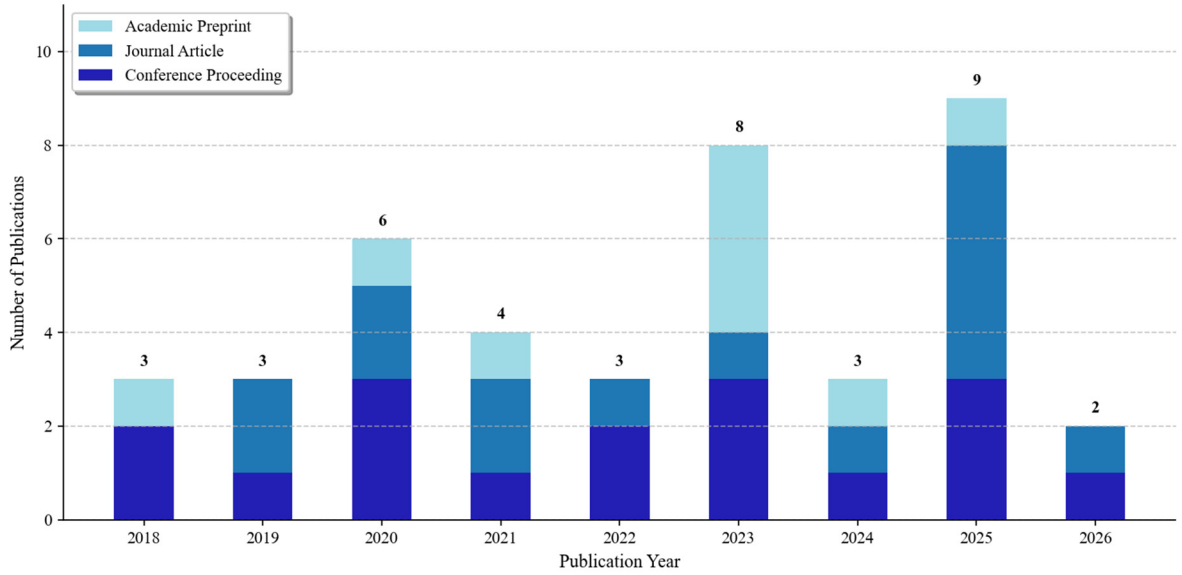


Figure 2. Reported temporal distribution of the 41 studies considered in the structured synthesis (2018–May 2026), categorised by publication format.

- ***Methodological landscape***

Multi-objective MARL concerns decision-making involving multiple, potentially competing objectives, whereas linear scalarisation, constraint-based optimisation, and Pareto-based optimisation represent mechanisms for resolving objective trade-offs. These mechanisms are not necessarily mutually exclusive and may coexist within individual formulations.

Figure 3a summarises the objective-resolution mechanisms identifiable from the methodological descriptions recorded in Supplementary File S2. Among the 41 mapped studies, linear scalarisation is explicitly identified in one study, Pareto-based optimisation in five studies, and constraint-based optimisation in one study. For 23 studies, the available descriptions do not sufficiently specify an objective-resolution mechanism, while 11 studies are classified as not applicable to this categorisation. These classifications reflect the information documented in S2 rather than independently verified methodological implementations. Consequently, the reported counts should not be interpreted as the prevalence of these optimisation strategies across the research field or as evidence of comparative algorithmic effectiveness.

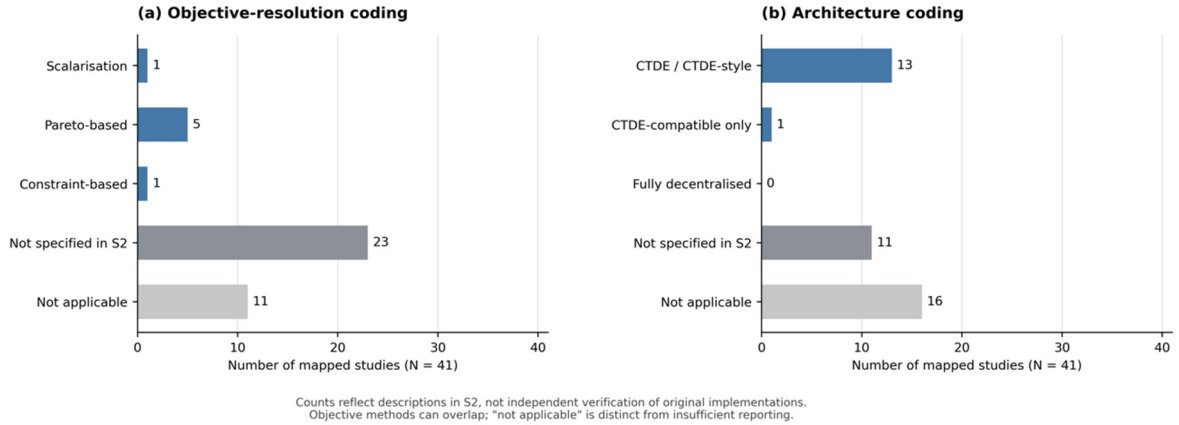


Figure 3. Distribution of documented objective-resolution mechanisms and MARL architectural characteristics across the 41 mapped studies. Note: Counts are derived from the methodological descriptions recorded in Supplementary File S2. Studies without sufficient classification information are distinguished from those for which the classification is not applicable. Objective-resolution mechanisms may overlap. CTDE-compatible approaches are reported separately from explicitly documented CTDE configurations. These descriptive findings do not establish comparative algorithmic effectiveness or application-specific validation in public transport.

- ***Architectural trends***

Figure 3b presents the MARL architectural characteristics documented in Supplementary File S2. CTDE or CTDE-style configurations are identified in 13 studies, while one additional study is described as CTDE-compatible without an explicitly established CTDE implementation. No study is explicitly classified as fully decentralised under the applied coding criteria. Architectural information is insufficiently specified for 11 studies, and the classification is not applicable to 16 studies.

Under CTDE, agents may use broader state or joint-action information during training while executing decentralised policies based on local observations. Fully decentralised approaches instead rely primarily on agent-level information, although their coordination mechanisms depend on the specific algorithm and application setting. The reported counts describe documented methodological characteristics rather than comparative architectural effectiveness. In particular, the absence of explicitly classified fully decentralised studies should not be interpreted as evidence that such approaches are absent from the broader literature.

Overall, Figures 2 and 3 provide descriptive summaries of publication patterns and methodological characteristics within the mapped corpus. Figure 3 additionally highlights limitations in the specificity and applicability of the available methodological descriptions. Section 2.6 complements these distributions through a study-level evidence mapping, while

Supplementary File S2 provides the corresponding study characteristics, evidence roles, and eligibility rationales.

2.6. Study-Level Evidence Mapping

This section presents the study-level mapping of the 41 articles retained for structured synthesis. Table 2 summarises their application domains, methodological approaches, evidence roles, and reported contributions, while Supplementary File S2 provides detailed characteristics, evaluation settings, evidence-eligibility classifications, and corresponding rationales.

The mapped literature comprises six synthesis categories: three direct public-transport applications, two transport-specific methodological or analytical support studies, 21 general MARL/MORL methodological contributions, eight adjacent transport applications, four deployment- or context-oriented studies, and three studies informing emerging methodological directions. These categories represent distinct evidential roles rather than equivalent levels of application-specific validation.

The three direct applications investigate bus holding, multi-objective bus coordination, and metro timetable rescheduling within simulation-based settings. Their operational and passenger-related findings are examined in Section 4.4, including distinctions among waiting time, travel time, passenger delay, headway regularity, and service reliability. The remaining studies inform methodological, operational, or contextual discussions without independently establishing MARL effectiveness in public or tourist transport fleets.

As described in Section 2.3, evidence eligibility was distinguished from synthesis role, while six reporting-quality dimensions informed the qualitative interpretation of reported methodological evidence. For direct public-transport applications, the relevance of comparison baselines, availability of statistical and robustness information, and realism of operational assumptions were considered when interpreting reported outcomes. Findings based on limited or insufficiently described comparison methods were restricted to the evaluated baselines rather than generalised as evidence of algorithmic superiority. Where statistical uncertainty, repeated evaluations, or robustness analyses were not adequately documented, the synthesis did not infer statistical significance, consistent performance, or stability under alternative conditions. Likewise, simulation-based findings were interpreted within their reported passenger-demand, vehicle-capacity, scheduling, and network assumptions rather than as evidence of validated real-world deployment.

For general MARL/MORL methodological contributions and adjacent-domain studies, the reporting-quality considerations primarily informed the discussion of methodological relevance and potential transferability. Results from general cooperative-learning benchmarks, traffic signal control, autonomous driving, or other adjacent settings were not treated as direct evidence of passenger-related benefits in public-transport fleets. Differences in application domains, algorithmic formulations, evaluation settings, and available reporting information further limited direct cross-study comparisons.

These considerations did not constitute a formal methodological-quality or risk-of-bias assessment, and no numerical quality scores, study weights, or comparative study rankings were assigned. The methodological distributions in Figure 3 and the conceptual coordination mechanisms discussed in Section 4 should therefore be interpreted as descriptive and conceptual representations, respectively, rather than evidence of comparative algorithmic effectiveness.

Table 2. Characteristics and contributions of the 41 studies included in the structured synthesis.

Study	Synthesis Role	Application Domain	Method/Approach	Contribution/Reported Focus
Foerster et al. [21]	Methodological	Cooperative MARL	COMA	Counterfactual credit assignment for cooperative multi-agent policy learning.
Song et al. [22]	Methodological	Multi-agent imitation learning	Multi-agent GAIL	Imitation learning from expert behaviour in multi-agent benchmark environments.
Tessler et al. [23]	Methodological	Constrained RL	Reward-constrained policy optimisation	Constraint-based handling of reward and operational requirements.
Chu et al. [6]	Adjacent transport	Urban traffic signals	Multi-agent deep RL	Coordinated traffic signal control and scalability in traffic simulations.
Das et al. [24]	Methodological	Multi-agent communication	TarMAC	Differentiable targeted communication between cooperative learning agents.
Wei et al. [25]	Adjacent transport	Network traffic signals	CoLight	Graph-attention-based coordination among traffic signal controllers.
Christianos et al. [26]	Methodological	Cooperative MARL	Selective parameter sharing	Parameter-sharing mechanisms for scalable multi-agent coordination.
Duan et al. [27]	Adjacent transport	Autonomous driving	Hierarchical RL	Hierarchical decomposition of strategic and operational driving decisions.
Lu et al. [28]	Adjacent transport	Vehicle decision-making	Hierarchical RL	Separation of long-horizon decisions and short-horizon vehicle control.

Table 2. *Cont.*

Study	Synthesis Role	Application Domain	Method/Approach	Contribution/Reported Focus
Rashid et al. [29]	Methodological	Cooperative MARL	Weighted QMIX	Weighted value-factorisation approach for cooperative multi-agent learning.
Rashid et al. [30]	Methodological	Cooperative MARL	QMIX	Monotonic value factorisation for cooperative action-value learning.
Wang & Sun [31]	Direct application	Fixed-route bus operations	Multi-agent deep RL	Simulation-based coordinated bus holding addressing headway regularity and delays.
Chen et al. [32]	Adjacent transport	Adaptive network routing	Multi-agent meta-RL	Adaptive routing under changing network conditions.
Rudskoy et al. [33]	Deployment support	Intelligent transportation systems	Digital Twin framework	Digital-twin integration and simulation-to-real-world considerations for ITS.
Suo et al. [34]	Deployment support	Urban traffic simulation	TrafficSim	Traffic-behaviour simulation for representing realistic and stochastic conditions.
Wu et al. [35]	Deployment support	Autonomous driving	Human-in-the-loop DRL	Incorporation of human feedback and supervisory interaction in learning-based control.
Fawaz et al. [36]	Methodological	Collaborative queuing	Graph convolutional RL	Graph-based multi-agent coordination for queuing-related objectives.
Pan et al. [37]	Methodological	Offline cooperative MARL	Offline MARL with actor rectification	Conservative offline policy learning under distribution shift.
Yu et al. [38]	Methodological	Cooperative MARL	MAPPO	Policy-gradient-based cooperative multi-agent learning in benchmark environments.

Table 2. *Cont.*

Study	Synthesis Role	Application Domain	Method/Approach	Contribution/Reported Focus
Bukharin et al. [39]	Methodological	Robust cooperative MARL	Adversarially regularised MARL	Adversarial robustness and performance considerations in multi-agent learning.
Cai et al. [40]	Methodological	Multi-objective RL	Distributional Pareto-optimal MORL	Pareto-based representation of competing objectives.
Fu et al. [41]	Adjacent transport	Traffic control	Multi-agent RL	Coordinated traffic control addressing efficiency and agent interactions.
Kwon et al. [42]	Emerging method	Reward design	Language-model-assisted reward design	Language-model-based support for reward specification.
Ma et al. [43]	Emerging method	Automated reward engineering	Eureka/LLM reward design	Automated reward-function generation in robotics and RL benchmarks.
Rocamonde et al. [44]	Emerging method	Reward modelling	Vision-language reward models	Vision-language-based reward representation and evaluation.
Wang & Sun [45]	Direct application	Multiline public-bus operations	Multi-objective multi-agent deep RL	Simulation-based coordinated bus control considering bunching, service performance, and shared-corridor constraints.
Zhou et al. [46]	Methodological	MARL architecture	CTDE analysis	Analysis of centralised training and decentralised execution in MARL.
Amato [47]	Methodological	Cooperative MARL	CTDE overview	Methodological discussion of coordination under partial observability.
Liu et al. [48]	Methodological	Graph-based MARL	Graph neural network MARL	Graph-based multi-agent information sharing and coordination.

Table 2. *Cont.*

Study	Synthesis Role	Application Domain	Method/Approach	Contribution/Reported Focus
Palma et al. [49]	Transport support	Public-transport route recommendation	POI and transport recommendation	Route recommendation incorporating POI accessibility and route-quality considerations.
Fang et al. [50]	Methodological	Sparse-reward MARL	Self-satisfaction MARL	Reward shaping and agent-satisfaction considerations in multi-agent learning.
Le & Ta [51]	Methodological	Multi-objective MARL	Strong Pareto-optimal policy learning	Pareto-oriented policy learning for multi-objective decision-making.
Li & He [52]	Methodological	Continuous-control MARL	MAPPO	Cooperative multi-agent policy learning in continuous-control benchmarks.
Liu et al. [53]	Transport support	Transit-system analytics	GTFS spatiotemporal network analysis	Operational transit-network representation and data-derived metrics.
Liu et al. [54]	Methodological	Multi-objective RL	Pareto set learning	Learning and representing Pareto solution sets in MORL.
Liu et al. [55]	Methodological	Multi-objective RL	Pareto-front discovery	Computational approaches for identifying Pareto trade-off solutions.
Mu et al. [56]	Methodological	Preference-based MORL	Preference-conditioned MORL	Preference-aware learning for multi-objective decision-making.
Sun et al. [57]	Adjacent transport	Vehicle eco-driving	Deep RL	Energy, comfort, and safety trade-offs in vehicle-level control.
Zhang et al. [58]	Deployment support	Ride-comfort evaluation	Intelligent comfort evaluation	Passenger-comfort assessment using vehicle vibration information.

Table 2. *Cont.*

Study	Synthesis Role	Application Domain	Method/Approach	Contribution/Reported Focus
Azari et al. [59]	Adjacent transport	Urban destination choice	Spatial agent-based modelling	Modelling spatial destination-choice behaviour and interactions.
Jufen et al. [60]	Direct application	Urban metro timetable rescheduling	Multi-agent deep RL/MAA2C	Simulation-based timetable rescheduling considering passenger delay and operational scheduling constraints.

Note: Evidence roles are abbreviated from the six synthesis categories documented in Supplementary File S2. "Direct application" refers to studies evaluating MARL within public-transport operations, whereas other categories provide supporting methodological, transport-related, deployment, or emerging-method insights. The contributions describe the original methodological or application focus of each study and do not imply that all approaches have been evaluated in public or tourist transport fleets. No pooled effect estimates or comparative algorithmic rankings are inferred from this table.

Overall, Table 2 provides the study-level basis for the subsequent synthesis of MARL architectures, multi-objective resolution strategies, coordination mechanisms, operational outcomes, and deployment challenges. The associated evidence-eligibility rationales are documented in Supplementary File S2, while limitations concerning evidence heterogeneity, single-author screening, methodological appraisal, and cross-domain transferability are discussed in Section 6.4.

3. Formal Problem Formulation: Multi-Objective Markov Games

This section synthesises a Multi-Objective Partially Observable Stochastic Game (MO-POSG) formulation to provide a common conceptual framework for examining multi-agent interactions and competing operational and passenger-related objectives in public transport [61,62]. The formulation organises methodological insights from the reviewed literature rather than introducing an experimentally validated controller or common benchmark for algorithmic comparison.

3.1. System Definition

Public transport coordination can be represented as an MO-POSG involving interacting agents, stochastic transitions, partial observability, and competing objectives. Centralised Training and Decentralised Execution (CTDE) provides a relevant learning paradigm in which broader state or joint-action information may be available during training, while agents execute decentralised policies using local observations [46,47]. Parameter sharing among homogeneous agents may also reduce model duplication, depending on the learning algorithm and task [26].

The system operates synchronously and is defined by the tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, O, \mathcal{Z}, \mathcal{P}, R, \gamma \rangle$, where:

- $\mathcal{N} = \{1, \ldots, n\}$ is the set of decision-making agents, representing vehicles or local transport controllers where appropriate [6,63].
- $\mathcal{S}$ is the global state space, potentially including network connectivity, congestion, time-dependent demand, vehicle locations, schedules, and operational constraints. Spatial information may be represented through networks, grids, or other application-specific structures [59,64].
- $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_n$ is the joint action space, where each agent selects $a_i \in \mathcal{A}_i$. Actions may be discrete, such as route selection, or continuous, such as dwell-time adjustment [31,65].
- $O = O_1 \times \ldots \times O_n$ is the joint observation space. Due to partial observability, the agent i cannot access the full state $s \in \mathcal{S}$ but receives a localised observation $o_i \in O_i$ determined by the observation function $\mathcal{Z} : \mathcal{S} \times \mathcal{A} \to \Delta O$.
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition probability function, governing the stochastic evolution of the traffic environment based on the joint actions executed by all agents.
- $\boldsymbol{R} = \{\boldsymbol{r_1}, \ldots, \boldsymbol{r_n}\}$ is the set of vector-valued reward functions. For each agent i, the reward function $r_i : \mathcal{S} \times \mathcal{A} \to R^m$ provides an m-dimensional feedback signal, quantifying the trade-offs between conflicting systemic goals [66].
- $\gamma \in [0, 1)$ is the discount factor for future rewards.

For discrete action spaces, the number of joint actions is $|\mathcal{A}| = \prod_{i=1}^{n} |\mathcal{A}_i|$, potentially growing exponentially with fleet size when individual action-set sizes remain fixed [67]. This can increase exploration and optimisation complexity, although CTDE does not necessarily require explicit enumeration of all joint actions. Factorised policies and hierarchical decision structures may alleviate some scalability challenges, as discussed in Section 4.1.

The formulation accommodates spatio-temporal transport characteristics through state variables, transition dynamics, feasible actions, and operational constraints. Network

restrictions, vehicle capacity, service schedules, and time windows may be represented through action restrictions, transition rules, or explicit constraints. Reward penalties alone do not guarantee feasibility. Thus, MO-POSG provides a complementary representation of sequential multi-agent decisions rather than a replacement for mixed-integer optimisation, time–space networks, or rolling-horizon approaches.

The multi-agent system seeks to learn a joint policy $\pi = \langle \pi_1, \ldots, \pi_n \rangle$, where $\pi_i(a_i \mid o_i)$ denotes agent (i)'s decentralised policy. The expected discounted vector return is defined as:

$$V^{\pi} = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \in R^m \quad (1)$$

Because V^{π} is vector-valued, competing objectives require an explicit optimisation criterion. Depending on the formulation, the goal may involve approximating a Pareto-optimal policy set or maximising a specified scalarisation function $\Phi : R^m \rightarrow R$ [54,55]. These alternatives do not inherently establish the superiority of any learning architecture.

Figure 4 illustrates the synthesised MO-POSG formulation under CTDE, connecting global states, local observations, decentralised policies, centralised training information, joint actions, and vector-valued rewards. It is a conceptual methodological representation rather than an implemented or experimentally validated public-transport controller.

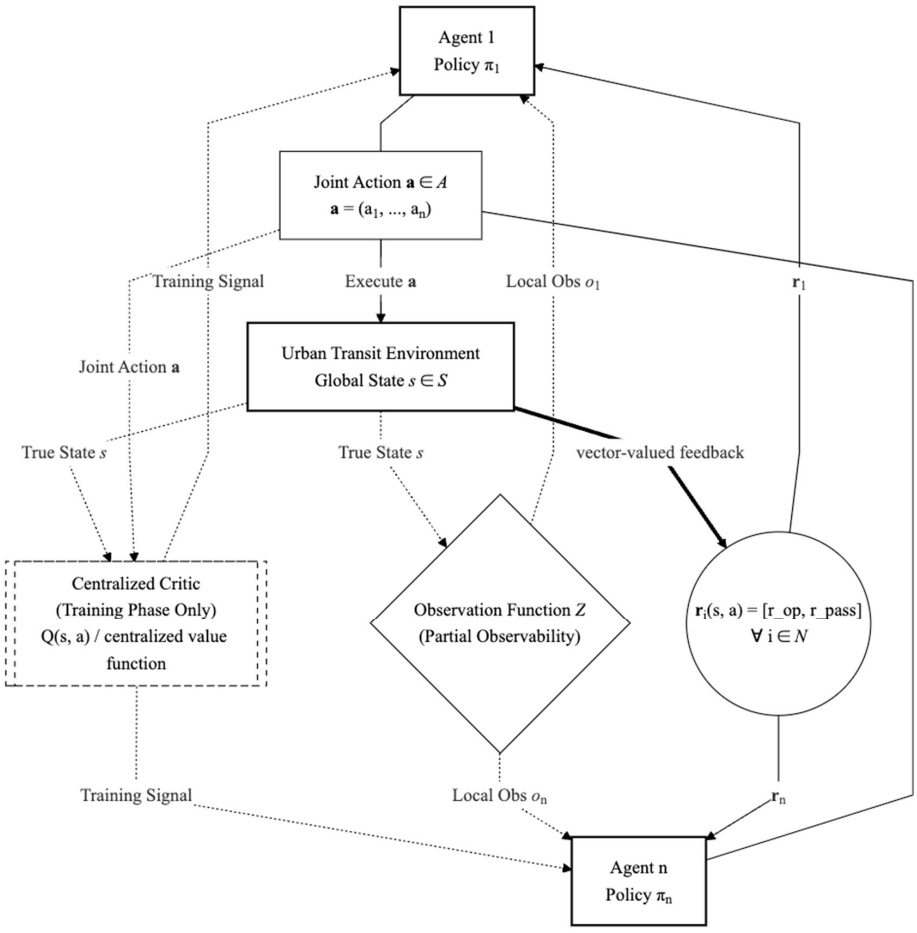


Figure 4. Conceptual representation of the synthesised Multi-Objective Partially Observable Stochastic Game (MO-POSG) formulation under the Centralised Training and Decentralised Execution (CTDE) paradigm. Note: The diagram illustrates the conceptual relationships among global states, local observations, decentralised policies, centralised training information, joint actions, and vector-valued rewards. It is a methodological synthesis rather than an implemented or experimentally validated MARL architecture in the present review.

*3.2. Vector Reward Design (**R**)*

Reward design is central to representing competing objectives in Multi-Agent Reinforcement Learning (MARL). Public and tourist transport operations involve operational efficiency and passenger-related considerations that may not be adequately represented by a single performance metric [14,16]. Within the conceptual MO-POSG formulation introduced in Section 3.1, each agent receives an (m)-dimensional reward vector $r_i(s,a) \in R^m$, which can represent distinct objectives without requiring their immediate aggregation into a scalar reward.

For illustration, the reward vector is decomposed into two components:

$$r_i(s,a) = \begin{bmatrix} r_i^{operational}(s,a) \\ r_i^{passenger}(s,a) \end{bmatrix} \quad (2)$$

- ***Operational Reward*** **($r_i^{operational}$):** Represents operational efficiency through travel duration and delay-related measures [64]:

$$r_i^{operational} = -(\alpha \cdot \text{travel_time} + \beta \cdot \text{delay}) \quad (3)$$

where α and β are weighting coefficients, while travel time and delay are operational indicators whose definitions depend on the application. Relevant information may be obtained from simulation outputs (e.g., SUMO) or transit data sources such as GTFS [53,68].

- ***Human-Centric Reward*** **($r_i^{passenger}$):** Represents passenger-related considerations, particularly scenic quality and travel comfort in tourist transport [50]:

$$r_i^{passenger} = \lambda \cdot \text{scenic_score} + \mu \cdot \text{comfort_index} \quad (4)$$

where λ and μ represent the relative importance of scenic quality and comfort. These attributes require measurable proxies: scenic quality may be approximated through Point-of-Interest (POI) accessibility, landmark exposure, or green-space proximity using spatial analysis tools such as OSMnx [49,69]. Comfort may be represented through vehicle-dynamics indicators, including harsh braking and acceleration [58,70].

Operational and passenger-related objectives may conflict. For example, a shorter route may reduce travel time but offer limited scenic value, whereas a scenic detour may increase travel duration or network exposure [71–74]. A vector-valued formulation preserves these distinctions for subsequent multi-objective analysis, although learning Pareto-optimal policies requires an appropriate resolution mechanism [66].

Because the reward components may differ substantially in scale and units, normalisation or standardisation can help balance their numerical influence during learning. Methods such as min–max scaling or z-score standardisation may be applied where appropriate, but their suitability depends on the reward distributions and learning formulation [75]. Table 3 summarises illustrative operational proxies, data sources, and normalisation approaches drawn from relevant literature.

The vector-valued formulation separates competing objectives conceptually; their combination, prioritisation, or constrained treatment depends on the objective-resolution strategy discussed in Section 3.3.

Table 3. Operationalisation of vector reward components in MARL-based public transport systems.

Reward Component	Sub-Objective/Specific Metrics	Operational Proxy & Data Source	Normalisation Strategy	Representative Literature
Operational Efficiency ($r_{operational}$)	- Minimise travel time - Minimise schedule delay (headway deviation)	- Real-time/Historical traffic simulators (e.g., SUMO) - Dynamic GTFS (General Transit Feed Specification) feeds	Min-max scaling or Z-score standardisation to align continuous time domains.	[53,64,68]
Human-Centric/Passenger Experience ($r_{passenger}$)	- Maximise Scenic score (POI density, green-space exposure) - Maximise Comfort index (smooth kinematics)	- Spatial analysis libraries (e.g., OSMnx) to evaluate route edges - Vehicle dynamics data (penalising hard braking/aggressive acceleration)	Min-max scaling or Z-score standardisation to convert arbitrary indices to uniform scales.	[49,50,58,69,70]

3.3. Multi-Objective Resolution Strategies

Multi-objective Multi-Agent Reinforcement Learning (MARL) concerns multi-agent decision-making with multiple, potentially competing objectives. This problem setting is distinct from the mechanisms used to resolve objective trade-offs. Linear scalarisation combines reward components using specified weights, constraint-based optimisation imposes explicit requirements while optimising selected objectives, and Pareto-based optimisation seeks non-dominated solutions. These mechanisms are not mutually exclusive and should not be interpreted as distinct or exhaustive families of MARL algorithms.

- ***Linear Scalarisation:*** Weighted-sum scalarisation combines vector-valued rewards into a scalar optimisation signal [16]. For the two-component reward formulation in Section 3.2, the scalarised reward is:

$$R_i(s,a) = w_1 \cdot r_i^{operational} + w_2 \cdot r_i^{passenger} \tag{5}$$

where $w_1 + w_2 = 1$ and $w_j \geq 0$ [57]. This approach provides a straightforward mechanism for incorporating specified objective preferences. However, fixed weights may not reflect changing operational priorities, and weighted-sum optimisation may fail to recover unsupported Pareto-optimal solutions in non-convex objective spaces [66,76]. These limitations concern scalarisation and should not be attributed automatically to the underlying MARL architecture.

- ***Constraint-Based MARL:*** Constrained formulations optimise a selected objective subject to explicit requirements, drawing on constrained Markov decision processes and related multi-agent formulations [77,78]. For example, operational efficiency may be maximised while maintaining a minimum passenger-service reward [45]:

$$\max_{\pi} E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_i^{operational}(s_t, a_t)\right] \tag{6}$$

$$\text{subject to } E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_i^{passenger}(s_t, a_t)\right] \geq \tau \tag{7}$$

where τ specifies a minimum acceptable discounted passenger-service reward. Such formulations may represent service-level requirements, operational restrictions, or other constraints relevant to transport systems [79,80]. However, feasibility and constraint satisfaction depend on the formulation and solution method. Restrictive constraints may reduce the feasible policy set, while Lagrangian optimisation can introduce training instability; PID-based dual updates have been investigated as a potential stabilisation mechanism [23,81,82].

These methodological developments do not independently establish reliable constraint satisfaction in public-transport deployment.

- ***Pareto-Based MARL:*** Pareto-based approaches seek to identify or approximate sets of non-dominated policies representing alternative trade-offs among competing objectives [83,84]. Unlike optimisation with a single fixed weight vector, these approaches can support exploration of multiple objective preferences, subject to the representation and learning method. Nevertheless, approximating diverse Pareto-optimal policies may entail computational and sample-efficiency challenges, and the resulting coverage depends on the algorithm and evaluation setting. General methodological findings do not establish superior operational performance in public-transport fleets.

Figure 5 conceptually illustrates trade-offs between operational efficiency and passenger-related service quality. Under suitable assumptions, fixed weighted-sum scalarisation may fail to recover unsupported Pareto-optimal solutions in non-convex objective spaces [76]. Constraint-based formulations specify acceptable performance thresholds, while Pareto-based approaches seek alternative non-dominated solutions that may support preference-dependent decisions [56]. However, their practical suitability depends on reward representation, computational requirements, and the underlying optimisation problem. The illustrated frontier and feasible region are hypothetical and do not represent empirical findings or demonstrate superior algorithmic performance in public transport.

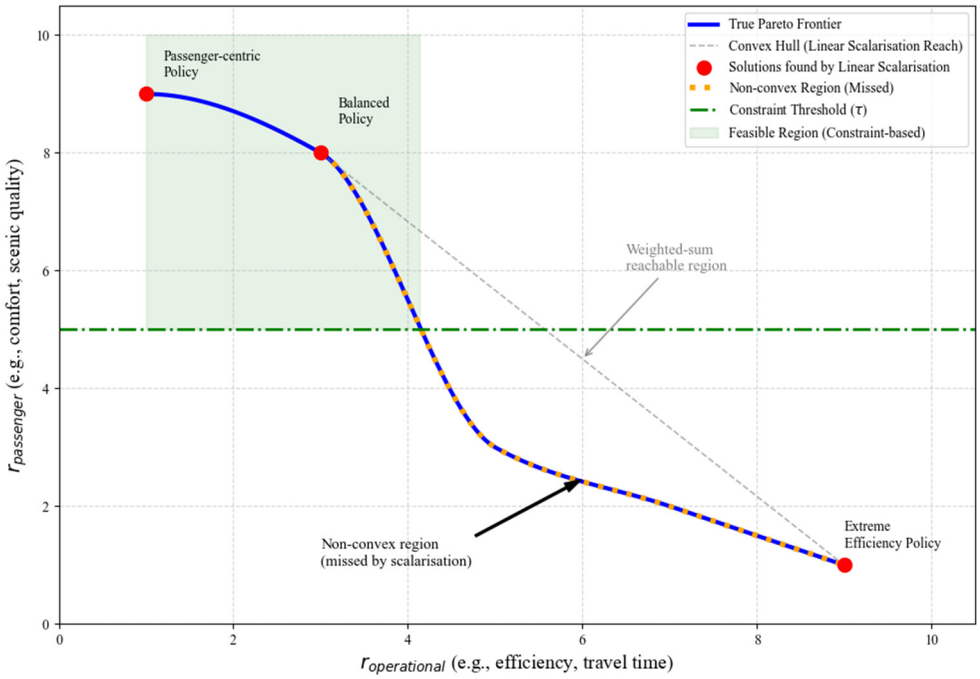


Figure 5. Conceptual illustration of multi-objective trade-offs between operational efficiency and passenger-related service quality. Note: The illustrated feasible region, Pareto boundary, preference directions, and constraint threshold are hypothetical and are not derived from numerical results or Pareto fronts reported by the 41 mapped studies. The diagram illustrates methodological distinctions between weighted-sum scalarisation, constraint-based optimisation, and Pareto-based optimisation rather than empirical algorithmic performance.

Table 4 summarises the mathematical formulations, potential advantages, limitations, and public-transport relevance of the three objective-resolution strategies.

Table 4. Comparison of Objective-Resolution Strategies Relevant to Multi-Objective MARL.

Resolution Strategy	Mathematical Approach	Potential Advantage	Main Limitation	Potential Public Transport Relevance
Linear Scalarisation [16,57]	$R_i = w_1 r^{op} + w_2 r^{pass}$	Straightforward optimisation using explicit preference weights	Fixed weighted sums may not recover unsupported Pareto-optimal solutions	Routing and service control with specified objective preferences
Constraint-Based Optimisation [45,77]	$\max(\text{Efficiency}) s.t. (\text{Satisfaction}) \geq \tau.$	Represents service or operational requirements as explicit constraints	Feasibility and constraint satisfaction depend on the formulation and solution method	Scheduling, fleet control and service-level restrictions
Pareto-Based Optimisation [55,84]	Identification or approximation of non-dominated solutions	Supports exploration of multiple trade-offs without a single fixed weighting	Potential computational and sample complexity; incomplete Pareto coverage	Comparing passenger-service and operational-efficiency trade-offs

Note: These strategies are not mutually exclusive algorithmic families. Multi-objective MARL defines the broader decision-making setting, while scalarisation, explicit constraints, and Pareto-based methods represent alternative or complementary mechanisms for handling competing objectives. The table provides a methodological comparison rather than an empirical ranking of algorithmic performance.

4. MARL Architectures and Algorithmic Insights

Building on the MO-POSG formulation in Section 3 and the study-level mapping in Section 2.6, this section synthesises methodological insights from the 41 reviewed studies. Their application domains, methods, and contributions are summarised in Table 2 and Supplementary File S2.

Section 4.1 compares value-based, policy-based, hierarchical, and offline MARL approaches; Section 4.2 examines explicit and implicit coordination mechanisms; Section 4.3 discusses scalability, non-stationarity, and multi-objective learning challenges; and Section 4.4 evaluates reported operational and passenger-related outcomes. Throughout, direct public-transport findings are distinguished from supporting methodological evidence. Reported benefits are interpreted according to the original evaluation settings, baselines, and performance measures rather than as universal algorithmic superiority or equivalent improvements across different passenger-related outcomes.

4.1. Comparative Analysis of Learning Paradigms

Centralised Training and Decentralised Execution (CTDE) is widely discussed in MARL methodology, although direct validation in public-transport fleets remains limited [9,46,47]. Under CTDE, agents may access broader information during training while executing decentralised policies using local observations. Value-based and policy-based approaches differ in their learning mechanisms, but their multi-objective capabilities depend on the complete reward formulation, algorithmic implementation, and evaluation setting. Connections between value-based and policy-based reinforcement learning have also been investigated through path-consistency learning, which relates value estimation and policy optimisation under entropy regularisation [85]. The following comparison draws primarily on methodological studies documented in Table 2 and Supplementary File S2 rather than assuming equivalent effectiveness in public transport.

- ***Value-Based Factorisation (VDN and QMIX):*** Value-Decomposition Networks (VDN) and QMIX represent cooperative joint action values through agent-level utility functions [18,29]. VDN uses additive decomposition:

$$Q_{tot}(\tau, a) = \sum_{i=1}^{n} Q_i(\tau_i, a_i) \tag{8}$$

This structure supports decentralised greedy action selection but restricts the representation of non-additive interactions among agents. Such limitations may be relevant to tightly coupled transport decisions involving shared infrastructure or congestion.

QMIX extends value factorisation through a nonlinear mixing network subject to a monotonicity constraint [30]:

$$\frac{\partial Q_{tot}}{\partial Q_i} \geq 0, \forall i \in \{1, \ldots, n\} \tag{9}$$

Monotonicity supports consistency between decentralised greedy actions and the joint action-value representation but restricts the class of representable joint value functions. Importantly, this representational limitation is distinct from the inability of fixed weighted-sum scalarisation to recover unsupported Pareto-optimal solutions in non-convex objective spaces [66,76]. Standard VDN and QMIX implementations typically optimise scalar rewards, while their multi-objective capabilities depend on reward representation, preference conditioning, and algorithmic extensions rather than value factorisation alone.

- ***Policy-Based Approaches (MAPPO):*** Multi-Agent Proximal Policy Optimisation (MAPPO) employs policy-gradient learning, commonly under CTDE, without explicitly factorising the joint action-value function as in VDN or QMIX [38,86]. Its clipped surrogate objective is:

$$L(\theta) = E_t\left[\min\left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{10}$$

Here, $r_t(\theta)$ is the probability ratio between updated and behaviour policies, and $\hat{A}_t$ denotes the estimated advantage. A centralised critic may use broader training information, potentially including network conditions, vehicle interactions, or passenger demand in transport-oriented formulations.

Unlike QMIX, MAPPO does not impose a monotonic value-factorisation constraint. Other multi-agent policy-gradient methods, such as off-policy decomposed policy gradients, integrate value-function decomposition into actor–critic learning to address credit assignment and coordination challenges [87]. However, this distinction does not establish superior multi-objective capabilities. Standard MAPPO generally optimises scalar feedback and requires additional reward formulation or learning mechanisms to represent preferences, constraints, or Pareto trade-offs. Its performance also depends on sample requirements, reward specification, training stability, and experimental conditions. General MARL benchmarks therefore do not establish superior passenger comfort or operational effectiveness in public transport [38,51].

- ***Hierarchical Reinforcement Learning (HRL):*** HRL decomposes sequential decision-making into temporally extended sub-tasks or policies operating at different abstraction levels [88,89]. Unlike the value-based versus policy-based distinction, hierarchy concerns decision organisation and may be combined with different learning algorithms.

In transport applications, high-level decisions could concern route planning or service allocation, while lower-level controllers manage vehicle movements or local operations [27,28]. Such decomposition may separate strategic and operational timescales, but its effectiveness depends on hierarchy design, reward structures, and inter-level coordination.

HRL does not inherently resolve multi-objective trade-offs or guarantee improved learning or operational performance.

- ***Offline MARL and Conservative Policy Rectification:*** Offline MARL learns policies from previously collected interaction data, reducing the need for exploratory interaction with live operational systems during training [90]. This is relevant to safety-sensitive transport settings where online exploration may disrupt service reliability or passenger safety [91].

Potential data sources include dispatcher records, vehicle trajectories, passenger-demand observations, and historical operations. However, offline learning requires sufficiently informative state, action, reward, and transition data. Incomplete coverage, distributional shift, and extrapolation errors remain important challenges, particularly when agents select actions poorly represented in the training dataset [92].

Pan et al. [37] proposed Offline Multi-Agent Reinforcement Learning with Actor Rectification (OMAR), combining conservative value estimation and actor rectification to support offline cooperative policy optimisation. This provides a methodological example of limiting unsupported policy updates rather than direct evidence of operational safety in public transport. Deployment would require application-specific validation, constraint handling, and human oversight.

- ***Trade-off Performance Synthesis:*** The reviewed approaches differ in their representational assumptions, learning mechanisms, and implementation requirements. VDN and QMIX employ additive or monotonic value factorisation, whereas MAPPO avoids these particular constraints but introduces other considerations involving reward design, sample requirements, and optimisation stability. HRL structures decisions across temporal scales, while offline MARL addresses the availability and use of historical interaction data.

These distinctions do not establish universal rankings of multi-objective capability. Performance depends on the complete formulation, including reward representation, scalarisation, preference modelling, operational constraints, and evaluation conditions. Accordingly, Table 5 provides a qualitative methodological comparison rather than a common experimental benchmark or evidence of superior public-transport performance.

4.2. Coordination Mechanisms and Spatial Conflict Resolution

Coordination among public transport agents may support decision-making across shared road networks, intersections, and passenger destinations. Under partial observability, agents relying solely on local information may struggle to anticipate neighbouring vehicles' actions, potentially contributing to inefficient routing or localised congestion [9,67,71,72].

MARL research has investigated coordination mechanisms involving differentiable communication, graph-based information aggregation, and cooperative value factorisation. Representative approaches include TarMAC, DGN, G2ANet, and VDQMix-related methods [6,24,36,48,86]. These approaches represent inter-agent dependencies through different mechanisms; importantly, value factorisation does not necessarily require direct communication. Their methodological contributions and original application domains are documented in Table 2 and Supplementary File S2.

Table 5. Qualitative comparison of MARL-related learning approaches and their potential relevance to multi-objective public transport.

Method	Methodological Characteristics	Main Limitations	Multi-Objective Considerations	Potential Public Transport Relevance
VDN [18]	Additive decomposition of cooperative joint action values into agent-level utilities	Additivity restricts the representation of non-additive joint action-value interactions	Standard implementations optimise scalar rewards; multi-objective extensions require additional formulation	Cooperative fleet decisions where additive value assumptions are appropriate
QMIX [29,30]	Nonlinear monotonic value mixing supporting decentralised greedy action selection	Monotonicity restricts representable joint action-value functions	Depends on reward representation and extensions; no inherent Pareto-front learning	Cooperative vehicle coordination under compatible value-factorisation assumptions
MAPPO [38,86]	Policy-gradient learning under CTDE; avoids explicit monotonic value factorisation	Performance depends on sample requirements, reward specification, and optimisation stability	Standard formulation optimises scalar feedback; extensions may support preferences or constraints	Coordinated fleet control with suitable observations, actions and reward formulations
HRL [27,28,89]	Temporal abstraction; decomposition of strategic and operational decisions	Performance depends on sub-task definitions, hierarchy design, and coordination between levels	Multi-objective capabilities depend on reward representation and the underlying learning algorithm	Multi-scale route planning and operational vehicle control
Offline MARL [37,90]	Policy learning from previously collected interaction data without online exploration during training	Sensitive to dataset coverage, distributional shift, and extrapolation errors; no general safety guarantee	Depends on offline reward formulation and constraint-handling methods	Data-supported policy development and simulation-based assessment before controlled deployment

Note: The table provides a qualitative comparison of methodological characteristics, representational assumptions, and potential limitations rather than a ranking of algorithmic performance. Multi-objective capabilities depend on the complete learning formulation, including reward representation, scalarisation, preference modelling, constraint handling, and algorithmic extensions. In particular, additive or monotonic value factorisation should not be conflated with limitations associated with weighted-sum scalarisation or the recovery of unsupported Pareto-optimal solutions. The public-transport applications are illustrative and do not establish empirical superiority or equivalent operational effectiveness across different transport settings.

Figure 6 conceptually organises explicit coordination, involving direct communication or information exchange, and implicit coordination, operating through shared environments, learning objectives, or incentives without requiring direct messaging. The taxonomy illustrates representative mechanisms from the reviewed literature rather than mutually exclusive classifications or quantitative distributions of the 41 studies. Its potential public-transport relevance should therefore be distinguished from application-specific empirical validation.

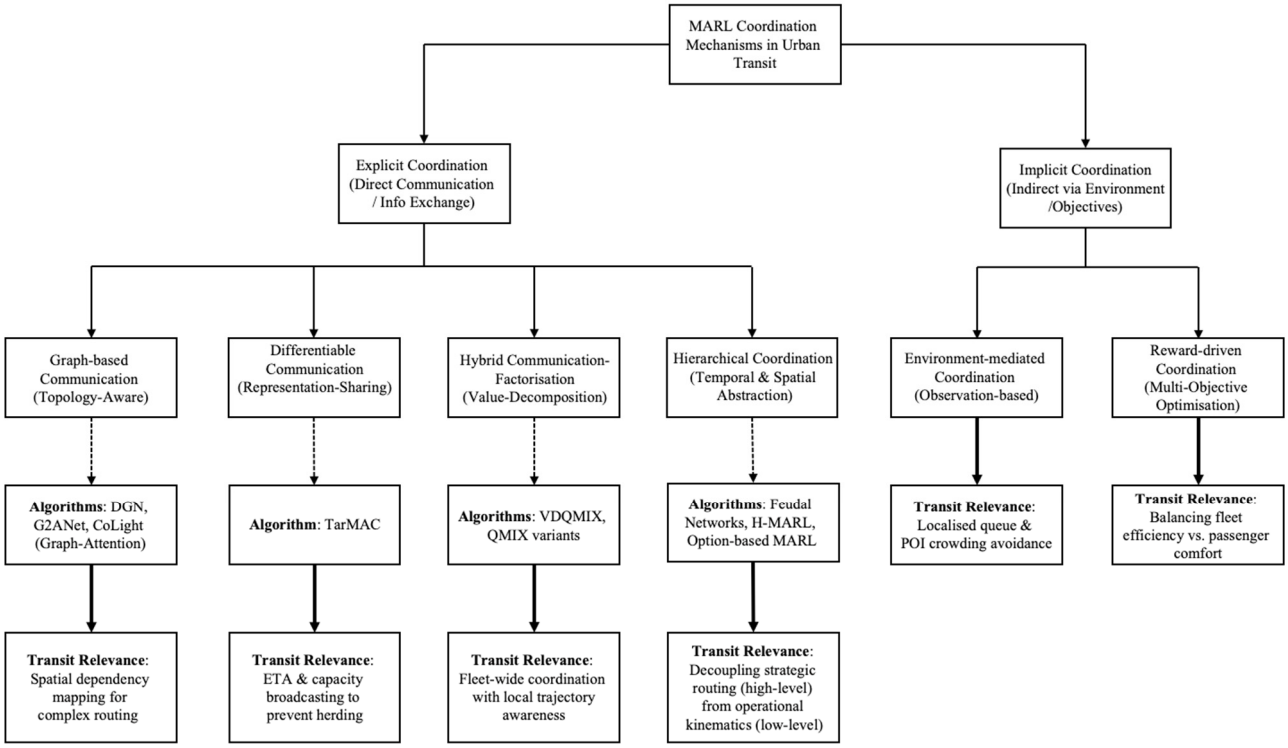


Figure 6. Conceptual taxonomy of explicit and implicit coordination mechanisms in multi-agent reinforcement learning. Note: The taxonomy illustrates representative coordination mechanisms, algorithmic examples, and potential public-transport applications derived from the methodological literature. The categories are conceptual and not necessarily mutually exclusive. The figure does not represent quantitative classifications of the 41 reviewed studies or establish empirical effectiveness in public-transport operations.

Graph-based coordination provides an example of potential methodological transferability. CoLight [25] uses graph-attention mechanisms to aggregate information from neighbouring traffic-signal controllers, assigning different importance to spatial interactions [93]. Similar representations could potentially incorporate congestion, queues, or shared-infrastructure conditions in public transport coordination. However, CoLight was developed and evaluated for traffic signal control rather than public transport fleet routing; its relevance to fleet-level MARL therefore remains methodological rather than direct empirical evidence [25,67,86].

To illustrate potential coordination challenges in tourist transport, consider several 45-seat tourist buses independently selecting the same high-value Point of Interest (POI). Concentrated arrivals at locations with limited access or stopping capacity could contribute to localised congestion and queuing, creating trade-offs between operational efficiency and passenger-related objectives [71,72]. This hypothetical scenario motivates an illustrative state–action–reward representation:

- ***State representation:*** Local observations could include congestion levels, POI queue lengths, and estimated vehicle dwell times, where available [45,64].
- ***Action selection:*** Agents could select feasible alternative routes, adjust permitted operating speeds, or modify dwell times subject to operational constraints [32,45,57].
- ***Reward signal:*** Reward components could account for congestion-related overlap, delays, schedule adherence, and passenger-related service quality, consistent with the multi-objective formulation in Section 3.

Two complementary coordination mechanisms may be considered. **Temporal staggering** adjusts feasible arrival times through dwell-time or permitted speed changes [31,64],

whereas **dynamic spatial rerouting** distributes arrivals across alternative feasible routes or POIs [14,32]. Both depend on service schedules, route feasibility, passenger preferences, and operational constraints.

Figure 7 illustrates these hypothetical coordination mechanisms by contrasting independent tourist-bus routing with potential information sharing, temporal staggering, and spatial rerouting. The scenario is a conceptual synthesis rather than a simulated or experimentally validated fleet controller, and the illustrated outcomes do not establish measured improvements in congestion, waiting time, or passenger satisfaction.

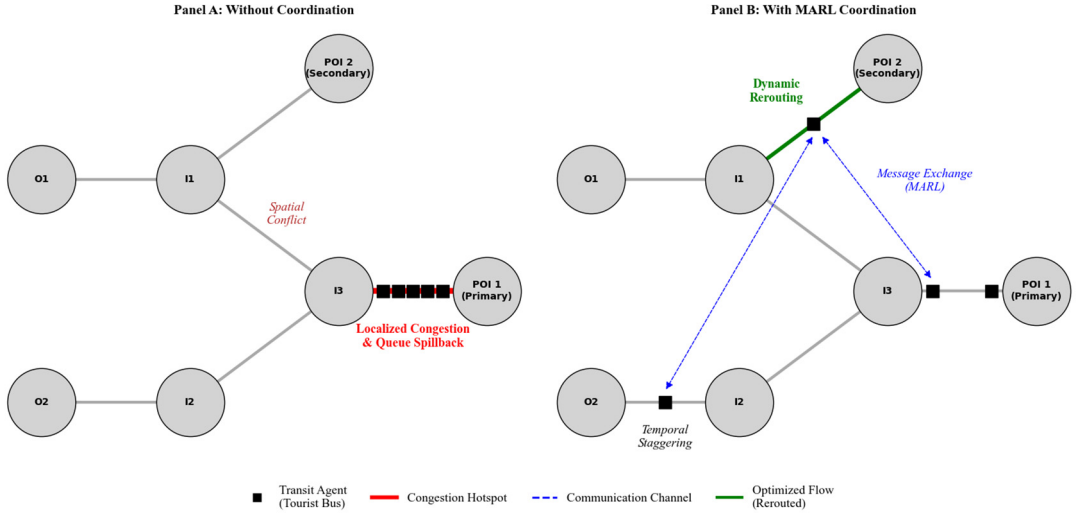


Figure 7. Hypothetical illustration of independent tourist-bus routing and potential MARL-based spatial and temporal coordination. (Note: This conceptual scenario illustrates possible mechanisms for information exchange, temporal staggering, and spatial rerouting under localised tourist demand concentration. It does not represent a simulation experiment, observed routing behaviour, or measured improvements in congestion, passenger waiting time, or service quality. O = Origin; I = Intersection/Intermediate Node; POI = Point of Interest).

Although communication may support coordinated decisions, it can introduce computational and communication overhead. Practical effectiveness depends on bandwidth, message latency, sensing uncertainty, communication reliability, scalability, and partial observability [47,67,86]. Consequently, the potential benefits of coordination mechanisms require application-specific evaluation before they can be interpreted as demonstrated improvements in public or tourist transport operations.

4.3. Algorithmic Challenges in Multi-Objective Settings

The algorithmic challenges discussed in this section are synthesised from the methodological studies and supporting literature identified in Table 2 and Supplementary File S2. They concern scalability, non-stationarity, and the representation of changing objective trade-offs rather than performance limitations established through a common public-transport benchmark.

Scalability and Computational Complexity. Multi-objective MARL can face substantial computational challenges when coordinating numerous agents with interdependent decisions [67,86]. Pareto-based approaches additionally seek to approximate multiple non-dominated policies, potentially increasing computational and sample requirements compared with optimising a single scalarised objective [16,84]. In transport networks involving shared intersections and Points of Interest (POIs), the growth of joint state–action spaces can further complicate learning and coordination [6,7].

Hierarchical Reinforcement Learning (HRL) and graph-based coordination offer potential mechanisms for managing these challenges. HRL separates strategic routing from lower-level operational control, while graph-attention mechanisms can prioritise relevant spatial interactions [25,27]. Their integration may reduce some representation and coordination burdens, but improved scalability or Pareto approximation in public-transport fleets remains a hypothesis requiring application-specific evaluation [62,67].

Non-Stationarity and Dynamic Objective Trade-offs. Public and tourist transport environments involve changing passenger demand, traffic conditions, and service requirements [15]. The relative importance of operational efficiency, ride comfort, and scenic quality may also vary with passenger preferences and travel contexts [14,73]. In multi-agent learning, simultaneously adapting policies can introduce additional non-stationarity from the perspective of individual agents [9].

These changes may complicate policy optimisation and the representation of multi-objective trade-offs. Policy updates can introduce learning instability, while changing reward preferences or environmental conditions may alter the attainable objective space [38,54,86]. Consequently, a fixed approximation of non-dominated policies may become less appropriate as operating conditions change. However, such shifts should not be interpreted as empirically demonstrated Pareto-front drift across the reviewed public-transport studies.

CTDE may alleviate certain learning difficulties by incorporating broader information during training, although its effectiveness depends on the algorithm and evaluation setting [46,47]. Dynamic preference conditioning and meta-learning offer potential directions for adapting to changing objectives and operating conditions [32,40,55,56]. Nevertheless, their ability to stabilise multi-objective learning or reduce retraining requirements in public-transport fleets requires further empirical validation.

4.4. Evidence-Based Operational Performance and Passenger Outcomes

The operational relevance of Multi-Agent Reinforcement Learning (MARL) in public transport depends on reported service and passenger-related outcomes. Passenger waiting time, in-vehicle travel time, passenger delay, headway regularity, and schedule adherence represent distinct performance measures that must be interpreted within their original evaluation settings. This section examines the three direct public-transport applications identified in Table 2 and Supplementary File S2, distinguishing demonstrated findings from potential benefits inferred from methodological research.

Passenger Waiting Time and Bus Holding Performance. Wang and Sun [31] investigated MARL-based dynamic bus holding, comparing their approach with no-holding, schedule-based, headway-based, and dynamic holding strategies in simulation. The reported results indicate more stable passenger waiting times, reduced unnecessary holding, and improved occupancy balance under the investigated conditions. However, improvements in headway regularity or holding efficiency do not necessarily correspond to equivalent reductions in passenger waiting time or total journey duration. These findings therefore support study-specific conclusions regarding coordinated holding performance rather than general superiority over alternative bus-control strategies.

Wang and Sun [45] extended coordinated bus control to multiple lines sharing a corridor. Their multi-objective MARL framework incorporated inter-line information and adaptive preference weighting. The reported findings indicate improved coordination and service regularity, particularly when preference weights emphasised passenger waiting-time objectives. These results remain specific to the study's simulation scenarios, baseline comparisons, and performance measures. They do not independently establish equivalent improvements across different public-transport networks or operating conditions.

Travel Time, Passenger Delay, and Schedule Reliability. Passenger journey time may include waiting, in-vehicle travel, holding, and transfer-related delays. Consequently, reductions in one operational measure cannot automatically be interpreted as improvements in another. The metro timetable-rescheduling study by Jufen et al. [60] provides application-specific evidence concerning scheduling and delay management. Under dense simulation scenarios, the reported MAA2C approach reduced total delay by approximately 35% relative to First-Come-First-Served (FCFS). This result concerns total delay and does not establish an equivalent reduction in passenger waiting time or in-vehicle travel time.

Synthesis of Reported Operational Evidence. Table 6 summarises the reported outcomes and evaluation baselines of the three direct public-transport studies. Wang and Sun [31] evaluated passenger waiting time and holding-related performance, Wang and Sun [45] examined multiline coordination and preference-dependent service regularity, and Jufen et al. [60] reported a quantitative total-delay improvement. Differences in transport modes, baselines, outcome definitions, and simulation settings prevent direct quantitative synthesis across these studies.

Table 6. Reported operational and passenger-related outcomes, comparison baselines, and evidence limitations of the three direct public-transport MARL applications.

Study	Evaluation Setting and Baseline	Performance Indicators	Reported Findings	Evidence Limitations
Wang & Sun (2020) [31]	Single-route bus-holding simulation; compared with no holding, headway-based, schedule-based, and dynamic holding strategies.	Average passenger waiting time; average holding duration; occupancy dispersion.	Coordinated MARL-based holding maintained more stable passenger waiting times, reduced unnecessary holding, and improved occupancy balance relative to conventional strategies.	Study-specific simulation findings; a verified numerical improvement in waiting time was not available for extraction.
Wang & Sun (2023) [45]	Two bus lines sharing a corridor; comparison of alternative coordination configurations and preference weighting.	Passenger waiting-time objective; bus bunching; service regularity.	Multiline coordination and preference adjustments were associated with improved coordination, reduced bunching, and improved service regularity.	No independently verified percentage reduction in passenger waiting time or journey time is reported in this synthesis.
Jufen et al. (2026) [60]	Flatland metro timetable-rescheduling simulation; MAA2C compared with First-Come-First-Served (FCFS) and DQN.	Total delay; passenger-delay objective; conflict resolution; computational performance.	MAA2C reduced reported total delay by approximately 35% relative to FCFS in dense simulation scenarios.	Total-delay reduction does not establish an equivalent percentage improvement in passenger waiting time or journey time.

Note: Findings are reported within the evaluation settings and baseline comparisons of the respective studies. Quantitative outcomes are included only where identifiable from the reported evidence. Waiting time, holding duration, service regularity, passenger delay, and total delay represent distinct performance measures and are not directly interchangeable. No pooled improvement estimates or general superiority of MARL over conventional optimisation methods is inferred.

Reporting-Quality Considerations and Evidence Interpretation. The six reporting-quality dimensions introduced in Section 2.3 provide a qualitative framework for interpreting the strength, scope, and transferability of the reported evidence. These considerations do not constitute a formal methodological-quality or risk-of-bias assessment. Instead, they clarify the extent to which the reported findings support particular conclusions under the original evaluation conditions.

Baseline adequacy was considered when interpreting comparative performance. For Wang and Sun [31], the reported holding-strategy comparisons support conclusions relative to the evaluated no-holding, schedule-based, headway-based, and dynamic holding alternatives. The findings from Wang and Sun [45] and Jufen et al. [60] were likewise interpreted within their respective reported comparison settings, including the FCFS baseline used in the latter study. These comparisons do not establish superiority over the broader range of conventional optimisation approaches, such as mixed-integer programming, time–space network models, or rolling-horizon optimisation, which were not evaluated under equivalent conditions.

Statistical reporting and robustness evaluation were considered separately from the magnitude of reported improvements. Numerical improvements were interpreted according to the repeated-run information, uncertainty estimates, statistical comparisons, and robustness analyses documented in the original studies. Where such information was absent or insufficiently specified in the available evidence, statistical significance, reproducibility across independent runs, and stability under alternative operating conditions were not inferred. Accordingly, the reported improvements were treated as study-specific findings rather than pooled evidence of consistent algorithmic effectiveness.

Evaluation setting, reproducibility information, and operational realism further informed the interpretation of application-specific findings. All three direct applications were evaluated in simulation-based settings, although their transport systems, operating assumptions, and performance measures differed. The availability of information concerning passenger demand, vehicle capacity, service schedules, network constraints, and experimental configurations affects the extent to which findings can be independently examined and transferred to other settings. Where operational assumptions or implementation details were simplified or incompletely documented, the synthesis restricted its conclusions to the reported evaluation conditions rather than inferring demonstrated real-world effectiveness. These considerations qualify the interpretation of reported benefits without independently establishing methodological validity or assigning numerical quality scores.

Comparison with Established Spatio-Temporal Optimisation Methods. The baselines in Table 6 represent the specific control strategies evaluated in the original studies, not the broader range of established spatio-temporal optimisation methods. The reviewed evidence does not provide a common benchmark comparing MARL with mixed-integer optimisation, time–space network formulations, or rolling-horizon approaches under equivalent operational conditions. Therefore, the reported improvements do not establish general superiority over conventional optimisation. Such comparisons require consistent scenarios, operational constraints, computational budgets, and passenger-level performance measures.

Evidence-Based Benefits and Operational Trade-offs. The three direct application studies indicate potential benefits in three areas. First, coordinated MARL-based control may improve the management of interacting vehicles, as illustrated by bus holding and multiline coordination [31,45]. Second, passenger-related objectives can be incorporated into coordinated decision-making, with reported improvements in waiting-time stability and service regularity under specific simulation conditions. Third, adaptive preference weighting and multi-agent timetable rescheduling illustrate mechanisms for addressing

competing service and scheduling objectives, including the reported total-delay reduction in Jufen et al. [60].

These benefits involve operational trade-offs. Bus holding may improve service regularity while increasing delays for passengers already on board, and coordinated control may require broader system information, communication, and appropriate reward design. Moreover, the available studies differ in their baselines, demand assumptions, statistical reporting, robustness evaluation, and operational configurations. Consequently, their findings do not support pooled estimates of MARL benefits, universal algorithmic rankings, or assumptions of equivalent passenger-level improvements.

Overall, the reviewed evidence supports the potential relevance of MARL for coordinated vehicle control and multi-objective public-transport management within the investigated simulation environments. However, the limited number of direct applications, differences in reported outcomes, and variability in evaluation and reporting conditions constrain the strength and generalisability of the conclusions. Broader claims of operational effectiveness and comparative superiority require further application-specific evaluations with clearly defined outcomes, appropriate baselines, transparent statistical reporting, robustness analyses, and realistic operational assumptions.

5. Practical Deployment Challenges (The Sim-to-Real Gap)

Transferring MARL-based transport policies from simulation to real-world operations introduces technical, environmental, and socio-technical challenges, including differences in data fidelity, sensing uncertainty, communication conditions, and human–AI interaction [91]. Figure 8 conceptually organises these interconnected barriers and potential feedback mechanisms based on the reviewed methodological and transport-related literature. It provides an interpretative synthesis rather than an experimentally validated deployment architecture or quantitative assessment of operational risks.

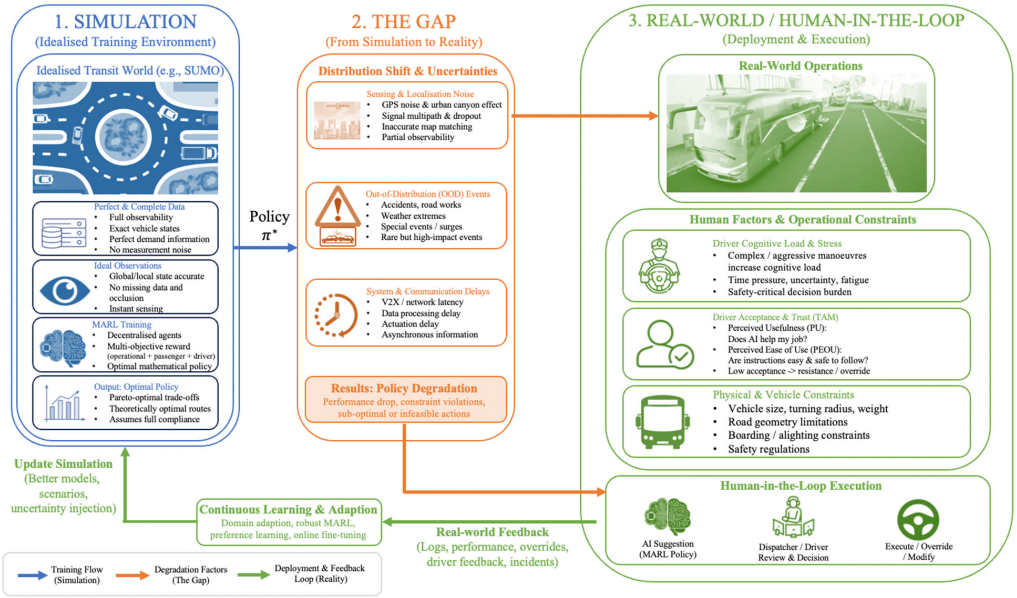


Figure 8. Conceptual synthesis of sim-to-real challenges relevant to MARL-based public transport operations. Note: The framework organises methodological and deployment-related challenges discussed in the reviewed literature. The relationships shown are interpretative and do not represent experimentally established causal pathways, quantified risk levels, or a validated implementation architecture. π^* denotes the policy optimised in the simulation environment, which may require further adaptation and validation before real-world deployment.

5.1. Simulation Environments vs. Real-World Noise

Simulation environments provide controlled settings for developing and evaluating MARL-based transport policies without exposing live operations to exploratory learning risks [4,13]. Microscopic traffic simulators, including SUMO, support multi-agent coordination experiments through structured traffic representations and configurable scenarios [6,68,94]. However, the fidelity of simulated observations and environmental dynamics varies across implementations, potentially limiting the transferability of learned policies.

In real-world transport operations, Global Positioning System (GPS) measurements and sensor observations may be affected by multipath errors, signal occlusion, transmission delays, and incomplete information, particularly in dense urban environments [95]. Policies trained under more accurate or timely observations may consequently experience state-estimation errors, potentially affecting routing and coordination decisions. These discrepancies motivate evaluation under realistic sensing noise, latency, and partial observability.

Beyond unintentional sensing errors, connected transport systems may face cyber-physical threats, including malicious vehicle-to-everything (V2X) messages and Global Navigation Satellite System (GNSS) spoofing [96]. Manipulated positioning or inter-agent information could distort local observations and influence coordinated decisions. Adversarial training offers a potential robustness mechanism by exposing policies to controlled disturbances during learning. Robust Adversarial Reinforcement Learning (RARL) and adversarial regularisation provide relevant methodological examples [39,97,98]. Nevertheless, their effectiveness against realistic V2X attacks or GNSS spoofing in multi-objective public-transport MARL remains to be empirically established.

Another challenge concerns out-of-distribution (OOD) conditions, including unexpected traffic disruptions, pedestrian surges, or changes in road availability that may be insufficiently represented in training scenarios [34]. Such changes can affect policy performance and the representation of operational trade-offs, although their impact depends on simulation fidelity, training coverage, and algorithmic robustness. They should not automatically be interpreted as demonstrated policy instability or Pareto-front drift in public-transport applications.

These challenges highlight the importance of controlled robustness evaluation before deployment. Potential safeguards include noise injection, domain randomisation, higher-fidelity simulation, safety constraints, and supervised policy adaptation [91,99]. Offline MARL provides a complementary approach by learning from previously collected interaction data without requiring exploratory interaction with live passenger services during training [90]. However, offline learning depends on dataset coverage and does not independently guarantee safe deployment. Subsequent validation should therefore consider operational constraints, uncertainty, and appropriate human oversight.

Overall, simulation-based performance alone is insufficient to establish real-world robustness. Reliable deployment requires application-specific testing under realistic sensing, communication, and environmental conditions, while distinguishing proposed robustness mechanisms from empirically demonstrated operational safety.

5.2. Driver Acceptance and Human-AI Interaction

Beyond algorithmic robustness and data fidelity, MARL deployment in public and tourist transport must account for human operators responsible for executing or supervising algorithmic decisions [35,100]. Even when a policy identifies favourable trade-offs between operational efficiency and passenger-related objectives, its practical effectiveness depends on route feasibility, driver acceptance, and operational constraints.

A potential challenge arises when algorithmically recommended routes involve narrow streets, complex manoeuvres, or demanding driving conditions. Such decisions may

increase cognitive workload or reduce operator acceptance, particularly when optimisation objectives do not adequately represent vehicle limitations and human factors [101–104]. These concerns highlight the distinction between mathematical optimisation and operational feasibility rather than demonstrating systematic driver non-compliance with MARL recommendations.

The Technology Acceptance Model (TAM) provides a complementary framework for examining perceived usefulness (PU) and perceived ease of use (PEOU) [105]. In MARL-assisted transport operations, these constructs could inform assessments of whether routing recommendations align with operator workflows and can be interpreted and executed safely. Where appropriately measured, driver-related preferences and operational requirements could also inform reward design, action restrictions, or evaluation criteria. However, TAM constructs do not directly constitute validated reward penalties or safety constraints without additional operationalisation and empirical assessment.

Human-centred deployment may further benefit from explicitly representing vehicle manoeuvrability, driver workload, and safety-related restrictions alongside passenger comfort considerations [58,70]. Constraint-based optimisation, including Lagrangian approaches discussed in Section 3.3, offers potential mechanisms for incorporating operational limits, but does not independently guarantee constraint satisfaction or real-world safety [82]. Human-in-the-loop reinforcement learning and preference-based MARL may support AI-assisted decision-making in which dispatchers and drivers can review, modify, or override policy recommendations [35,56]. Their effectiveness in professional public-transport operations nevertheless requires application-specific evaluation of usability, acceptance, operational performance, and safety.

6. Open Challenges and Future Directions

Despite methodological advances in Multi-Agent Reinforcement Learning (MARL), transferring multi-objective transport coordination approaches from simulation to real-world operations remains challenging. Key limitations concern scalability, stochasticity, partial observability, uncertain passenger preferences, and operational constraints. Building on the evidence synthesised in the preceding sections, this section identifies potential research directions in human-centred reward modelling, foundation-model integration, and approaches to narrowing the simulation-to-reality gap. These directions represent prospective methodological opportunities rather than empirically established solutions for public-transport fleets.

6.1. Quantifying Subjective Human Models

Human-centred reward components, such as the passenger-related reward illustrated in Section 3.2, may be represented through measurable proxies, including Point-of-Interest (POI) accessibility, scenic characteristics, and vehicle-kinematic indicators of comfort. However, these proxies may not fully reflect heterogeneous or context-dependent passenger preferences. Human satisfaction can involve latent behavioural and psychological factors that are not directly observable from operational data [106,107].

Future multi-objective MARL research could therefore examine uncertainty-aware and preference-dependent reward models rather than assuming fixed relationships between observable indicators and passenger satisfaction. Evaluation should also consider variability across passengers, operational scenarios, and experimental runs, where appropriate, to assess the robustness of reported outcomes.

Three complementary research directions may support more representative human-centred reward modelling:

- ***Learning-based reward inference and policy distillation:*** Inverse Reinforcement Learning (IRL) could infer latent preferences from observed route choices or behavioural feedback [108]. Multi-agent imitation learning could additionally learn coordination policies from suitably documented expert demonstrations, including historical dispatcher decisions [22]. However, inferred preferences and imitated policies depend on the quality, coverage, and representativeness of the available observations and do not necessarily reproduce the underlying intentions of human decision-makers.
- ***Real-time adaptive rewards:*** Context-dependent reward formulations could incorporate changing environmental and operational conditions, including weather, congestion, or crowd density, where reliable data are available. Such adaptation may help represent changing service priorities, but requires careful reward specification, uncertainty handling, and evaluation under dynamic conditions.
- ***Human-in-the-loop feedback integration:*** Active preference learning could incorporate feedback from passengers, dispatchers, or other relevant operators to update reward models and assess the acceptability of recommended decisions [35,56]. Its practical effectiveness depends on feedback quality, interaction requirements, and the reliability of the resulting preference representations.

Together, these directions provide potential mechanisms for improving the representation of passenger-related objectives in MARL-based transport coordination. However, their ability to improve passenger satisfaction, operational performance, or real-world deployment remains to be established through application-specific evaluation.

6.2. LLMs and Foundation Models for Automated Reward Shaping

Building upon the challenges of human-centred reward modelling discussed in Section 6.1, translating high-level operational requirements and passenger preferences into structured reward functions remains a potential challenge for MARL-based transport coordination. For example, operator instructions concerning scenic routes, congestion avoidance, or service priorities may require interpretation into the vector-valued reward components introduced in Section 3.2. Large Language Models (LLMs) and multimodal foundation models offer prospective mechanisms for supporting this translation, although their reliability, interpretability, computational requirements, and susceptibility to reward misspecification require careful evaluation [109,110].

Four complementary research directions may be considered:

- ***Natural Language to Reward (NL2Reward) Synthesis:*** LLMs could assist in translating operator instructions or natural-language service requirements into structured reward specifications or parameterised objective functions [42,43]. Rather than directly generating unconstrained control policies, such systems could produce candidate reward components subject to predefined operational constraints and independent verification. However, generated specifications would require validation to establish their consistency with intended objectives and safety requirements.
- ***Foundation Models as Auxiliary Evaluators:*** Vision–Language Models (VLMs) and LLMs could be investigated as auxiliary evaluators or sources of weak supervision for passenger-related attributes, including scenic quality, perceived comfort, or contextual service conditions [44,111]. Multimodal information from visual observations, event descriptions, or other available data could potentially supplement conventional spatial and kinematic proxies. Nevertheless, such model outputs should not be treated as validated passenger-satisfaction measures without independent assessment.
- ***Semantic-Driven Trade-off Calibration:*** Foundation models could support higher-level adjustment of scalarisation weights or constraint thresholds in response to changing service requirements, as discussed in Section 3.3. For example, an operator-approved

adaptation could modify the relative emphasis on operational efficiency, passenger comfort, or scenic quality under adverse conditions. Such adjustments would require explicit operational bounds, consistency checks, and evaluation of their effects on downstream MARL policies rather than assuming stable or safe adaptation.

- ***LLM-Guided Hierarchical Policies:*** Building upon the Hierarchical Reinforcement Learning (HRL) approaches discussed in Section 4.1, LLMs could serve as auxiliary high-level semantic planners that interpret operator instructions and translate them into strategic objectives or structured sub-tasks [112,113]. Lower-level MARL agents could remain responsible for time-sensitive routing and operational decisions. This separation may reduce reliance on frequent LLM inference within the control loop, although the reliability, latency, and feasibility of generated strategic instructions would require independent validation.

Collectively, these directions suggest potential roles for foundation models in reward specification, contextual evaluation, preference adaptation, and hierarchical decision support. However, the underlying methodological studies do not establish equivalent effectiveness or operational safety in public-transport MARL. Future research should therefore evaluate generated reward consistency, computational overhead, constraint satisfaction, human oversight, and application-specific performance before considering real-world deployment.

6.3. Digital Twins for Urban Networks: Bridging the Continuous Sim-to-Real Gap

As discussed in Section 5.1, transferring MARL policies from simulation to real-world public transport requires accounting for differences in traffic dynamics, sensing uncertainty, and out-of-distribution (OOD) conditions. Although microscopic simulators such as SUMO can represent stochastic traffic scenarios, their fidelity depends on model calibration, available data, and the operational conditions represented. Urban Digital Twins (DTs) offer a potential extension by integrating simulation models with observations from physical transport networks, supporting updated state representations and scenario-based evaluation [33,114].

However, maintaining sufficiently accurate and timely DT representations introduces challenges concerning sensor reliability, communication latency, computational requirements, and model consistency. Moreover, continuous policy adaptation within a DT does not independently guarantee safe deployment in physical transport systems. Future research could investigate three complementary directions:

- ***Continuous Offline-to-Online Fine-Tuning:*** DTs could integrate historical operational records with updated traffic observations, passenger demand, and simulation-generated trajectories to support periodic policy evaluation and adaptation. Offline MARL and conservative actor rectification, discussed in Section 4.1, provide potential mechanisms for limiting unsupported policy updates under distributional shift [37,90]. Within a CTDE framework, updated information could inform centralised training while decentralised policies remain subject to operational constraints and independent validation. Nevertheless, conservative policy adaptation does not guarantee safe actions or eliminate risks arising from incomplete data and model mismatch.
- ***Dynamic Uncertainty and Noise Replication:*** DT environments could incorporate configurable sensing disturbances and communication uncertainties, including GPS multipath errors, V2X message loss, and delayed observations. They could also support controlled evaluation under adversarial scenarios, such as GNSS spoofing or manipulated cooperative messages, extending the robustness considerations discussed in Section 5.1. Such testing may help identify policy sensitivities and potential fail-

ure modes, although robustness demonstrated in simulated environments does not independently establish real-world security or operational safety.

- ***Edge-Assisted and Federated Digital Twins:*** Distributed DT architectures could be investigated to reduce reliance on fully centralised processing and support local data management at transport hubs or connected vehicles [115]. Edge computing may enable more localised state estimation and coordination updates, potentially supporting time-sensitive fleet decisions. However, its effectiveness depends on communication infrastructure, computational resources, synchronisation requirements, and system architecture. Federated approaches additionally require careful consideration of model consistency, privacy, and heterogeneous data availability.

Collectively, these directions suggest how DTs could support iterative simulation, uncertainty-aware evaluation, and controlled MARL policy adaptation. Nevertheless, their effectiveness in public-transport fleet coordination remains application-dependent and requires validation under realistic operational conditions. DT-based testing should therefore complement, rather than replace, independent safety assessment, operational constraints, and appropriate human oversight.

6.4. Review Limitations

This systematic review has several methodological limitations. First, the original literature search was restricted to English-language publications published between January 2018 and May 2026. Although earlier foundational studies were selectively cited for theoretical context, relevant research outside this period or in other languages may have been missed. In addition, differences in database coverage, search functionality, and metadata quality across the five databases and search platforms may have affected retrieval consistency. The original database-specific executable queries, exact search dates, and filtering configurations were not consistently preserved, limiting independent reproduction of the historical identification counts.

In particular, the original IEEE Xplore identification process reported 410 records in May 2026, while an independent reviewer reported 510 records in October 2026. A subsequent author-initiated search conducted on 11 November 2026 retrieved 516 records, comprising 218 journal articles, 296 conference papers, one book, and one magazine. As detailed in Section 2.1 and Supplementary File S3, these retrievals represent distinct search events rather than directly comparable identification datasets. The updated search extended beyond the original May 2026 publication cut-off, and equivalence among the historical and updated search configurations could not be established. Consequently, differences in retrieval counts cannot be attributed solely to database growth or interpreted as evidence that a specific number of eligible studies was omitted. Because record-level overlap and eligibility assessment across the retrieval datasets were not established, the possibility that relevant studies were missed during the original identification process cannot be excluded. The 41 studies retained for synthesis therefore represent the corpus selected through the documented original PRISMA process rather than a demonstrably exhaustive collection of all potentially eligible publications.

Second, study screening, eligibility assessment, and data extraction were conducted by a single author without independent verification, introducing potential selection and classification biases. Although the eligibility criteria and study-level rationales in Supplementary File S2 improve transparency, they do not eliminate these limitations.

Third, the 41 mapped studies are heterogeneous in application domain, methodological formulation, optimisation objectives, evaluation settings, and performance measures. Consequently, statistical meta-analysis was not considered appropriate, and the reported distributions should be interpreted as descriptive characteristics of the selected literature rather than

pooled estimates or evidence of algorithmic effectiveness. Moreover, no formal risk-of-bias or methodological-quality assessment was conducted. The six reporting-quality dimensions discussed in Sections 2.3 and 2.6, evaluation setting, baseline adequacy, statistical reporting, robustness evaluation, reproducibility information, and operational realism, were used as qualitative interpretative considerations rather than validated methodological-quality appraisal criteria or study-level quality scores. Accordingly, the synthesis distinguishes reported methodological findings from the strength and completeness of their supporting evidence, and reporting completeness should not be equated with methodological validity or low risk of bias.

Fourth, the distinction between direct public-transport evidence and supporting methodological literature limits cross-study comparisons. Only three mapped studies directly evaluated MARL in public-transport applications, while the remaining studies primarily informed methodological or contextual discussions. Findings from general MARL benchmarks and adjacent transport domains cannot independently establish equivalent performance in public or tourist transport fleets. Consequently, the comparative analysis does not provide universal rankings of MARL architectures or learning paradigms.

Fifth, operational and passenger-related outcomes discussed in Section 4.4 are not directly interchangeable. The three direct public-transport studies differ in their definitions of passenger waiting time, travel time, delay, headway regularity, baseline strategies, and simulation conditions. Improvements in service regularity or operational delay cannot automatically be interpreted as equivalent reductions in passenger waiting time or total journey duration. These differences preclude pooled estimates of passenger-level improvements and limit the generalisability of the reported operational benefits beyond the evaluated settings.

A further limitation is the absence of systematic, like-for-like comparisons between MARL approaches and established spatio-temporal optimisation methods. Although mixed-integer optimisation, time–space network models, and rolling-horizon approaches provide established mechanisms for transport planning and control, the reviewed direct applications primarily selected holding, scheduling, or learning-based baselines. The available evidence therefore does not establish the relative computational efficiency, solution quality, robustness, or passenger-level benefits of MARL against conventional optimisation under equivalent conditions.

The conceptual illustrations in Figures 4–8, including the Multi-Objective Partially Observable Stochastic Game (MO-POSG) representation, multi-objective trade-off diagram, coordination taxonomy, hypothetical tourist-bus scenario, and simulation-to-real-world transfer framework, serve to organise methodological insights rather than present original experimental findings or validated operational improvements. Quantitative classifications reported elsewhere in the review must likewise be interpreted according to their underlying study-level coding and should not be conflated with these conceptual illustrations.

Finally, the literature may have evolved beyond the original May 2026 search cut-off, particularly in areas such as offline MARL, foundation models, and Digital Twins. Although the November 2026 IEEE Xplore retrieval provides a subsequent search observation, it was not incorporated into a new systematic screening and eligibility-assessment process and therefore does not constitute a comprehensive update of the review corpus. Future reviews would benefit from broader database coverage, fully documented and reproducible search strategies, independent multi-reviewer screening and extraction, transparent study classification, and more standardised reporting of passenger-level outcomes and operational baselines. Where feasible, structured methodological-quality appraisal should also be considered, while maintaining clear distinctions between reporting completeness, methodological validity, and application-specific evidence strength.

7. Conclusions

This systematic review examined 41 studies relevant to Multi-Agent Reinforcement Learning (MARL) for urban public and tourist transport through structured descriptive and comparative synthesis. The reviewed literature encompasses direct public-transport applications, general MARL/MORL methodological contributions, adjacent transport studies, and supporting deployment-related research, as documented in Supplementary File S2. By distinguishing application-specific evidence from broader methodological insights, the review examined MARL architectures, multi-objective reward design, trade-off resolution strategies, operational and passenger-related outcomes, and real-world deployment challenges.

The methodological synthesis highlights Centralised Training and Decentralised Execution (CTDE), value-based factorisation, policy-based learning, and hierarchical and offline MARL as relevant approaches to multi-agent coordination. Their capabilities and limitations depend on reward formulation, operational constraints, and evaluation conditions rather than architectural distinctions alone. Similarly, linear scalarisation, constraint-based optimisation, and Pareto-based approaches provide alternative or complementary mechanisms for addressing competing objectives, including travel time, service reliability, passenger comfort, scenic quality, and Point-of-Interest (POI) accessibility. However, the available evidence does not establish universally superior learning architectures or objective-resolution strategies for public-transport operations.

The three direct public-transport applications provide simulation-based evidence concerning coordinated bus holding, multiline bus control, and metro timetable rescheduling. Reported findings indicate potential improvements in service regularity, fleet coordination, and selected delay-related outcomes under study-specific conditions. Nevertheless, passenger waiting time, in-vehicle travel time, headway regularity, and operational delay are distinct measures. Improvements in one indicator cannot automatically be interpreted as equivalent improvements in another. Differences in baselines and evaluation settings limit quantitative cross-study comparisons, while the absence of systematic comparisons with established spatio-temporal optimisation methods prevents conclusions about MARL's general superiority.

Transferring MARL-based approaches to real-world public transport remains challenging because of sensing uncertainty, changing traffic conditions, communication limitations, operational constraints, and human acceptance. Potential research directions include uncertainty-aware evaluation, human-centred reward modelling, Offline MARL, conservative policy learning, foundation-model-assisted decision support, and Digital Twin environments. These approaches require independent assessment of operational feasibility, robustness, and safety before deployment.

Overall, MARL offers a potentially useful framework for investigating coordinated transport decisions involving competing operational and passenger-related objectives. However, the current evidence remains heterogeneous, with limited direct public-transport validation. Future progress requires transparent and reproducible review practices, consistent passenger-level performance measures, stronger application-specific comparisons, including conventional optimisation baselines, and careful integration of operational constraints and human-centred requirements.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/vehicles8110304/s1. Supplementary File S1 (PRISMA 2020 Checklist); Supplementary File S2 (Study Characteristics and Methodological Coding of the 41 Studies Included in the Structured Synthesis); Supplementary File S3 (Database-Specific Search Strategies).

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this systematic review were derived from the published literature cited in the manuscript. Study-level characteristics, evidence classifications, and methodological coding of the 41 included studies are provided in Supplementary File S2. Search strategies, retrieval information, and reproducibility documentation are provided in Supplementary File S3, while the PRISMA 2020 checklist is available in Supplementary File S1. No primary participant-level data were collected.

Acknowledgments: Declaration of Generative AI and AI-assisted Technologies in the Writing Process: During the preparation of this work, the author(s) used ChatGPT (GPT-6 Pro, OpenAI) to improve grammar and language readability. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Guihaire, V.; Hao, J.-K. Transit network design and scheduling: A global review. *Transp. Res. Part A Policy Pract.* **2008**, *42*, 1251–1273. [CrossRef]
2. Ibarra-Rojas, O.J.; Delgado, F.; Giesen, R.; Muñoz, J.C. Planning, operation, and control of bus transport systems: A literature review. *Transp. Res. Part B Methodol.* **2015**, *77*, 38–75. [CrossRef]
3. Le, P.G.B.; Moridpour, S.; Dinh, M.N. Tourist route planning in urban road networks: A systematic review of optimisation models and data integration. *J. Adv. Transp.* **2026**, *2026*, 8706872. [CrossRef]
4. Haydari, A.; Yılmaz, Y. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 11–32. [CrossRef]
5. Michailidis, P.; Michailidis, I.; Lazaridis, C.R.; Kosmatopoulos, E. Traffic signal control via reinforcement learning: A review on applications and innovations. *Infrastructures* **2025**, *10*, 114. [CrossRef]
6. Chu, T.; Wang, J.; Codecà, L.; Li, Z. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1086–1095. [CrossRef]
7. El-Tantawy, S.; Abdulhai, B.; Abdelgawad, H. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): Methodology and large-scale application on downtown Toronto. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 1140–1150. [CrossRef]
8. Lin, K.; Zhao, R.; Xu, Z.; Zhou, J. Efficient large-scale fleet management via multi-agent deep reinforcement learning. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1774–1783.
9. Lowe, R.; Wu, Y.I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
10. Shalev-Shwartz, S.; Shammah, S.; Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv* **2016**, arXiv:1610.03295.
11. Le, P.G.B.; Moridpour, S.; Dinh, M.N.; Nguyen, B. Macroscopic impacts of optimised tourist bus routes on urban road networks: A spatiotemporal analysis. *Transp. Eng.* **2026**, *26*, 100461. [CrossRef]
12. Qin, Z.T.; Zhu, H.; Ye, J. Reinforcement learning for ridesharing: An extended survey. *Transp. Res. Part C Emerg. Technol.* **2022**, *144*, 103852. [CrossRef]
13. Zong, Z.; Feng, T.; Wang, J.; Xia, T.; Li, Y. Deep reinforcement learning for demand-driven services in logistics and transportation systems: A survey. *ACM Trans. Knowl. Discov. Data* **2025**, *19*, 1–42. [CrossRef]
14. Gavalas, D.; Konstantopoulos, C.; Mastakas, K.; Pantziou, G. A survey on algorithmic approaches for solving tourist trip design problems. *J. Heuristics* **2014**, *20*, 291–328. [CrossRef]
15. Ronald, N.; Thompson, R.; Winter, S. Simulating demand-responsive transportation: A review of agent-based approaches. *Transp. Rev.* **2015**, *35*, 404–421. [CrossRef]
16. Hayes, C.F.; Rădulescu, R.; Bargiacchi, E.; Källström, J.; Macfarlane, M.; Reymond, M.; Verstraeten, T.; Zintgraf, L.M.; Dazeley, R.; Heintz, F.; et al. A practical guide to multi-objective reinforcement learning and planning. *Auton. Agents Multi-Agent Syst.* **2022**, *36*, 26. [CrossRef]
17. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; A Akl, E.; E Brennan, S.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, 372. [CrossRef] [PubMed]
18. Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W.M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J.Z.; Tuyls, K.; et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv* **2017**, arXiv:1706.05296.

19. Qin, Z.; Tang, X.; Jiao, Y.; Zhang, F.; Xu, Z.; Zhu, H.; Ye, J. Ride-hailing order dispatching at didi via reinforcement learning. *Inf. J. Appl. Anal.* **2020**, *50*, 272–286. [CrossRef]
20. Sun, J.; Jin, H.; Ding, R.; Fan, G.; Wei, Y.; Su, L. Multi-objective order dispatch for urban crowd sensing with for-hire vehicles. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*; IEEE: New York, NY, USA, 2023; pp. 1–10.
21. Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; Whiteson, S. Counterfactual multi-agent policy gradients. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
22. Song, J.; Ren, H.; Sadigh, D.; Ermon, S. Multi-agent generative adversarial imitation learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
23. Tessler, C.; Mankowitz, D.J.; Mannor, S. Reward constrained policy optimization. *arXiv* **2018**, arXiv:1805.11074.
24. Das, A.; Gervet, T.; Romoff, J.; Batra, D.; Parikh, D.; Rabbat, M.; Pineau, J. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*; PMLR: Sydney, Australia, 2019; pp. 1538–1546.
25. Wei, H.; Xu, N.; Zhang, H.; Zheng, G.; Zang, X.; Chen, C.; Xu, K.; Li, Z. Colight: Learning network-level cooperation for traffic signal control. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 1913–1922.
26. Christianos, F.; Papoudakis, G.; Rahman, M.A.; Albrecht, S.V. Scaling multi-agent reinforcement learning with selective parameter sharing. In *International Conference on Machine Learning*; PMLR: Sydney, Australia, 2021; pp. 1989–1998.
27. Duan, J.; Eben Li, S.; Guan, Y.; Sun, Q.; Cheng, B. Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data. *IET Intell. Transp. Syst.* **2020**, *14*, 297–305. [CrossRef]
28. Lu, Y.; Xu, X.; Zhang, X.; Qian, L.; Zhou, X. Hierarchical reinforcement learning for autonomous decision making and motion planning of intelligent vehicles. *IEEE Access* **2020**, *8*, 209776–209789. [CrossRef]
29. Rashid, T.; Farquhar, G.; Peng, B.; Whiteson, S. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 10199–10210.
30. Rashid, T.; Samvelyan, M.; De Witt, C.S.; Farquhar, G.; Foerster, J.; Whiteson, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *J. Mach. Learn. Res.* **2020**, *21*, 1–51.
31. Wang, J.; Sun, L. Dynamic holding control to avoid bus bunching: A multi-agent deep reinforcement learning framework. *Transp. Res. Part C Emerg. Technol.* **2020**, *116*, 102661. [CrossRef]
32. Chen, L.; Hu, B.; Guan, Z.H.; Zhao, L.; Shen, X. Multiagent meta-reinforcement learning for adaptive multipath routing optimization. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 5374–5386. [CrossRef] [PubMed]
33. Rudskoy, A.; Ilin, I.; Prokhorov, A. Digital twins in the intelligent transport systems. *Transp. Res. Procedia* **2021**, *54*, 927–935. [CrossRef]
34. Suo, S.; Regalado, S.; Casas, S.; Urtasun, R. Trafficsim: Learning to simulate realistic multi-agent behaviors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 10400–10409.
35. Wu, J.; Huang, Z.; Huang, C.; Hu, Z.; Hang, P.; Xing, Y.; Lv, C. Human-in-the-loop deep reinforcement learning with application to autonomous driving. *arXiv* **2021**, arXiv:2104.07246.
36. Fawaz, H.; Lesca, J.; Quang, P.T.A.; Leguay, J.; Zeghlache, D.; Medagliani, P. Graph convolutional reinforcement learning for collaborative queuing agents. *IEEE Trans. Netw. Serv. Manag.* **2022**, *20*, 1363–1377. [CrossRef]
37. Pan, L.; Huang, L.; Ma, T.; Xu, H. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *International Conference on Machine Learning*; PMLR: Sydney, Australia, 2022; pp. 17221–17237.
38. Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24611–24624. [CrossRef]
39. Bukharin, A.; Chen, Z.; Li, Y.; Yu, Y.; Zhang, C.; Zhang, Q.; Zhang, S.; Zhao, T.; Zuo, S. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 68121–68133. [CrossRef]
40. Cai, X.Q.; Zhang, P.; Zhao, L.; Bian, J.; Sugiyama, M.; Llorens, A. Distributional pareto-optimal multi-objective reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 15593–15613. [CrossRef]
41. Fu, X.; Chen, S.; Liang, Q.; Li, Y. Research on Multi-Agent Reinforcement Learning Traffic Control. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*; IEEE: New York, NY, USA, 2023; pp. 231–239.
42. Kwon, M.; Xie, S.M.; Bullard, K.; Sadigh, D. Reward design with language models. *arXiv* **2023**, arXiv:2303.00001.
43. Ma, Y.J.; Liang, W.; Wang, G.; Huang, D.A.; Bastani, O.; Jayaraman, D.; Zhu, Y.; Fan, J.; Anandkumar, A. Eureka: Human-level reward design via coding large language models. *arXiv* **2023**, arXiv:2310.12931.
44. Rocamonde, J.; Montesinos, V.; Nava, E.; Perez, E.; Lindner, D. Vision-language models are zero-shot reward models for reinforcement learning. *arXiv* **2023**, arXiv:2310.12921.
45. Wang, J.; Sun, L. Multi-objective multi-agent deep reinforcement learning to reduce bus bunching for multiline services with a shared corridor. *Transp. Res. Part C Emerg. Technol.* **2023**, *155*, 104309. [CrossRef]

46. Zhou, Y.; Liu, S.; Qing, Y.; Chen, K.; Zheng, T.; Song, J.; Song, M. Is centralized training with decentralized execution framework centralized enough for marl? *arXiv* **2023**, arXiv:2305.17352.
47. Amato, C. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning. *arXiv* **2024**, arXiv:2409.03052.
48. Liu, Z.; Zhang, J.; Shi, E.; Liu, Z.; Niyato, D.; Ai, B.; Shen, X. Graph neural network meets multi-agent reinforcement learning: Fundamentals, applications, and future directions. *IEEE Wirel. Commun.* **2024**, *31*, 39–47. [CrossRef]
49. Palma, Á.; Morais, P.; Alves, A. Route Recommendation Based on POIs and Public Transportation. *KDIR* **2024**, 391–398. [CrossRef]
50. Fang, B.; Tang, D.; Wang, Z.; Wang, H. Multi-agent reinforcement learning based on self-satisfaction in sparse reward scenarios. *Int. J. Bio-Inspir. Comput.* **2025**, *25*, 56–67. [CrossRef]
51. Le, B.G.; Ta, V.C. Toward finding strong pareto optimal policies in multi-agent reinforcement learning. *Mach. Learn.* **2025**, *114*, 60. [CrossRef]
52. Li, H.; He, H. Bridging Theory and Practice of MAPPO for Cooperative Multi-Agent Continuous Control. In *2025 International Joint Conference on Neural Networks (IJCNN)*; IEEE: New York, NY, USA, 2025; pp. 1–7.
53. Liu, D.; Guo, J.; Gu, Y.; King, M.; Han, L.D.; Brakewood, C. Analyzing Transit Systems Using General Transit Feed Specification (GTFS) by Generating Spatiotemporal Transit Networks. *Information* **2025**, *16*, 24. [CrossRef]
54. Liu, E.; Wu, Y.C.; Huang, X.; Gao, C.; Wang, R.J.; Xue, K.; Qian, C. Pareto set learning for multi-objective reinforcement learning. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 18789–18797. [CrossRef]
55. Liu, R.; Pan, Y.; Xu, L.; Song, L.; You, P.; Chen, Y.; Bian, J. Efficient discovery of Pareto front for multi-objective reinforcement learning. In Proceedings of the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
56. Mu, N.; Luan, Y.; Jia, Q.S. Preference-based multi-objective reinforcement learning. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 18737–18749. [CrossRef]
57. Sun, X.; Lu, Y.; Wu, Q.; Li, H.; Xu, Z.; Zheng, L.; Wang, Q. An eco-driving strategy of vehicle comfort-safety and energy efficiency based on deep reinforcement learning. *Alex. Eng. J.* **2025**, *132*, 252–263. [CrossRef]
58. Zhang, J.; Zhou, S.; Nie, L.; Guo, W. An Intelligent Evaluation Approach for Riding Comfort Incorporating Vehicle Vibration Acceleration Distribution Characteristics. *Int. J. Pavement Res. Technol.* **2025**, *19*, 1362–1381. [CrossRef]
59. Azari, M.; Moridpour, S.; Hatami, M.; Kashi, S.M.H. A Spatial Agent-Based Approach for Modeling and Mapping Multi-Locality Destination Choices. *Sustainability* **2026**, *18*, 1904. [CrossRef]
60. Jufen, Y.; Haozhe, Y.; Weikang, W.; Chengyang, X. Large-Scale Metro Train Timetable Rescheduling via Multi-Agent Deep Reinforcement Learning: A High-Dimensional Optimization Approach in Flatland Environment. *Appl. Sci.* **2026**, *16*, 3338. [CrossRef]
61. Littman, M.L. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*; Morgan Kaufmann: San Francisco, CA, USA, 1994; pp. 157–163.
62. Rădulescu, R.; Mannion, P.; Roijers, D.M.; Nowé, A. Multi-objective multi-agent decision making: A utility-based analysis and survey. *Auton. Agents Multi-Agent Syst.* **2020**, *34*, 10. [CrossRef]
63. Chen, C.X.; Chen, W.Y.; Chen, Z.Y. A Multi-Agent Reinforcement Learning approach for bus holding control strategies. *Adv. Transp. Stud.* **2015**, *2 (Special Issue)*, 41–54.
64. Chen, W.; Zhou, K.; Chen, C. Real-time bus holding control on a transit corridor based on multi-agent reinforcement learning. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*; IEEE: New York, NY, USA, 2016; pp. 100–106.
65. Joe, W.; Lau, H.C. Deep reinforcement learning approach to solve dynamic vehicle routing problem with stochastic customers. *Proc. Int. Conf. Autom. Plan. Sched.* **2020**, *30*, 394–402. [CrossRef]
66. Roijers, D.M.; Vamplew, P.; Whiteson, S.; Dazeley, R. A survey of multi-objective sequential decision-making. *J. Artif. Intell. Res.* **2013**, *48*, 67–113. [CrossRef]
67. Nguyen, T.T.; Nguyen, N.D.; Nahavandi, S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Trans. Cybern.* **2020**, *50*, 3826–3839. [CrossRef] [PubMed]
68. Flitsch, C.; Kastner, K.H.; Bósa, K.; Neubauer, M. Calibrating traffic simulation models in sumo based upon diverse historical real-time traffic data–lessons learned in its upper austria. *Epic. Ser. Eng.* **2018**, *2*, 25–42. [CrossRef]
69. Boeing, G. Modeling and analyzing urban networks and amenities with OSMnx. *Geogr. Anal.* **2025**, *57*, 567–577. [CrossRef]
70. Nahvi, H.; Fouladi, M.H.; Nor, M.M. Evaluation of whole-body vibration and ride comfort in a passenger car. *Int. J. Acoust. Vib.* **2009**, *14*, 143–149. [CrossRef]
71. Bianchin, G.; Pasqualetti, F. Routing apps may cause oscillatory congestions in traffic networks. In *2020 59th IEEE Conference on Decision and Control (CDC)*; IEEE: New York, NY, USA, 2020; pp. 253–260.
72. Cabannes, T.; Sangiovanni, M.; Keimer, A.; Bayen, A.M. Regrets in routing networks: Measuring the impact of routing apps in traffic. *ACM Trans. Spat. Algorithms Syst.* **2019**, *5*, 1–19.
73. Denstadli, J.M.; Jacobsen, J.K.S. The long and winding roads: Perceived quality of scenic tourism routes. *Tour. Manag.* **2011**, *32*, 780–789. [CrossRef]

74. Mokhtarian, P.L.; Salomon, I. How derived is the demand for travel? Some conceptual and measurement considerations. *Transp. Res. Part A Policy Pract.* **2001**, *35*, 695–719. [CrossRef]
75. Rahmad Ramadhan, L.; Anne Mudya, Y. A comparative study of Z-score and min-max normalization for rainfall classification in Pekanbaru. *J. Data Sci.* **2024**, *2024*, 1–8. [CrossRef]
76. Vamplew, P.; Dazeley, R.; Berry, A.; Issabekov, R.; Dekker, E. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Mach. Learn.* **2011**, *84*, 51–80. [CrossRef]
77. Altman, E. *Constrained Markov Decision Processes*; Routledge: London, UK, 2021.
78. Kushwaha, A.; Ravish, K.; Lamba, P.; Kumar, P. A survey of safe reinforcement learning and constrained mdps: A technical survey on single-agent and multi-agent safety. *arXiv* **2025**, arXiv:2505.17342.
79. Marcucci, E.; Gatta, V. Quality and public transport service contracts. *Eur. Transp.* **2007**, *36*, 92–106.
80. Sohail, M.; Maunder, D.A.C.; Cavill, S. Effective regulation for sustainable public transport in developing countries. *Transp. Policy* **2006**, *13*, 177–190. [CrossRef]
81. Achiam, J.; Held, D.; Tamar, A.; Abbeel, P. Constrained policy optimization. In *International Conference on Machine Learning*; PMLR: Sydney, Australia, 2017; pp. 22–31.
82. Stooke, A.; Achiam, J.; Abbeel, P. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*; PMLR: Sydney, Australia, 2020; pp. 9133–9143.
83. Silva, G.D. Analysis and Improvements of Multi Objective Reinforcement Learning Algorithms Based on Pareto Dominating Policies. Bachelor's Thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, 2023.
84. Van Moffaert, K.; Nowé, A. Multi-objective reinforcement learning using sets of pareto dominating policies. *J. Mach. Learn. Res.* **2014**, *15*, 3483–3512.
85. Nachum, O.; Norouzi, M.; Xu, K.; Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
86. Zhang, K.; Yang, Z.; Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control*; Springer: Cham, Switzerland, 2021; pp. 321–384.
87. Wang, Y.; Han, B.; Wang, T.; Dong, H.; Zhang, C. Dop: Off-policy multi-agent decomposed policy gradients. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
88. Ghavamzadeh, M.; Mahadevan, S.; Makar, R. Hierarchical multi-agent reinforcement learning. *Auton. Agents Multi-Agent Syst.* **2006**, *13*, 197–229. [CrossRef]
89. Sutton, R.S.; Precup, D.; Singh, S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.* **1999**, *112*, 181–211. [CrossRef]
90. Levine, S.; Kumar, A.; Tucker, G.; Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv* **2020**, arXiv:2005.01643.
91. Dulac-Arnold, G.; Levine, N.; Mankowitz, D.J.; Li, J.; Paduraru, C.; Gowal, S.; Hester, T. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Mach. Learn.* **2021**, *110*, 2419–2468. [CrossRef]
92. Kumar, A.; Zhou, A.; Tucker, G.; Levine, S. Conservative q-learning for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1179–1191.
93. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
94. Lopez, P.A.; Behrisch, M.; Bieker-Walz, L.; Erdmann, J.; Flötteröd, Y.P.; Hilbrich, R.; Lucken, L.; Rummel, J.; Wagner, P. Microscopic traffic simulation using sumo. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*; IEEE: New York, NY, USA, 2018; pp. 2575–2582.
95. Budiyono, A. Principles of GNSS, inertial, and multi-sensor integrated navigation systems. *Ind. Robot An. Int. J.* **2012**, *39*. [CrossRef]
96. Petit, J.; Shladover, S.E. Potential cyberattacks on automated vehicles. *IEEE Trans. Intell. Transp. Syst.* **2014**, *16*, 546–556. [CrossRef]
97. Pinto, L.; Davidson, J.; Sukthankar, R.; Gupta, A. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*; PMLR: Sydney, Australia, 2017; pp. 2817–2826.
98. Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*; PMLR: Sydney, Australia, 2019; pp. 7472–7482.
99. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2017; pp. 23–30.
100. Dietvorst, B.J.; Simmons, J.P.; Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. General.* **2015**, *144*, 114. [CrossRef] [PubMed]
101. Batool, Z.; Younis, M.W.; Yasir, A.; Rehman, A.U.; Dilawar, M.; Yasin, M.; Hamza, M.; Shahzad, S.; Ali, M.S.; Jamil, A.; et al. Effects of safety pattern, cabin ergonomics, and sleep on work-related stress and burnout of city and transit bus drivers in Lahore, Pakistan. *Ergonomics* **2022**, *65*, 704–718. [CrossRef] [PubMed]

102. De Waard, D. The Measurement of Drivers' Mental Workload. Ph.D. Thesis, University of Groningen, Haren, The Netherlands, 1996.
103. Kompier, M.A.; Di Martino, V. Review of bus drivers' occupational stress and stress prevention. *Stress Med.* **1995**, *11*, 253–262. [CrossRef]
104. Parasuraman, R.; Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors* **1997**, *39*, 230–253. [CrossRef]
105. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef] [PubMed]
106. Ben-Akiva, M.; McFadden, D.; Train, K.; Walker, J.; Bhat, C.; Bierlaire, M.; Bolduc, D.; Boersch-Supan, A.; Brownstone, D.; Bunch, D.S.; et al. Hybrid choice models: Progress and challenges. *Mark. Lett.* **2002**, *13*, 163–175. [CrossRef]
107. Chorus, C.G.; Arentze, T.A.; Timmermans, H.J. A random regret-minimization model of travel choice. *Transp. Res. Part B Methodol.* **2008**, *42*, 1–18. [CrossRef]
108. Ziebart, B.D.; Maas, A.L.; Bagnell, J.A.; Dey, A.K. Maximum entropy inverse reinforcement learning. *Aaai* **2008**, *8*, 1433–1438.
109. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565.
110. Skalse, J.; Howe, N.; Krasheninnikov, D.; Krueger, D. Defining and characterizing reward gaming. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9460–9471. [CrossRef]
111. Du, Y.; Watkins, O.; Wang, Z.; Colas, C.; Darrell, T.; Abbeel, P.; Gupta, A.; Andreas, J. Guiding pretraining in reinforcement learning with large language models. In *International Conference on Machine Learning*; PMLR: Sydney, Australia, 2023; pp. 8657–8677.
112. Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; Julian, R.; et al. Do as i can, not as i say: Grounding language in robotic affordances. In Proceedings of the Conference on Robot Learning, Atlanta, GA, USA, 6–9 November 2023; pp. 287–318.
113. Gong, R.; Huang, Q.; Ma, X.; Noda, Y.; Durante, Z.; Zheng, Z.; Terzopoulos, D.; Fei-Fei, L.; Gao, J.; Vo, H. Mindagent: Emergent gaming interaction. *Find. Assoc. Comput. Linguist. NAACL* **2024**, *2024*, 3154–3183. [CrossRef]
114. Fuller, A.; Fan, Z.; Day, C.; Barlow, C. Digital twin: Enabling technologies, challenges and open research. *IEEE Access* **2020**, *8*, 108952–108971. [CrossRef]
115. Lim, W.Y.B.; Luong, N.C.; Hoang, D.T.; Jiao, Y.; Liang, Y.C.; Yang, Q.; Niyato, D.; Miao, C. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2031–2063. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.