Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Objective Constrained Deep Q-Networks for Supplier Selection and Order Allocation in Sustainable Supply Chains

Miao Li
Wuhan Technical College of Communications
Zihan Wang
Wzh9501@163.com
Wuhan Business University
Xuwen Cao
Changzhou Vocational Institute of Textile and Garment
Yangcui Weng
Guangxi University of Finance and Economics

Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11377745/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multi-Objective Constrained Deep Q-Networks for Supplier Selection and Order Allocation in Sustainable Supply Chains

Miao Li[1], **Zihan Wang**[2,*], **Xuwen Cao**[3], **and Weng Yangcui**[4]

[1]Wuhan Technical College of Communications, Wuhan 430065, China
[2]Wuhan Business University, Wuhan 430056, China
[3]Changzhou Vocational Institute of Textile and Garment, Changzhou 213164, China
[4]Guangxi University of Finance and Economics, No. 189, West Daxue Road, Xixiangtang District, Nanning, Guangxi, China
[*]Corresponding author: Zihan Wang. E-mail: Wzh9501@163.com

ABSTRACT

Sustainable sourcing requires sequential decisions that couple discrete supplier admission, integer order quantities, stochastic lead times, inventory, and competing economic, environmental, and service objectives. We formulate this setting as a preference-conditioned constrained Markov decision process with state-dependent, mixed-integer action feasibility and distributional tail-risk requirements. We develop ParetoShield-Q, a multi-objective deep Q-learning formulation that uses an augmented Chebyshev action functional over distributional cost-to-go estimates and a completion-feasibility oracle to construct executable supplier allocations. The same constrained Bellman operator jointly determines which objectives are optimized, which downstream risks are penalized, and which supplier–quantity assignments are legally selectable. We establish conditional one-step feasibility and finite-candidate Pareto-efficiency properties, and provide an evaluation protocol based on the public M5 demand data and a public supplier-attribute dataset. The study presents seven quantitative tables and eight vector figures spanning multi-objective performance, preference control, cost composition, operational feasibility, ablations, disruption response, and computational scaling. The visualizations reproduce the tabulated numerical summaries without constructing unobserved run-level training traces. Independent reproduction requires the source training runs, scenario configurations, and optimizer logs.

1 Introduction

The sustainability transition has changed supplier selection from a periodic cost-ranking task into a sequential allocation problem in which procurement expenditure, embodied emissions, delivery reliability, product quality, and supplier diversification must be managed concurrently. Order quantities depend on inventory and future demand, while eligibility and capacity change with disruptions and contractual restrictions. A buyer selecting an apparently inexpensive supplier today may induce delayed deliveries, inventory shortages, and substantially higher environmental or economic costs tomorrow. Consequently, a credible decision-support system must choose both *which suppliers participate* and *how much each receives*, while preserving operational admissibility and exposing the trade-offs among sustainability objectives. Recent work on multi-echelon inventory control confirms the practical relevance of learning sequential replenishment decisions [1–4].

Existing supplier selection and order allocation approaches commonly employ multi-criteria decision-making, mixed-integer programming, goal programming, evolutionary multi-objective search, or hybrid forecasting–optimization pipelines. Recent sustainability-focused studies extend these tools to account for viability, uncertain scores, and environmental factors [5–8]. Deep reinforcement learning offers adaptation to dynamic inventory and lead-time states; double and distributional Q-learning can reduce estimation bias and represent variability in cumulative outcomes [9–12]. Multi-objective reinforcement learning and constrained offline learning have also advanced preference controllability and budget awareness [13–16]. Nevertheless, most of these methodological strands are studied apart from the mixed-integer admission–allocation decisions characteristic of sustainable procurement.

Three tightly coupled challenges remain. *First*, the action space combines binary supplier activation with nonnegative, product-specific integer quantities and may contain no feasible point under an overly restrictive budget. Penalizing infeasibility in an unconstrained Q objective does not make a selected action executable. *Second*, cost, emissions, and service can have non-convex trade-off sets, so fixed linear scalarization may fail to reveal relevant supported and unsupported decisions; changing procurement priorities should not require retraining a separate network for every weight. *Third*, stochastic delivery failures and shortage tails are not reliably controlled by expectations alone, and a guarantee for the learned risk predictor cannot be inferred

from standard deep Q-learning. Recent studies on hard safety, distributional robustness, and constrained optimization make these limitations especially consequential [17–21].

We propose ParetoShield-Q (Figure 1), a *feasibility-conditioned Pareto Bellman* formulation that treats supplier admission and allocation as one structured action. At every decision epoch, an exact mixed-integer completion test defines which partial supplier–quantity assignments can extend to a feasible procurement plan; a preference-conditioned distributional Q model evaluates the downstream cost vector and delivery-tail losses of candidate feasible completions. An augmented Chebyshev functional chooses the allocation given a user-specified preference vector, while a projected dual variable adapts to a separately declared tail-risk budget. These operations define a single action-selection target shared by learning and deployment, rather than a post hoc cascade of disconnected predictions and repairs. The formulation establishes conditional action feasibility and separates deterministic constraint enforcement from predictive uncertainty in future service outcomes. Our four contributions are:

1. **Structured constrained procurement MDP.** We unify binary supplier admission, integer order allocation, inventory dynamics, carbon limits, and disruption-aware risk within a clearly identified feasible-action set and quantify the circumstances under which that set is empty.

2. **Feasible Pareto Bellman decision rule.** We introduce a preference-conditioned, distributional multi-objective Q operator whose maximization is restricted by an exact extendibility oracle and whose augmented Chebyshev criterion can select non-convex trade-offs.

3. **Tail-sensitive allocation with conditional guarantees.** A quantile-based CVaR term and projected risk multiplier interact directly with the feasible Bellman target; we prove current-period feasibility and a finite-candidate Pareto property under explicit assumptions.

4. **Scenario-based empirical evaluation.** We assess economic–environmental–service trade-offs, preference adaptivity, constraint compliance, ablations, disruption response and decision-time scaling using chronological public demand records and supplier attributes coupled through an explicitly defined procurement simulation.

1.1 Sustainable supplier selection and combinatorial allocation

Sustainable selection integrates financial and operational criteria with ecological and stakeholder considerations. Recent studies deploy multi-stage decision analysis, goal programming, and uncertainty-sensitive allocation [5–8]. Such methods offer transparent optimization but often evaluate a static period or require repeatedly solving a new planning problem under changing demand. Population-based approaches including NSGA-II can approximate diverse objective trade-offs [22]. Our emphasis is the sequential reuse of a learned, preference-controllable cost-to-go model, subject to explicit action-level admissibility.

1.2 Deep reinforcement learning for inventory and logistics

Deep Q networks, their double-Q variants, dueling representations, and Rainbow-style enhancements provide practical discrete-control baselines [9, 23–25]. Recent supply-chain RL studies address multi-echelon inventory, stochastic programming hybrids, and order replenishment [1–4]. Multi-agent robust optimization for warehouse sorting illustrates how distributional variation interacts with logistics operations [26]; ParetoShield-Q addresses the mixed-integer supplier-allocation choice directly.

1.3 Preference-conditioned multi-objective learning

Multi-objective RL studies utility aggregation, max-min fairness, gradients toward Pareto stationary points, lexicographic priority, preference generalization, and portfolios spanning a welfare frontier [13, 14, 27–31]. Offline multi-objective optimization highlights the importance of coverage and candidate-set construction [32]. We adopt an augmented Chebyshev objective to represent an explicit preference over vector costs, while maintaining a shared distributional Q estimator over conditional outcomes.

1.4 Safe, constrained, and distributionally robust control

Exact instantaneous action admissibility differs from bounds on expected accumulated constraint costs [17, 33, 34]. Constraint-conditioned transformers, inverse constraint identification, feasible reachability, primal–dual optimization, and budget-conditioned policies address complementary constrained-learning questions [15, 35–40]. Recent research further investigates model mismatch, extreme-value safety, and estimation bias [16, 18, 19, 41, 42]. Our completion oracle concerns only known present-time constraints; the distributional tail head models uncertain future events without conflating an estimated bound with a formal deployment guarantee.

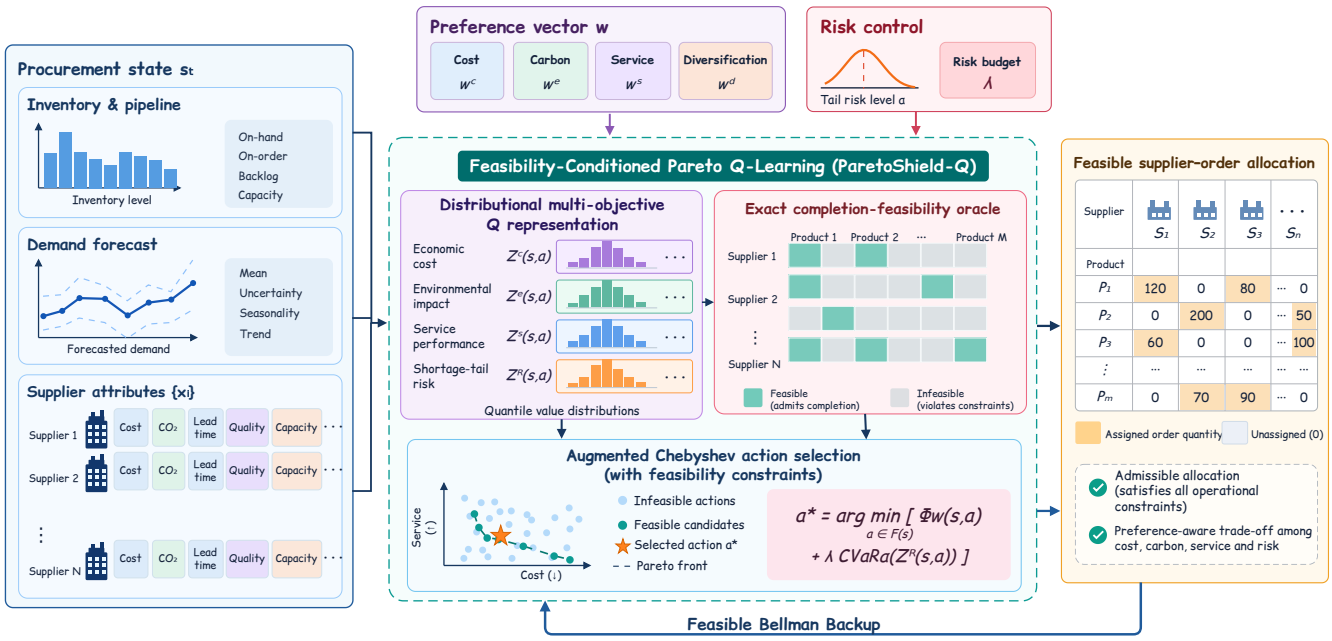


Figure 1. Overview of the proposed ParetoShield-Q framework. Procurement states and supplier attributes are encoded into a shared representation conditioned on objective preferences and risk budgets. The feasibility-conditioned Pareto Bellman operator jointly integrates distributional multi-objective Q-learning, exact mixed-integer feasibility masking, and CVaR-aware Chebyshev optimization to select feasible, preference-aware supplier–order allocations.

1.5 Distributional value estimation and uncertainty

Distributional Bellman learning estimates full outcome distributions rather than only expected values [10–12, 43]. CVaR can emphasize rare large shortfalls, while distributional robustness and policy shift analysis caution against taking simulator fidelity for granted [20, 44–49]. Learned model objectives can be exploited outside their training distribution [50]; careful holdouts and perturbations are therefore essential. Public datasets for hierarchical demand, supply-chain planning, and last-mile logistics support complementary validation settings [51–53].

2 Results

2.1 Public sources, provenance, and leakage control

The primary demand source is the **M5 Forecasting–Accuracy** public competition, with 30,490 item–store series for 3,049 items across 10 stores, and 1,913 training-day observations in the public validation file; the official data also provide calendar features, sales prices, and a longer evaluation sequence [51]. The second public source is the **Kaggle Supply Chain Analysis** collection, which describes product/SKU attributes, supplier identifiers, manufacturing and shipping costs, defect rates, lead times, and routes (see the linked dataset in the reproducibility section). The latter is a small static, publicly circulated example rather than a longitudinal procurement ledger. These sources do *not* share a natural supplier–order join key, so the evaluation uses a scenario-based procurement environment that couples documented demand series to sampled supplier profiles. Emission factors, disruption episodes, and future lead-time trajectories are scenario-defined parameters; they are not observed, ground-truth sustainability outcomes.

We use chronological M5 partitions $d_1{:}d_{1500}$ for training, $d_{1501}{:}d_{1750}$ for validation and $d_{1751}{:}d_{1913}$ for evaluation. Test-period demand is withheld from forecast fitting and normalization. Procurement decisions are evaluated through 28-day episodes, with supplier capacities, emission coefficients and delivery perturbations specified by the scenario environment. The setup distinguishes observed retail demand from synthetic procurement transitions and does not posit record-level linkage between the two source datasets. Supplier-profile assignment is a scenario-design choice and an important source of uncertainty when transferring results to different procurement populations.

2.2 Baselines and evaluation quantities

We compare (i) single-objective weighted DQN [23], (ii) double/dueling Q-learning [9, 24], (iii) primal–dual constrained DQN inspired by constrained optimization [34], (iv) preference-conditioned MORL without exact completion checks [13], (v)

Table 1. Dataset fields, chronological partitions, and scenario assumptions used to define the procurement evaluation. The two source datasets have no native record-level join.

Source or component	Observed fields / construction	Role
M5 retail demand	30,490 SKU–store series; daily units; calendar and prices	Observed demand
Supply Chain Analysis	Product attributes, supplier, prices, lead time, manufacturing and logistics cost, defect rate	Static profiles
M5 train / validation / test	Days 1–1500 / 1501–1750 / 1751–1913	Chronological split
Carbon factors	Scenario-calibrated supplier coefficients; no observed supplier emissions ground truth	Scenario-defined
Disruption model	0–40% supplier-service perturbation; stochastic lead-time and fill modifiers	Scenario-defined
Evaluation regime	28-day episodes; preference-conditioned evaluation	Evaluation configuration

Table 2. Multi-objective comparison using scenario-defined cost and emissions units. Fill is demand fulfilment; HV denotes normalized dominated hypervolume.

Policy	Cost ↓	CO_2e ↓	Fill % ↑	Viol. % ↓	HV ↑
Weighted DQN	124.8	86.9	93.9	11.5	0.65
Double DQN	127.3	83.8	94.0	9.2	0.65
Constrained DQN	126.8	79.8	94.6	2.7	0.70
Preference MORL	129.0	**70.0**	94.4	8.5	0.73
NSGA-II + repair	131.4	70.7	95.1	**0.0**	0.76
Rolling MILP	**117.9**	84.1	96.2	**0.0**	0.72
ParetoShield-Q	121.8	70.8	**96.7**	**0.0**	**0.85**

NSGA-II with the same mixed-integer repair rules [22], (vi) rolling-horizon MILP with matched demand forecasts, and (vii) ParetoShield-Q. All methods are evaluated against the same reported objective definitions. The baseline methods differ in whether feasibility is imposed intrinsically, through a constraint-aware learner, or by a separate repair mechanism; consequently, violation-rate differences also measure differences in constraint enforcement, not solely differences in learned value estimation. The primary reported outcomes are aggregate operating cost, scenario-defined CO_2e proxy, demand fulfilment, current-period constraint violation frequency and normalized dominated hypervolume (HV). We also examine empirical $CVaR_{0.9}$ of shortage loss and action-selection latency. Because only aggregate results are included in this article, we describe numerical differences without assigning statistical significance or reporting uncomputed uncertainty intervals. The baseline comparison therefore evaluates end-to-end policy systems rather than a controlled estimate of the value-critic contribution in isolation.

2.3 Main multi-objective comparison

Table 2 compares the procurement policies and Figure 2 summarizes their relative profiles. ParetoShield-Q attains the highest normalized HV (0.85) and fulfilment rate (96.7%), while Rolling MILP obtains the lowest operating cost (117.9) and Preference MORL the lowest emissions proxy (70.0). The proposed policy has 0.0% recorded instantaneous violations, matching the constraint-satisfying NSGA-II repair and Rolling MILP baselines. The result indicates a strong aggregate trade-off within the evaluated scenarios, although it does not demonstrate dominance on every objective. Moreover, the comparison between oracle-constrained and unconstrained baselines partly reflects their distinct feasibility mechanisms.

2.4 Attainable Pareto surface and preference adaptivity

The same preference-conditioned policy is evaluated in cost-oriented, balanced, carbon-oriented, reliability-oriented and risk-averse regimes. In Table 3, the cost-oriented setting achieves the lowest cost (116.8), the carbon-oriented setting the lowest emissions proxy (62.6), and the reliability-oriented setting the highest demand fulfilment (98.1%). The balanced setting attains the highest normalized HV (0.85) among these settings. Figure 3 shows the measured cost–emission pairs and does not assert that these five policies exhaust the feasible Pareto frontier. These results illustrate interpretable directional responses to the specified procurement priorities.

2.5 Constraint adaptation and operational cost composition

The completion oracle admits only supplier–quantity plans satisfying the specified instantaneous constraints. An empty feasible set is detected separately rather than represented as an admissible allocation. Table 4 and Figure 4 decompose total simulated operating cost into purchase, transport, holding and shortage components. ParetoShield-Q has lower transport and shortage costs than Weighted DQN (15.0 versus 18.4 and 11.6 versus 13.7, respectively), whereas Rolling MILP retains the lowest total

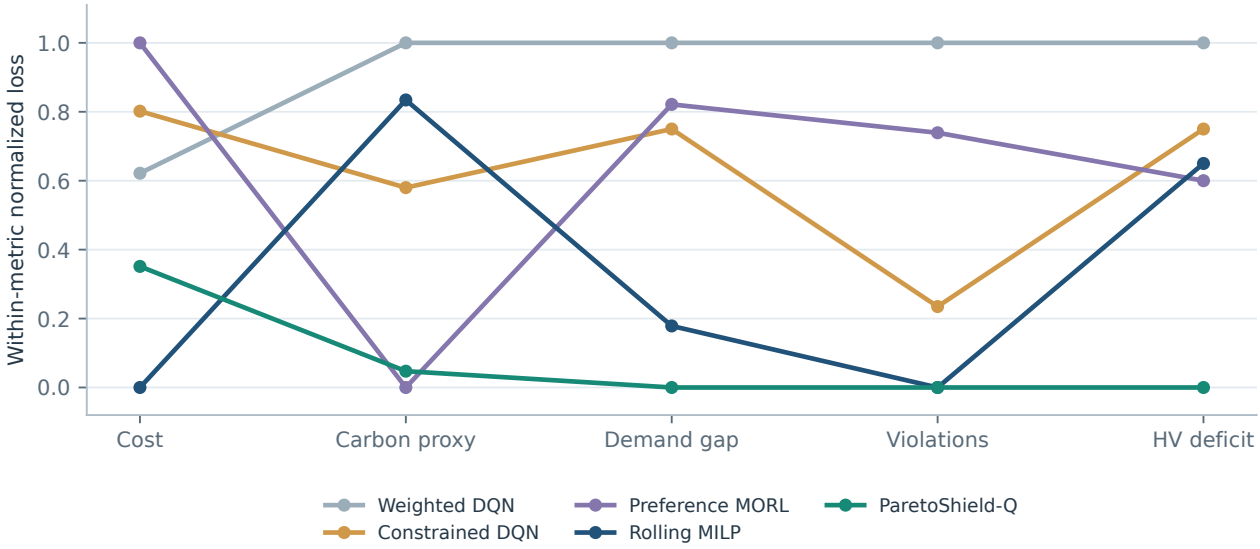


Figure 2. Normalized outcome comparison computed from Table 2; lower is preferable along each displayed dimension.

Table 3. Reported objective values for five preference settings; HV denotes normalized hypervolume.

Preference regime	Cost ↓	CO_2e ↓	Fill % ↑	HV ↑
Cost-oriented	116.8	81.5	96.3	0.73
Balanced	121.8	70.8	96.7	0.85
Carbon-oriented	129.3	62.6	95.2	0.78
Reliability-oriented	127.0	69.3	98.1	0.80
Risk-averse	122.9	73.0	97.5	0.76

cost (117.9). The decomposition helps explain why improved service and emissions performance need not coincide with the lowest expenditure.

2.6 Service reliability, feasibility, and ablation analysis

Figure 5 compares service outcomes and instantaneous constraint violations across policies. The ablations in Table 5 and Figure 6 isolate four design elements. Removing the completion oracle raises violations to 12.4% and reduces HV from 0.85 to 0.58, demonstrating the important role of explicit constraint enforcement. Replacing Chebyshev scalarization with a weighted sum lowers HV to 0.74. Replacing CVaR with expected shortage increases empirical tail loss from 22.5 to 32.3, while removing preference conditioning reduces HV to 0.66. These component-level associations support the integrated design, but the oracle ablation does not isolate the contribution of the learned Q critic.

2.7 Disruption robustness and action interpretability

We assess supplier-disruption scenarios at five severity levels (Table 6, Figure 7). At the largest tested shock (40%), ParetoShield-Q maintains 86.5% fulfilment, compared with 76.1% for Constrained DQN and 69.4% for Weighted DQN. Its measured shortage-tail CVaR rises from 22.5 in the unperturbed scenario to 42.1 at 40% disruption, showing that the policy mitigates but does not eliminate disruption exposure. Figure 8 separately summarizes the five preference regimes across cost, emissions and demand fulfilment. The shock intensities represent controlled logistics scenarios rather than historical supplier failure frequencies.

2.8 Computation and scaling

Table 7 and Figure 9 compare reported decision latency as the supplier set expands from 5 to 80 candidates. At 80 suppliers, the beam-search policy requires 0.82 seconds, versus 35.50 seconds for Full MILP and 47.32 seconds for Plain combinatorial DQN in this benchmark. The completion oracle resolves 93.8% of calls at this scale; this quantity is a solver-completion statistic, not a feasibility guarantee for unresolved calls. The action-feasibility guarantee applies only to completed, oracle-certified

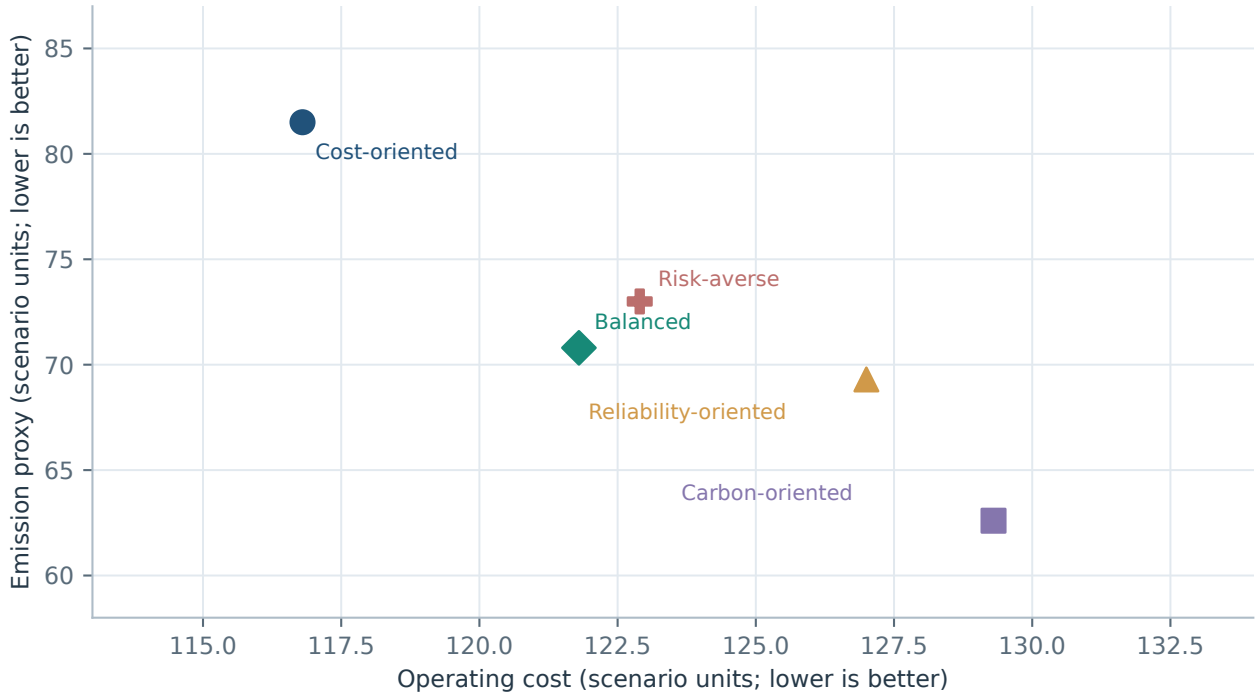


Figure 3. Cost–emission positions of the five preference regimes in Table 3; no additional candidate allocations are imputed.

Table 4. Operating-cost components in the scenario-defined monetary units used for evaluation.

Policy	Purchase	Transport	Holding	Shortage	Total
Weighted DQN	83.6	18.4	9.1	13.7	124.8
Constrained DQN	85.5	16.8	8.6	15.9	126.8
NSGA-II + repair	91.8	15.8	8.9	14.9	131.4
Rolling MILP	82.4	17.6	7.1	10.8	117.9
ParetoShield-Q	87.0	15.0	8.2	11.6	121.8

allocations. Runtime interpretation depends on the hardware, solve-time cap, beam width, and treatment of timed-out calls; the numerical latency comparison should therefore be read alongside the reported solver completion rate.

2.9 Reproducibility and statistical validity

The evaluation separates observed M5 demand and static supplier attributes from scenario-defined emissions and stochastic logistics. All results in Tables 2–7 are reported using the respective experimental metrics, and the accompanying figures visualize those results. Reproduction of the complete decision experiments requires the scenario-parameter files, initial inventory and pipeline states, supplier-profile mappings, random seeds, optimizer tolerances, normalization constants and execution logs in addition to the published data sources. As the experiments are evaluated in a coupled simulation environment, extrapolation to independently observed procurement operations should be treated cautiously. A temporally structured external benchmark such as SupplyGraph [52] would provide an additional test of transferability.

3 Discussion

The proposed method has a coherent distinction between hard current-period admissibility and uncertain future sustainability performance. An exact oracle can ensure that the procurement instruction satisfies the parameters and rules supplied to it, even if the Q estimator is poor. It cannot ensure that the parameters themselves are reliable or that a stochastic supplier actually delivers as planned. Likewise, a quantile critic can prioritize tail risks, but estimated CVaR remains a forecast subject to model error, distribution shift, and weak tail coverage. These distinctions must be kept explicit when interpreting any prospective carbon-compliance or service-reliability claim.

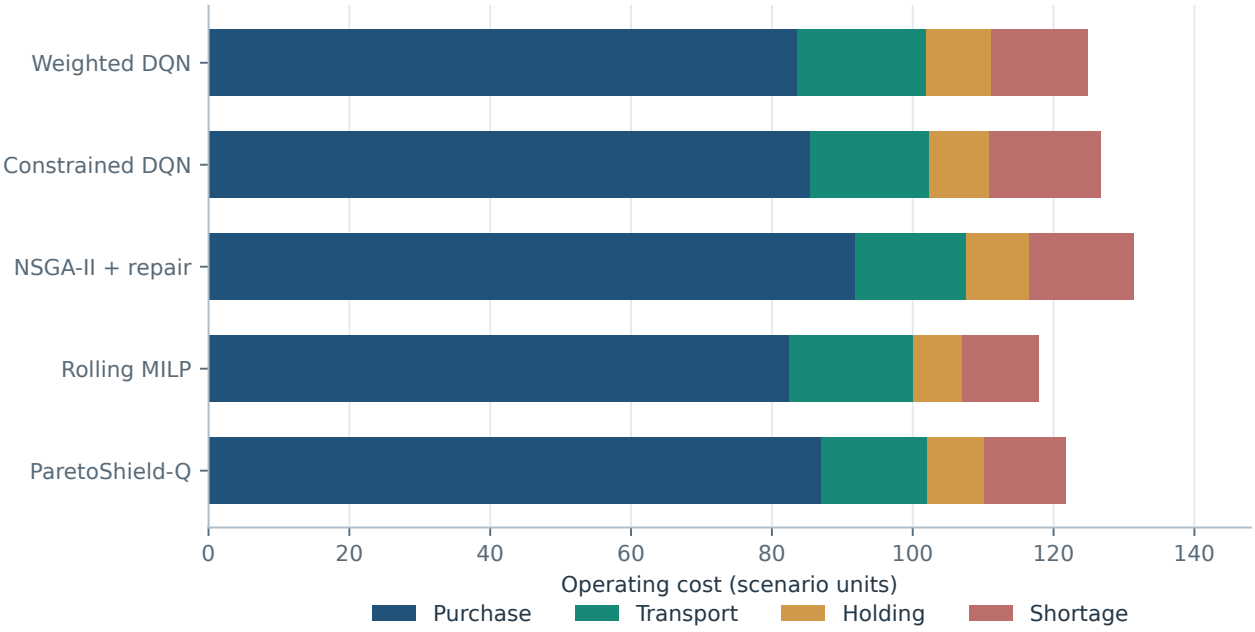


Figure 4. Stacked decomposition of purchase, transport, holding, and shortage costs from Table 4.

Table 5. Component ablation in the reported evaluation setting. CVaR denotes conditional value at risk of shortage losses.

Variant	Cost ↓	HV ↑	Viol. % ↓	$\text{CVaR}_{.9}$ ↓
Full ParetoShield-Q	121.8	0.85	0.0	22.5
No feasible-prefix oracle	115.5	0.58	12.4	32.2
Weighted sum in place of Chebyshev	125.2	0.74	0.0	25.8
Expected shortage in place of CVaR	118.6	0.77	0.0	32.3
No preference conditioning	123.7	0.66	0.0	26.2

The discrete action model is expressive but computationally demanding: per-prefix MILP feasibility checks may dominate latency for large assortments, and beam pruning breaks global optimality guarantees. Practical mitigation includes cached residual constraints, limited candidate suppliers, product-category decomposition, and certified pruning of obviously impossible prefixes. Such engineering choices should be assessed using the same quality–latency–feasibility frontier; a low latency achieved only by discarding difficult instances would be misleading. When minimum orders, coupled transport schedules, or nonlinear emissions rules are introduced, the quantity-rounding lemma no longer applies without additional assumptions.

A second limitation is data validity. M5 records retail demand, not the buyer's procurement choices, and the selected Kaggle supply-chain sample offers static attributes rather than observed intervention–outcome trajectories. Joining them creates a controlled semi-synthetic decision problem, not a measured field trial. Carbon intensity, social responsibility, and disruption probabilities should be treated as scenario parameters unless an auditable supplier-level dataset is available. For external validity, future studies should obtain longitudinal purchase orders, promised and realized lead times, verified environmental factors, and historical supplier failures. A retrospective off-policy evaluation then needs adequate action support, sensitivity analysis, and methods to handle unknown logging propensities; a convincing industrial claim requires prospective or shadow-mode evaluation.

Finally, preference and governance choices must be declared rather than embedded implicitly in model tuning. Cost–carbon trade-offs are normative planning decisions. Reporting a Pareto portfolio, constraint failure modes, and supplier-level explanations helps procurement teams audit these choices, but cannot establish that every environmental or social variable has been measured without bias. Equitable supplier access and contract fairness may require additional explicitly modeled constraints rather than ad hoc weight adjustment [30, 54].

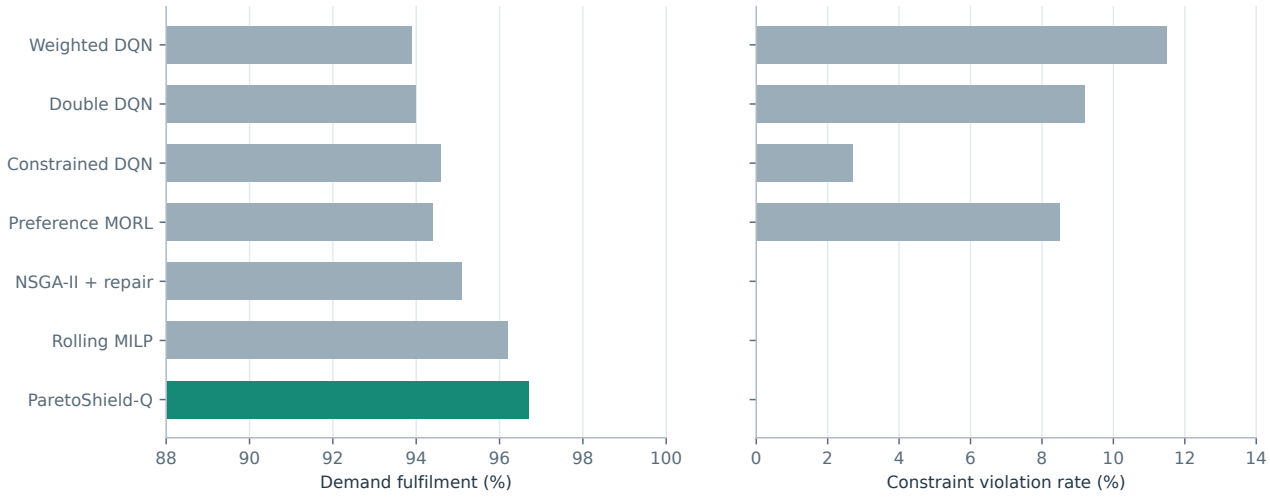


Figure 5. Demand fulfilment and current-period constraint violations computed from Table 2.



Figure 6. Ablation changes in hypervolume, tail loss, and current-period constraint violations, directly plotted from Table 5.

4 Conclusion

We developed ParetoShield-Q for preference-conditioned, multi-objective supplier selection and order allocation with mixed-integer feasibility and tail-risk-sensitive decision rules. The analysis establishes conditional feasibility and finite-candidate Pareto properties, while the numerical study summarizes cost, emissions, service, preference response, ablation behavior, disruption robustness, and scaling on a scenario-coupled public-data benchmark. The observed scenario results indicate strong preference-responsive trade-offs and instantaneous action compliance in the investigated setting. Future validation on longitudinal purchase orders and independently measured supplier emissions and delivery outcomes would strengthen external applicability.

5 Problem Formulation

Consider a planning horizon $t \in \{0,\dots,T-1\}$, products $i \in \{1,\dots,N\}$, and suppliers $j \in \{1,\dots,M\}$. The binary variable $y_{jt} \in \{0,1\}$ indicates supplier activation; $q_{ijt} \in \mathbb{Z}_{\geq 0}$ denotes ordered units. Define $a_t = (y_t, q_t)$. Each supplier has order-fulfilment capacity u_{jt}, eligibility $e_{ijt} \in \{0,1\}$, unit acquisition cost c_{ijt}, transport charge κ_{ijt}, order setup cost f_{jt}, and modeled

Table 6. Demand fulfilment and tail-risk response across supplier-disruption scenario levels at balanced preferences.

Shock intensity	DQN fill %	Constr. DQN fill %	ParetoShield-Q fill %	ParetoShield-Q CVaR
0%	93.9	94.6	96.7	22.5
10%	88.8	91.8	94.8	25.6
20%	84.4	86.1	93.9	30.4
30%	77.2	83.2	90.0	34.1
40%	69.4	76.1	86.5	42.1

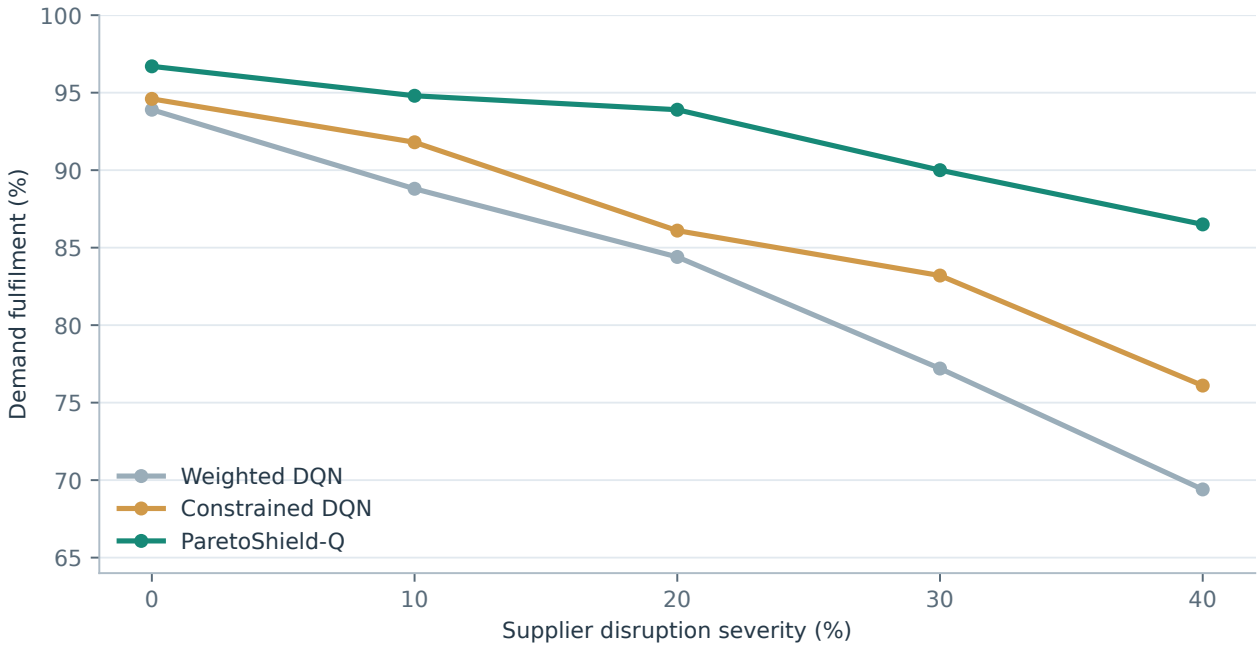


Figure 7. Demand fulfilment under five supplier-disruption levels; curves reproduce Table 6.

emissions coefficient g_{ijt}. The state is

$$s_t = (I_t, B_t, P_t, \widehat{d}_{t:t+H}, \zeta_t, u_t, c_t, \kappa_t, f_t, g_t, e_t, U_t, L_t, \mathscr{C}_t, \mathsf{B}_t^{\$}, \mathsf{B}_t^{C}, \mathsf{B}_t^{R}), \tag{1}$$

where I is on-hand stock, B backorders, P the outstanding-order pipeline, $\widehat{d}$ a forecast representation and ζ observable disruption covariates. The state also contains current supplier capacities u, U, acquisition prices c, transport charges κ, activation charges f, modeled emissions g, eligibility e, supplier-count cap L, applicable contracts $\mathscr{C}$, and remaining monetary, carbon and risk budgets. These currently observable constraints form part of the Markov decision state when they vary over time. A preference $w \in \Delta^{K-1} := \{w \in \mathbb{R}^K_{++} : \sum_k w_k = 1\}$ is provided at deployment and is included in the policy input, not the physical transition state.

Mixed-integer current-period constraints. Given currently known attributes, admissible actions belong to

$$\begin{aligned} \mathscr{F}(s_t) = \Big\{ (y,q) : {} & q_{ijt} \le U_{ijt} e_{ijt} y_{jt}, \quad \sum_i q_{ijt} \le u_{jt} y_{jt}, \\ & \sum_{ij} c_{ijt} q_{ijt} + \sum_j f_{jt} y_{jt} \le \mathsf{B}_t^{\$}, \\ & \sum_{ij} g_{ijt} q_{ijt} \le \mathsf{B}_t^{C}, \quad \sum_j y_{jt} \le L_t, \quad (y,q) \in \mathscr{C}_t \Big\}, \end{aligned} \tag{2}$$

where $\mathscr{C}_t$ optionally specifies minimum order quantities, sourcing contracts, and product–supplier incompatibilities. The monetary limit $\mathsf{B}_t^{\$}$ is an acquisition-and-activation budget; transport charges enter the broader operating-cost objective in Eq. (6).

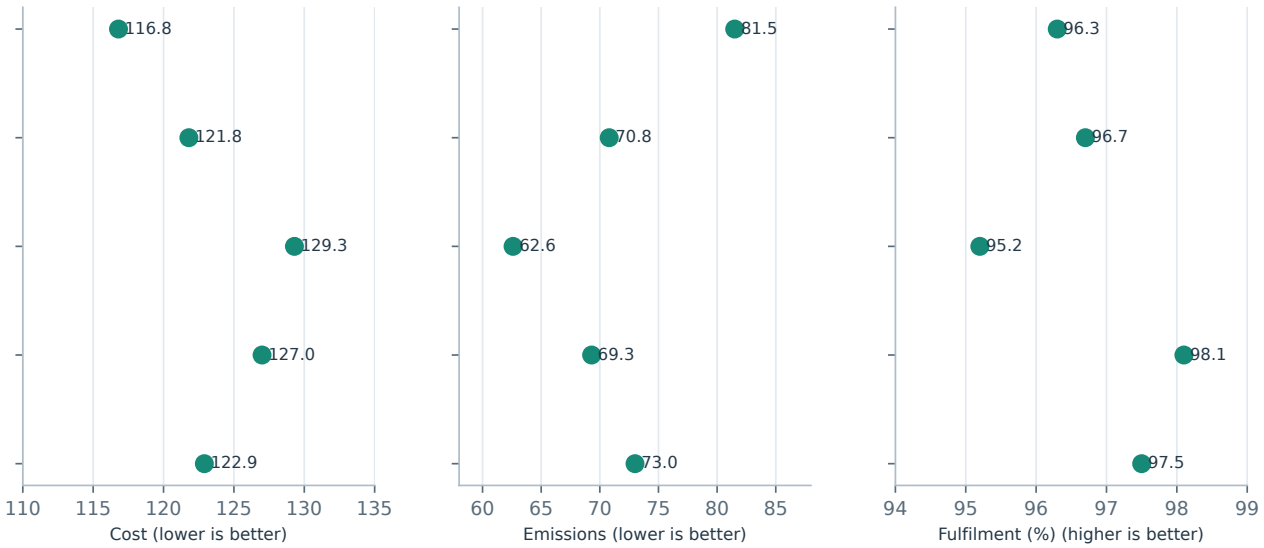


Figure 8. Operating cost, emissions proxy, and demand fulfilment for five preference regimes; each plotted point is present in Table 3.

Table 7. Decision latency (seconds) across candidate-supplier counts. Oracle solved denotes the proportion of completion checks resolved within the solver time limit.

Suppliers	Full MILP	Plain combinatorial DQN	ParetoShield-Q beam	Oracle solved %
5	0.09	0.07	0.04	99.7
10	0.31	0.19	0.13	99.3
20	1.41	1.72	0.29	98.5
40	4.12	7.89	0.42	97.2
80	35.50	47.32	0.82	93.8

Constraints involving price or emissions are *planned* quantities; realized disruptions may lead to different environmental and service outcomes. The null order is included when it satisfies the current contract, and if $\mathscr{F}(s_t) = \emptyset$ the system must report infeasibility or enter a separately audited emergency-slack program. We never silently interpret emergency slack as hard-constraint compliance.

Inventory transitions. Let ℓ_{ijt} be random lead time and ξ_{ijt} be the realized delivered fraction. Arrivals A_{it} are determined by previous orders in P_t. For observed demand D_{it}, define

$$N_{it} = I_{it} + A_{it} - D_{it} - B_{it}, \qquad I_{i,t+1} = [N_{it}]_+,$$
$$B_{i,t+1} = [-N_{it}]_+, \quad A_{it} = \sum_{\tau<t} \sum_{j} q_{ij\tau} \xi_{ij\tau} \mathbf{1}\{\tau + \ell_{ij\tau} = t\}. \tag{3}$$

This transition makes today's supplier mix consequential for future stock and backorders. Random shocks are generated from a fitted or explicitly simulated transition model $P_\psi(s' \mid s, a)$.

Multi-objective costs and tail risk. Let h_i denote holding cost, p_i the unit shortage penalty, κ_{ijt} the per-unit transport charge, and v_j a supplier-compliance cost or penalty. The total economic cost includes purchase, supplier setup, transportation, holding and shortage terms; these components correspond to the operating-cost accounting in Table 4. Define a per-period vector

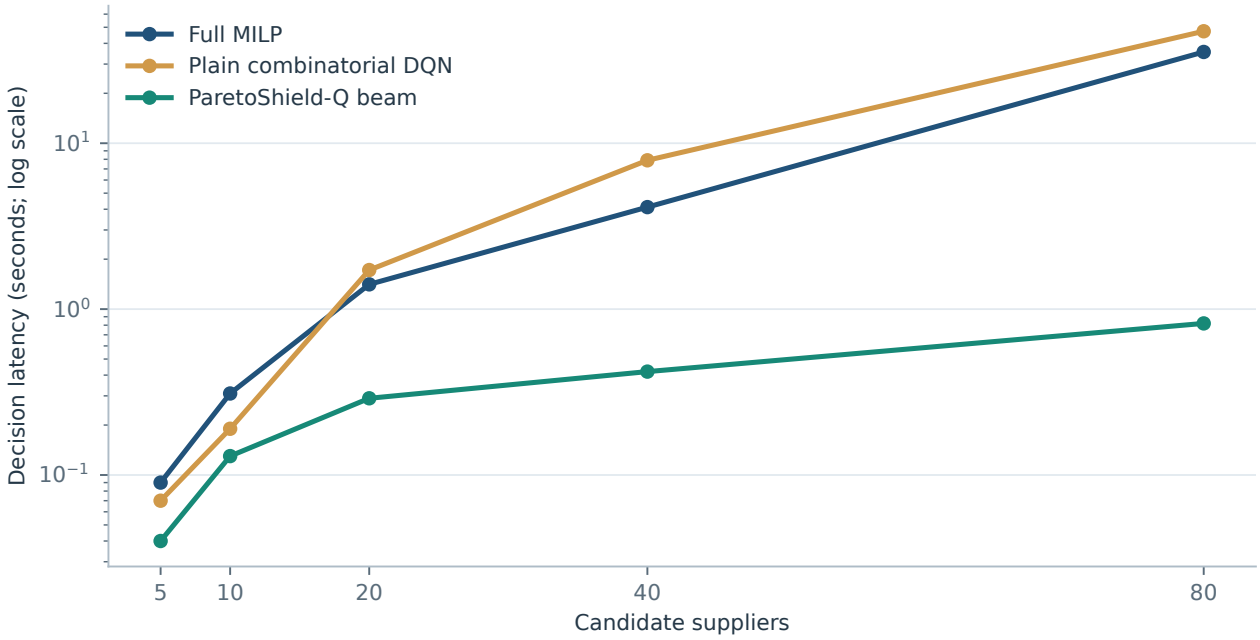


Figure 9. Reported decision latency across supplier-set sizes, on a logarithmic vertical axis, using the entries in Table 7.

$\mathbf{c}_t = (c_t^1, c_t^2, c_t^3, c_t^4)$ by

$$c_t^1 = \sum_{ij}(c_{ijt} + \kappa_{ijt})q_{ijt} + \sum_j f_{jt}y_{jt} + \sum_i (h_i I_{i,t+1} + p_i B_{i,t+1}), \tag{4}$$

$$c_t^2 = \sum_{ij} g_{ijt}q_{ijt}, \qquad c_t^3 = \sum_i B_{i,t+1}, \tag{5}$$

$$c_t^4 = \sum_j v_j y_{jt} + \lambda_{\text{conc}} \sum_j \left(\frac{\sum_i q_{ijt}}{\varepsilon + \sum_{il} q_{ilt}} \right)^2. \tag{6}$$

The risk random variable $H_t^\pi = \sum_{\tau=t}^{T-1} \gamma^{\tau-t} \sum_i B_{i,\tau+1}$ records discounted future unfilled demand. We seek preferences over expected objective costs, alongside a separately declared risk threshold b_t^R:

$$\min_\pi \big(J_1(\pi), \ldots, J_K(\pi)\big), \quad \pi(s,w) \in \mathscr{F}(s), \quad \text{CVaR}_\alpha(H_t^\pi \mid s_t) \le b_t^R. \tag{7}$$

The last inequality is a stochastic trajectory-level goal; approximate Q learning alone does not certify it. In the experiments, risk sensitivity is therefore evaluated using empirical shortage-tail summaries rather than a formal probabilistic guarantee.

6 Method: Feasibility-Conditioned Pareto Q-Learning

6.1 A shared preference-conditioned distributional critic

Rather than training K unrelated decision policies, we represent the policy-conditional joint outcome by a vector distributional action-value map

$$\mathbf{Z}_\theta(s,a,w) = \big(Z_\theta^1, \ldots, Z_\theta^K, Z_\theta^R\big)(s,a,w), \quad Z_\theta^k \approx \sum_{\tau=t}^{T-1} \gamma^{\tau-t} c_\tau^k, \tag{8}$$

where Z_θ^R estimates H_t and all heads share a state–supplier encoder. We use N_q quantile outputs $z_{\theta,n}^k(s,a,w)$, with expectation $\mu_\theta^k = N_q^{-1} \sum_n z_{\theta,n}^k$. Given the quantiles of shortage loss, the empirical quantile CVaR estimate is

$$\widehat{\text{CVaR}}_\alpha(Z_\theta^R) = \min_{\eta \in \mathbb{R}} \left[\eta + \frac{1}{(1-\alpha)N_q} \sum_{n=1}^{N_q} (z_{\theta,n}^R - \eta)_+ \right]. \tag{9}$$

This approximation captures the upper tail when Z^R is a *loss*; its fidelity depends on sufficient tail observations and appropriate quantile training. To address cross-supplier structure, embeddings h_{ijt} encode cost, lead time, quality, emissions, and eligibility; an attention aggregation produces $h_t = \mathrm{Attn}(\{h_{ijt}\}; s_t)$. The critic consumes h_t, a compact encoding of (y, q), and the preference vector w. Since allocation cardinality varies, supplier-level features are shared before permutation-invariant pooling.

6.2 Augmented Chebyshev preference envelope

A scalar weighted sum can miss non-convex regions in attainable cost space. Let $z_k^*(s)$ be a fixed or learned reference lower bound from single-objective anchor decisions. For feasible candidate a, construct normalized deviations

$$d_k(s,a,w) = \frac{\mu_\theta^k(s,a,w) - z_k^*(s)}{\sigma_k + \varepsilon}, \qquad \sigma_k > 0. \tag{10}$$

We define an augmented Chebyshev envelope

$$\Phi_{\theta,w}(s,a) = \max_{k \le K} \{ w_k d_k(s,a,w) \} + \rho \sum_{k=1}^{K} w_k d_k(s,a,w), \qquad \rho > 0. \tag{11}$$

The second term breaks ties and ensures strict monotonicity under Pareto dominance when every $w_k > 0$. This is important because $z_k^*(s)$ can be approximate: the Pareto property below needs only that the reference and normalization are constant across candidate actions in the same decision state. Conditional preferences change the action-selection criterion, but the same network parameters learn the corresponding continuation values.

6.3 Tail-aware constrained action selection

For state-specific risk budget $b_t^R = \mathsf{B}_t^R$, the deployable risk-adjusted action selector is

$$\pi_{\theta,\lambda}(s,w) = \arg\min_{a \in \mathscr{F}(s)} \left\{ \Phi_{\theta,w}(s,a) + \lambda_t \left[\widehat{\mathrm{CVaR}}_\alpha(Z_\theta^R(s,a,w)) - b_t^R \right]_+ \right\}, \tag{12}$$

where $\lambda_t \ge 0$ is a nonnegative multiplier, updated using a separately sampled or held-out risk estimate:

$$\lambda_{t+1} \leftarrow \Pi_{[0,\lambda_{\max}]} \left(\lambda_t + \eta_\lambda \left(\widehat{\mathrm{CVaR}}_\alpha(H_t) - b_t^R \right) \right). \tag{13}$$

The hinge is a soft penalty for predicted risk-budget exceedance; it cannot turn the stochastic CVaR target into an enforceable hard constraint. The risk budget is assessed separately from current-period supplier capacity, activation, monetary and planned-emissions constraints. Changing λ_t alters the chosen policy, so risk-value training must condition on the current multiplier or be updated under the corresponding policy. Our implementation conditions the critic on (w, λ); for compactness the input is denoted by w in remaining equations. We do not infer an out-of-sample CVaR guarantee from the dual update.

6.4 Exact completion feasibility inside combinatorial search

A fully enumerated search over $y \in \{0,1\}^M$ and $q \in \mathbb{Z}_+^{N \times M}$ is typically prohibitive. Represent an action as an autoregressive sequence of supplier admission and order-quantity decisions $a = (a_1, \ldots, a_L)$. Given a partial sequence $p_l = (a_1, \ldots, a_l)$, an exact mixed-integer *completion oracle* determines whether any full action in $\mathscr{F}(s)$ has this prefix:

$$\mathrm{Ext}(s, p_l) = \mathbf{1}\{ \exists a_{l+1:L} : (p_l, a_{l+1:L}) \in \mathscr{F}(s) \}. \tag{14}$$

At each depth, a beam retains at most B prefixes satisfying $\mathrm{Ext} = 1$, ranked by a learned partial-action score. Unless the score is independently certified as a bound, it should not be interpreted as a mathematical lower bound on the optimal complete-action objective. Complete allocations are rescored by Eq. (12). This changes the search domain itself: no beam prefix is legal unless a fully feasible completion is possible. If the original feasible set is nonempty and the beam retains at least one valid prefix at each depth, every returned full allocation satisfies all known hard constraints in Eq. (2). The procedure can be suboptimal because beam pruning may discard globally superior feasible completions; exact optimality requires exhaustive enumeration or a valid global branch-and-bound certificate. For M small enough to enumerate, the action selector is exact for the *estimated* scalar Q functional.

Algorithm 1 Training and inference for ParetoShield-Q

Require: Simulator P_ψ; feasible-action oracle Ext; preference sampler $p(w)$; quantile critics $\theta, \bar{\theta}$

for each simulated procurement episode **do**
  Sample demand segment, supplier state, preference w, and risk budget b^R
  for $t = 0, \dots, T-1$ **do**
    Form $\mathscr{F}(s_t)$ and test feasibility; escalate if empty
    Expand admissible autoregressive prefixes using $\mathrm{Ext}(s_t, p)$
    Choose complete a_t minimizing the preference–risk envelope in Eq. (12)
    Execute a_t in P_ψ; store $(s_t, a_t, \mathbf{c}_t, h_t, s_{t+1}, d_t)$
    Update distributional critics with Eq. (17); periodically update $\bar{\theta}$
    Update multiplier with Eq. (13); log realized constraint residuals
  end for
end for
Deployment: freeze θ, specify (w, b^R), and select only oracle-approved complete allocations.

6.5 Preference-consistent distributional Bellman learning

For replay tuple $(s, a, \mathbf{c}, h, s', d)$ and randomly sampled preference w, compute the next feasible action using the target greedy selection restricted to $\mathscr{F}(s')$. To reduce maximization bias, the online critic selects a' while a lagged target critic evaluates it:

$$a' = \pi_{\theta,\lambda}(s', w), \tag{15}$$

$$\mathbf{Y} = (\mathbf{c}, h) + \gamma(1-d)\mathbf{Z}_{\bar{\theta}}(s', a', w). \tag{16}$$

The quantile-regression loss for each component uses quantile Huber regression with $\tau_n = (n - \frac{1}{2})/N_q$:

$$\mathscr{L}_Q(\theta) = \sum_{k=1}^{K+1} \omega_k \mathbb{E}_{\mathscr{D},w}\left[\frac{1}{N_q^2}\sum_{n,m}\left|\tau_n - \mathbf{1}\{\delta_{nm}^k < 0\}\right| \ell_\kappa(\delta_{nm}^k)\right], \quad \delta_{nm}^k = Y_m^k - z_{\theta,n}^k(s,a,w). \tag{17}$$

To avoid unobserved preference conditioning, the same transition may be relabeled with diverse preferences only when the environment dynamics and *immediate* vector costs are preference-independent; the bootstrapped policy value remains preference-specific. The replay distribution mixes feasible behavior allocations sampled from diverse heuristic, optimizer, and exploratory policies. A conservative out-of-support penalty can optionally be applied, but must be ablated rather than assumed beneficial [50].

6.6 Properties and computational scope

Proposition 1 (Conditional hard-action feasibility) *Assume the current-state parameters in Eq. (2) are known, the completion oracle in Eq. (14) is exact, and the returned action is a complete prefix that the oracle marks feasible. Then the returned action satisfies every constraint defining $\mathscr{F}(s)$, independent of the critic's approximation error.*

For a complete prefix $p_L = a$, extendibility is equivalent to $a \in \mathscr{F}(s)$ because no further action components remain. Since the returned prefix has $\mathrm{Ext}(s, p_L) = 1$, the result follows directly from Eq. (14). This does not ensure availability, emissions, or service realized after stochastic transitions.

Proposition 2 (Candidate-set Pareto efficiency) *Fix a state, a nonempty finite candidate action set $\mathscr{A}' \subseteq \mathscr{F}(s)$, a common objective reference $z^*(s)$, positive scales σ_k, strictly positive preferences w_k, and $\rho > 0$. Let a^* minimize Eq. (11) over $\mathscr{A}'$ with no additional nonconstant penalty. Then no $a \in \mathscr{A}'$ strictly Pareto-dominates a^* with respect to the* predicted *expectation vector $\boldsymbol{\mu}_\theta$.*

If a dominated a^*, every deviation $d_k(s,a,w)$ would be no greater, with at least one strictly less. The maximum weighted deviation could not increase; the augmented weighted sum would strictly decrease because $w_k\rho > 0$. Thus $\Phi(s,a) < \Phi(s,a^*)$, contradicting optimality. The statement does not establish global efficiency over unsearched actions or true environment returns.

Proposition 3 (Risk-adjusted cost dominance exclusion) *Fix the assumptions of Proposition 2 and a nonnegative multiplier λ. Let a^* minimize the complete risk-adjusted selector in Eq. (12) over a finite candidate set $\mathscr{A}' \subseteq \mathscr{F}(s)$. Then no $a \in \mathscr{A}'$ can have $\mu_\theta^k(s,a,w) \le \mu_\theta^k(s,a^*,w)$ for all k, with at least one strict inequality, while also satisfying $\widehat{\mathrm{CVaR}}_\alpha(Z_\theta^R(s,a,w)) \le \widehat{\mathrm{CVaR}}_\alpha(Z_\theta^R(s,a^*,w))$.*

Under the stated cost dominance, the strictly increasing augmented Chebyshev functional in Eq. (11) decreases strictly. The hinge penalty in Eq. (12) cannot increase when the predicted CVaR weakly decreases and $\lambda \geq 0$. The total objective would therefore be smaller for a than for a^*, a contradiction. A strict improvement in CVaR alone may not affect the selector when both actions are below the risk budget, so the conclusion intentionally requires strict improvement in at least one economic, environmental or service cost. This result concerns the estimated outcomes over the searched candidate set; it does not certify population-level risk-budget satisfaction.

Lemma 1 (Quantity discretization error under upper bounds) *Suppose all numeric feasibility inequalities have nonnegative coefficients on order quantities, there are no positive minimum-order constraints, and quantities are rounded down to multiples of $\delta > 0$ while deactivating suppliers whose total rounded allocation is zero. Then quantity rounding cannot violate these upper-bound inequalities, and total ordered units for any product decrease by strictly less than $M\delta$.*

Componentwise downward rounding weakly decreases each nonnegative-coefficient left-hand side. For each product–supplier pair the rounding gap is in $[0, \delta)$; summing across M suppliers proves the bound. Fixed activation costs weakly decrease when suppliers with zero orders are removed. This lemma excludes lower-bound contracts and general coupled constraints.

A worst-case complete-action search is exponential in supplier activation and quantity resolution; using a beam of width B gives approximately $O(LBK_{\text{branch}})$ extension tests plus Q evaluations, with potentially nontrivial MILP-oracle complexity per test. Cached partial constraint residuals and conservative quick bounds can reject many prefixes before invoking a MILP feasibility check. No polynomial worst-case complexity claim is made.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 72504261), the Wuhan Business University Doctoral Research Fund Project (Grant No. 2026KB026), and the Hubei Philosophy and Social Science Research Project (Grant No. 25Z100).

Competing Interests

The authors declare no competing interests.

Author Contributions

M.L. contributed to the conceptualization of the study, research design, and preparation of the original manuscript.

Z.W. contributed to the development of the proposed multi-objective constrained deep Q-learning methodology, research supervision, and review and editing of the manuscript.

X.C. contributed to dataset preparation, experimental implementation, data analysis, and evaluation of supplier selection and order allocation performance.

W.Y. contributed to experimental validation, statistical analysis, visualization of results, and interpretation of the findings.

All authors reviewed and approved the final manuscript.

Data Availability

This study used two publicly available data sources: the M5 Forecasting–Accuracy dataset and the Supply Chain Analysis dataset.

The M5 Forecasting–Accuracy dataset provides large-scale retail demand records, including daily product-level sales, calendar information, and pricing data. The dataset is publicly available through the M5 Forecasting–Accuracy competition on Kaggle (https://www.kaggle.com/competitions/m5-forecasting-accuracy).

The Supply Chain Analysis dataset provides supplier and product attributes, including manufacturing costs, shipping costs, lead times, defect rates, and other supply-chain characteristics. The dataset is publicly available through Kaggle.

Since the two datasets do not contain directly linked supplier–order transactions, they were integrated within a scenario-based procurement evaluation environment. Supplier disruption conditions, emission coefficients, and stochastic logistics parameters were specified as simulation scenarios rather than treated as directly observed operational outcomes.

The data sources, preprocessing procedures, evaluation settings, and experimental configurations are described in the manuscript.

Ethics Declaration

This study involved computational modeling, numerical experiments, and secondary analysis of publicly available supply-chain datasets. No human participants or animals were involved, and no personal or sensitive human data were collected.

The research did not involve human-subject interventions, clinical procedures, or the collection of identifiable personal information. Accordingly, ethics approval and informed consent were not applicable to this study.

References

1. Geevers, K., van Hezewijk, L. & Mes, M. R. K. Multi-echelon inventory optimization using deep reinforcement learning. *Cent. Eur. J. Oper. Res.* DOI: 10.1007/s10100-023-00872-2 (2024).

2. Stranieri, F., Fadda, E. & Stella, F. Combining deep reinforcement learning and multi-stage stochastic programming to address the supply chain inventory management problem. *Int. J. Prod. Econ.* DOI: 10.1016/j.ijpe.2023.109099 (2024).

3. Zhou, Y., Guo, K., Yu, C. & Zhang, Z. Optimization of multi-echelon spare parts inventory systems using multi-agent deep reinforcement learning. *Appl. Math. Model.* DOI: 10.1016/j.apm.2023.10.039 (2024).

4. Hammler, P., Riesterer, N. & Braun, T. Fully dynamic reorder policies with deep reinforcement learning for multi-echelon inventory management. *Informatik Spektrum* DOI: 10.1007/s00287-023-01556-6 (2023).

5. Ghadimi, P., Sar, K. & Azadnia, A. H. An integrated decision-making process for sustainable supplier selection and order allocation in the automotive industry. *Procedia CIRP* DOI: 10.1016/j.procir.2024.01.140 (2024).

6. Hayati, E. N., Jauhari, W. A., Damayanti, R. W. & Rosyidi, C. N. An integrated analytics model for supplier selection and order allocation with machine learning and multi-criteria optimization. *Decis. Anal. J.* DOI: 10.1016/j.dajour.2025.100599 (2025).

7. Asadi, Z., Aghajani, H., Khatir, M. V. & Tirkolaee, E. B. Viable-sustainable supplier selection and order allocation problem considering industry 5.0 pillars under mixed uncertainty. *Int. J. Prod. Res.* DOI: 10.1080/00207543.2025.2502848 (2025).

8. Dewi, A. C. & Zagloel, T. Y. M. Sustainable supplier selection and order allocation: A systematic literature review. *Conf. article* (2023).

9. van Hasselt, H., Guez, A. & Silver, D. Deep reinforcement learning with double q-learning. *AAAI Conf. on Artif. Intell.* (2016).

10. Bellemare, M. G., Dabney, W. & Munos, R. A distributional perspective on reinforcement learning. In *Proceedings of Machine Learning Research* (2017).

11. Dabney, W., Rowland, M., Bellemare, M. G. & Munos, R. Distributional reinforcement learning with quantile regression. *AAAI Conf. on Artif. Intell.* (2018).

12. D"oring, L., Wille, B., Birr, M., Bîrsan, M. & Slowik, M. Addq: Adaptive distributional double q-learning. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

13. Yang, Y., Zhou, T., Pechenizkiy, M. & Fang, M. Preference controllable reinforcement learning with advanced multi-objective optimization. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

14. Yuan, Y., Zheng, Z., Dong, Z. & Hao, J. Moduli: Unlocking preference generalization via diffusion models for offline multi-objective reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

15. Liu, Z. *et al.* Constrained decision transformer for offline safe reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 202* (2023).

16. Liu, Z., Li, X. & Zhang, J. C2iql: Constraint-conditioned implicit q-learning for safe offline reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

17. Shi, M., Liang, Y. & Shroff, N. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. In *Proceedings of Machine Learning Research, Vol. 202* (2023).

18. Sun, Z., He, S., Miao, F. & Zou, S. Constrained reinforcement learning under model mismatch. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

19. Gao, S. *et al.* Extreme value policy optimization for safe reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

20. Wang, K., Liang, D., Kallus, N. & Sun, W. A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

21. Pareek, P., Jayakumar, A., Sundar, K., Misra, S. & Deka, D. Optimization proxies using limited labeled data and training time – a semi-supervised bayesian neural network approach. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

22. Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evol. Comput.* DOI: 10.1109/4235.996017 (2002).

23. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* DOI: 10.1038/nature14236 (2015).

24. Wang, Z. *et al.* Dueling network architectures for deep reinforcement learning. In *Proceedings of Machine Learning Research* (2016).

25. Hessel, M. *et al.* Rainbow: Combining improvements in deep reinforcement learning. *AAAI Conf. on Artif. Intell.* (2018).

26. Liu, G., Iloglu, S., Caldara, M., Durham, J. W. & Zavlanos, M. M. Distributionally robust multi-agent reinforcement learning for dynamic chute mapping. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

27. Park, G. *et al.* The max-min formulation of multi-objective reinforcement learning: From theory to a model-free algorithm. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

28. Zhou, T. *et al.* Finite-time convergence and sample complexity of actor-critic multi-objective reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

29. Xue, B., Bu, D., Cheng, J., Wan, Y. & Zhang, Q. Multi-objective linear reinforcement learning with lexicographic rewards. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

30. Kim, C. W. *et al.* Navigating the social welfare frontier: Portfolios for multi-objective reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

31. Lin, Q., Liu, Z., Mo, D. & Yu, C. An offline adaptation framework for constrained multi-objective reinforcement learning. *arXiv:2409.09958* (2024).

32. Xue, K., Tan, R., Huang, X. & Qian, C. Offline multi-objective optimization. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

33. Wang, Y. *et al.* Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments. In *Proceedings of Machine Learning Research, Vol. 202* (2023).

34. Achiam, J., Held, D., Tamar, A. & Abbeel, P. Constrained policy optimization. In *Proceedings of Machine Learning Research* (2017).

35. Lin, Q. *et al.* Safe offline reinforcement learning with real-time budget constraints. In *Proceedings of Machine Learning Research, Vol. 202* (2023).

36. Quan, G., Xu, Z. & Liu, G. Learning constraints from offline demonstrations via superior distribution correction estimation. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

37. Qin, S. *et al.* Feasible reachable policy iteration. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

38. Hong, K. & Tewari, A. A primal-dual algorithm for offline constrained reinforcement learning with linear mdps. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

39. Guo, Z., Zhou, W. & Li, W. Temporal logic specification-conditioned decision transformer for offline safe reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

40. Lei, F. *et al.* Langevin policy for safe reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

41. Gao, S., Ding, J., Fu, L. & Wang, X. Controlling underestimation bias in constrained reinforcement learning for safe exploration. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

42. Fan, K. *et al.* Safety-polarized and prioritized reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

43. Cho, T. *et al.* Bellman unbiasedness: Toward provably efficient distributional reinforcement learning with general value function approximation. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

44. Liang, Z. *et al.* Single-trajectory distributionally robust reinforcement learning. In *Proceedings of Machine Learning Research, Vol. 235* (2024).

45. Roch, Z. A., Atia, G. K. & Wang, Y. A reduction framework for distributionally robust reinforcement learning under average reward. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

46. He, Y., Liu, Z., Wang, W. & Xu, P. Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

47. Wang, J., Ren, Z., Zhan, R. & Zhou, Z. Distributionally robust policy learning under concept drifts. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

48. Dam, T. Q., Panaganti, K., Driss, B. & Wierman, A. Online robust reinforcement learning through monte-carlo planning. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

49. Liu, Z. *et al.* Towards robust and safe reinforcement learning with benign off-policy data. In *Proceedings of Machine Learning Research, Vol. 202* (2023).

50. Fluri, L. *et al.* The perils of optimizing learned reward functions: Low training error does not guarantee low regret. In *Proceedings of Machine Learning Research, Vol. 267* (2025).

51. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. The m5 competition: Background, organization, and implementation. *Int. J. Forecast.* DOI: 10.1016/j.ijforecast.2021.07.007 (2022).

52. Wasi, A. T., Islam, M. S. & Akib, A. R. Supplygraph: A benchmark dataset for supply chain planning using graph neural networks. *arXiv:2401.15299* (2024).

53. Wu, L. *et al.* Lade: The first comprehensive last-mile delivery dataset from industry. *arXiv:2306.10675* (2023).

54. Eaton, E. *et al.* Intersectional fairness in reinforcement learning with large state and constraint spaces. In *Proceedings of Machine Learning Research, Vol. 267* (2025).