Value Decomposition Fails in Anti-Coordination: A Systematic MARL Comparison for Dynamic Spectrum Access in Dense 6G Networks

Anonymous Authors[1]

Abstract

Dynamic spectrum access (DSA) requires secondary users to select *different* channels, making it an anti-coordination problem. We present a systematic MARL comparison for DSA, evaluating sixteen methods across seven paradigms. Value decomposition (VDN, QMIX, QPLEX) achieves only 0.34–0.43 bps/Hz, dramatically below Random (1.10 bps/Hz). This failure persists across all interventions tested: IGM-complete decomposition, Boltzmann execution, hyperparameter sweeps (optimizer, GRU size, buffer size), agent-ID symmetry breaking, increased replay ratios, PU sensing, extended training ($2.2\times$ budget), and idle actions. Full COMA with exact counterfactual marginalization—including a joint-action critic variant—performs *worse* than single-sample COMA-lite, suggesting gradient noise helps escape degenerate equilibria in anti-coordination. Policy gradient methods (IPPO, MAPPO) are the only learned methods to consistently match stochastic baselines across all densities.

1. Introduction

6G wireless systems must support unprecedented device densities (Letaief et al., 2019; Tataria et al., 2021). Dynamic spectrum access (DSA), where secondary users (SUs) opportunistically use bands vacated by primary users (PUs), is key to spectral efficiency (Samsung Research, 2020). Deep RL enables agents to learn channel access directly from interaction (Luong et al., 2019; Wang et al., 2018; Naparstek & Cohen, 2019), but realistic deployments require multi-agent coordination without explicit communication.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

MARL offers several paradigms: independent learning (Tan, 1993), value decomposition (VDN (Sunehag et al., 2018), QMIX (Rashid et al., 2018), QPLEX (Wang et al., 2021)), counterfactual credit assignment (COMA (Foerster et al., 2018)), and centralized-critic methods (MAPPO (Yu et al., 2022)). Value decomposition excels on cooperative benchmarks (Samvelyan et al., 2019) and wireless tasks with richer observations (Naderializadeh et al., 2021), but DSA is structurally different: agents must select *different* channels, making it an *anti-coordination* task where one agent's optimal action becomes suboptimal if others copy it.

We present the largest systematic comparison of MARL for DSA with the following contributions:

- Sixteen methods spanning seven paradigms (baselines, bandits, independent learners, value decomposition with ε-greedy and Boltzmann execution, counterfactual credit assignment with approximate and exact marginalization, centralized-critic PG), evaluated across five seeds with six metrics.
- A strong negative result: value decomposition underperforms Random at all densities tested ($N = 5$ to $N = 50$). Boltzmann execution does not close the gap (VDN-Boltz: 0.317, QMIX-Boltz: 0.484, QPLEX-Boltz: 0.288, all $\ll$ Random: 1.103), demonstrating the problem is value *representation*, not execution policy.
- Full COMA with exact counterfactual marginalization (0.151 ± 0.176) is worse than single-sample COMA-lite (0.493 ± 0.540), showing that reduced gradient noise can trap agents in degenerate equilibria.
- An analytical result (Remark 1) explaining the constant PU collision rate across all methods, plus twelve ablations.

2. Related Work

RL for spectrum access. Wang *et al.* (Wang et al., 2018) and Naparstek and Cohen (Naparstek & Cohen, 2019) demonstrated DQN-based multichannel access. Xu *et al.* (Xu et al., 2018; 2020) studied DRL under partial observability. These works evaluated only independent learners

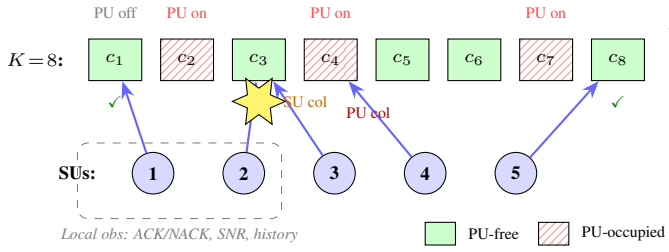

Figure 1. DSA environment schematic ($N = 5$, $K = 8$). SUs select channels with local observations only. A transmission succeeds only when the SU is alone on a PU-free channel (✓). Co-selecting agents cause SU collisions; selecting PU-occupied channels causes PU collisions.

without CTDE comparisons.

Multi-agent RL. Value decomposition factorizes joint Q-functions: VDN (Sunehag et al., 2018) (additive), QMIX (Rashid et al., 2018) (monotonic), QPLEX (Wang et al., 2021) (IGM-complete). COMA (Foerster et al., 2018) uses counterfactual baselines; MAPPO (Yu et al., 2022) uses centralized critics. Papoudakis *et al.* (Papoudakis et al., 2021) showed task structure strongly modulates relative performance.

Anti-coordination and bandits. Musical Chairs (Rosenski et al., 2016) and SIC-MMAB (Boursier & Perchet, 2019) address distributed arm differentiation. The monotonicity constraint in QMIX ($\partial Q_{\text{tot}}/\partial Q_n \geq 0$) conflicts with anti-coordination, where one agent's good action reduces others' values. Boltzmann exploration has been studied in multi-agent settings (Claus & Boutilier, 1998; Wei & Luke, 2016); we test it directly for value decomposition in DSA.

3. Problem Formulation and Methods

3.1. Multi-Agent DSA Environment

We model DSA as a Dec-POMDP with N SUs and K channels. Each channel's PU follows a two-state Markov chain ($p_{\text{on}} = 0.3$, $p_{\text{off}} = 0.5$, steady-state occupancy $\pi_{\text{on}} = 0.375$). Channel quality follows Rayleigh fading ($\gamma_k \sim \text{Exp}(\bar{\gamma})$, $\bar{\gamma} = 10$, noise $\sigma^2 = 0.1$). Each SU must select exactly one channel per timestep (no idle action). A transmission succeeds iff the SU is alone on a PU-inactive channel, yielding rate $r_n^t = \log_2(1 + \gamma_{a_n^t}^t/\sigma^2)$; otherwise $r_n^t = 0$. Each agent observes only local ACK/NACK, SNR, and action history ($W = 5$). We report six metrics: throughput (TP), collision rate, SU/PU collision rates, spectrum utilization, and Jain's fairness (Jain et al., 1984). Full environment details are in Appendix A.

Remark 1 (PU collision rate under rational agents)**.** *Under uniform channel selection,* $\mathbb{E}[\text{PU Col}] = \pi_{\text{on}} = p_{\text{on}}/(p_{\text{on}} + p_{\text{off}}) = 0.375$. *With perfect PU sensing, agents could target only* $K_{\text{free}} \sim$ *Binomial*$(K, 1 - \pi_{\text{on}})$ *PU-free channels (*$\mathbb{E}[K_{\text{free}}] = 5$*), achieving* PU Col $= 0$. *However, the probability of successful lone transmission drops:* $P(\text{alone}) = (4/5)^9 \approx 0.134$ *vs.* $P(\text{alone and free}) = 0.625 \times (7/8)^9 \approx 0.188$ *under uniform spreading. Since* $\mathbb{E}[\log_2(1+\text{SNR})] \approx 5.88$*, PU avoidance yields* 0.788 *bps/Hz vs. uniform's* 1.106 *bps/Hz—a 28.8% throughput reduction. Rational agents therefore do not avoid PU channels when* $\lambda = 0$.

3.2. Methods

We evaluate sixteen methods (details in Appendix A): **(1) Non-learning:** Random, Greedy (EMA + ε-greedy), S-ALOHA, p-CSMA. **(2) Bandits:** Musical Chairs (Rosenski et al., 2016), SIC-MMAB (Boursier & Perchet, 2019). **(3) Independent:** IDQN (Mnih et al., 2015), IDRQN (Hausknecht & Stone, 2015), IPPO (Schulman et al., 2017). **(4) VD ε-greedy:** VDN (Sunehag et al., 2018), QMIX (Rashid et al., 2018), QPLEX (Wang et al., 2021). **(5) VD Boltzmann:** VDN-Boltz, QMIX-Boltz, QPLEX-Boltz, using $\pi(a) = \text{softmax}(Q/\tau)$ with $\tau : 1.0 \to 0.1$. **(6) Counterfactual:** COMA-lite (single-sample), COMA-Full (exact marginalization over $K = 8$ actions: $b_i = \sum_a \pi_i(a) Q(s, (a_{-i}, a))_i$). **(7) Centralized-critic PG:** MAPPO (Yu et al., 2022). All CTDE methods share the same global state ($s \in \mathbb{R}^{2K+NK}$) during training. Default: 200 episodes $\times$ 120 steps, 5 seeds, GRU (48 hidden).

4. Experiments

4.1. Main Comparison ($N = 10$, $K = 8$)

Table 1 shows three clear performance tiers. **Tier 1:** Greedy (1.938 ± 0.014), exploiting channel quality. **Tier 2:** Stochastic baselines, bandits, and policy gradient ($\sim$1.10 bps/Hz). **Tier 3:** All value-based methods (≤ 0.60). The IPPO–QPLEX gap is significant ($t = 28.4$, $p < 0.001$, Welch's test). COMA-lite is bimodal: 2/5 seeds match IPPO ($\sim$1.09), 3/5 collapse ($\sim$0.09).

4.2. Scalability ($N \in \{5, 10, 20, 50\}$)

Figure 2 shows per-agent throughput across four densities. At $N = 50$, Greedy collapses (0.183) while Musical Chairs leads (0.360). Value decomposition remains below Random at all densities: at $N = 50$, VDN: 0.157, QMIX: 0.164, QPLEX: 0.168 vs. Random: 0.299. Full results in Appendix B.

4.3. Ablations

Table 2 summarizes ablations (details in Appendix C). No intervention resolves value decomposition failure. Ex-

Table 1. Main comparison ($N = 10$, $K = 8$). Mean ± 95% CI over 5 seeds. Best learned method underlined. Metrics: throughput (TP), total collision rate (Col), secondary-user collision rate (SU Col), primary-user collision rate (PU Col), spectrum utilization (Util), and Jain's fairness index (Jain).

Method	TP ↑	Col ↓	SU Col ↓	PU Col	Util ↑	Jain ↑
Random	1.103 ± 0.006	0.813	0.436	0.377	0.374	0.177
Greedy	$\mathbf{1.938 \pm 0.014}$	**0.671**	**0.294**	0.377	**0.659**	**0.308**
S-ALOHA	1.114 ± 0.014	0.810	0.431	0.379	0.380	0.180
p-CSMA	1.119 ± 0.013	0.810	0.436	0.374	0.380	0.180
MusChairs	1.772 ± 0.301	0.809	0.433	0.376	0.384	0.181
SIC-MMAB	1.216 ± 0.026	0.810	0.434	0.376	0.380	0.180
IDQN	0.604 ± 0.012	0.899	0.521	0.378	0.204	0.098
IDRQN	0.269 ± 0.065	0.947	0.572	0.374	0.107	0.052
IPPO	$\underline{1.101 \pm 0.018}$	<u>0.814</u>	<u>0.439</u>	0.375	<u>0.372</u>	<u>0.177</u>
VDN	0.342 ± 0.035	0.937	0.564	0.373	0.125	0.061
QMIX	0.359 ± 0.058	0.936	0.569	0.367	0.129	0.063
QPLEX	0.434 ± 0.041	0.926	0.553	0.374	0.147	0.072
COMA-lite	0.493 ± 0.540	0.862	0.487	0.375	0.290	0.136
MAPPO	1.104 ± 0.020	0.813	0.436	0.377	0.376	0.178

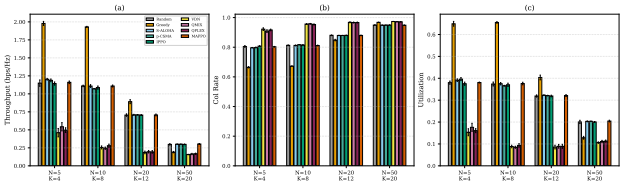

Figure 2. Per-agent throughput across densities ($N \in \{5, 10, 20, 50\}$, $K = 8$). Value decomposition remains below Random at all N. Greedy collapses at high density due to contention.

Table 2. Summary of ablations ($N = 10$, $K = 8$). No intervention resolves value decomposition failure.

Ablation	Key result	Sec.
Agent ID	VDN: 0.332 → 0.317	App. C
Replay ratio	VDN rr4: 0.606, still ≪ IPPO	App. C
Extended (2.2×)	VDN: 0.342 → 0.220 (worse)	App. C
PU sensing	No method improves	App. C
Idle action	VDN+idle: 0.300 ≪ IPPO: 1.109	App. C
Entropy (β)	IPPO: 1.086–1.090 (stable)	App. C
Critic (MAPPO)	1.113 vs. IPPO: 1.116 (no diff)	App. C
Penalty (λ)	TP drops, PU Col unchanged	App. C
Non-stationarity	IPPO adapts; IDRQN: -53.9%	App. C

tended training *worsens* VD (VDN: 0.342 → 0.220 at 2.2× budget). A hyperparameter sweep across optimizers (Adam, RMSProp), GRU sizes (48, 64, 128), and buffer sizes (10K, 20K) for VDN/QMIX finds that the best configuration (VDN-RMSProp: 0.293) still falls 3.8× below Random (Table 3), ruling out tuning artifacts (Hu et al., 2021). PU sensing, idle actions, agent IDs, entropy coefficients, and centralized critics all fail to close the gap.

Table 3. VD hyperparameter sweep ($N = 10$, $K = 8$). No configuration approaches Random (1.103) or IPPO (1.101).

Method	Optimizer	GRU	Buffer	TP
VDN	Adam	48	20K	0.239
VDN	RMSProp	48	20K	**0.293**
VDN	Adam	64	20K	0.259
VDN	Adam	128	20K	0.278
VDN	Adam	48	10K	0.239
QMIX	Adam	48	20K	0.234
QMIX	RMSProp	48	20K	0.237
QMIX	Adam	64	20K	0.243

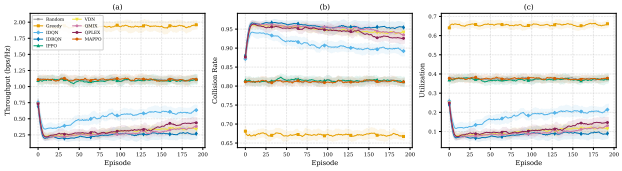

Figure 3. Learning curves ($N = 10$, $K = 8$) over 200 episodes. Value decomposition methods converge well below Random (dashed). Policy gradient (IPPO, MAPPO) converge to the stochastic baseline level. Shaded regions: ±1 std over 5 seeds.

4.4. Boltzmann Execution for Value Decomposition

A natural hypothesis is that ε-greedy execution forces near-deterministic action selection, causing all agents to converge to the same channel. Boltzmann execution replaces argmax with $\pi(a) = \text{softmax}(Q/\tau)$, directly injecting channel diversity.

Table 4 shows this does not resolve the failure. VDN-Boltz (0.317 ± 0.027) is comparable to standard VDN (0.342). QMIX-Boltz (0.484 ± 0.042) shows modest im-

Table 4. Boltzmann vs. ε-greedy execution ($N = 10$, $K = 8$). Boltzmann does not resolve VD failure; the Q-value representation is the bottleneck.

Method	TP (bps/Hz)	SU Col	PU Col
VDN	0.342 ± 0.035	0.564	0.373
VDN-Boltz	0.317 ± 0.027	0.571	0.375
QMIX	0.359 ± 0.058	0.569	0.367
QMIX-Boltz	0.484 ± 0.042	0.541	0.376
QPLEX	0.434 ± 0.041	0.553	0.374
QPLEX-Boltz	0.288 ± 0.091	0.573	0.378
Random	1.103 ± 0.006	0.436	0.377
IPPO	1.101 ± 0.018	0.439	0.375

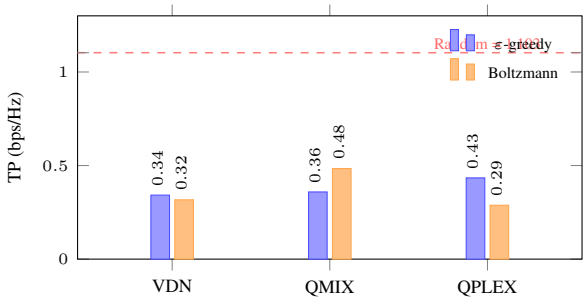

Figure 4. ε-greedy vs. Boltzmann execution for VD methods. Boltzmann (softmax) sampling does not close the gap to Random (dashed), confirming the learned Q-values are degenerate under additive/monotonic constraints.

provement over QMIX (0.359) but remains far below Random. QPLEX-Boltz (0.288 ± 0.091) is actually *worse* than standard QPLEX (0.434). SU collision rates are nearly identical across execution strategies (> 0.54), confirming the Q-value landscape itself is pathological: even stochastic sampling from incorrect Q-values does not produce effective anti-coordination.

4.5. Full COMA with Exact Counterfactual Marginalization

COMA-lite's bimodality raised whether single-sample approximation was the instability source. COMA-FULL computes the exact baseline by enumerating all $K = 8$ actions per agent. We test two critic designs: (1) the acting agent's one-hot action as input (COMA-FULL), and (2) the full $N \times K$ joint action as input (COMA-Full-Joint), directly addressing whether critic expressivity limits performance.

Table 5 shows both variants are *worse* than COMA-LITE. COMA-FULL (0.151 ± 0.176) partially converges on 3/5 seeds. COMA-Full-Joint (0.066 ± 0.057) is even worse: no seed exceeds 0.122, and the richer critic input does not help. This rules out critic under-parameterization as the explanation. The seeds succeeding for COMA-LITE

Table 5. COMA variants per-seed TP (bps/Hz). Exact marginalization and richer critics both degrade performance relative to single-sample COMA-lite.

Seed	COMA-Full	Full-Joint	COMA-lite
42	0.304	0.097	1.081
123	0.001	0.073	1.106
456	0.205	0.122	0.092
789	0.244	0.038	0.101
1024	0.001	0.000	0.083
Mean	0.151 ± 0.176	0.066 ± 0.057	0.493 ± 0.540

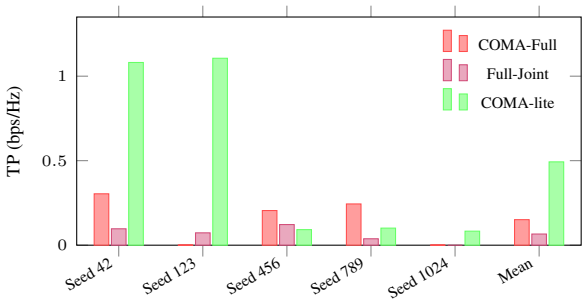

Figure 5. Per-seed comparison of COMA variants. COMA-lite (green) succeeds on 2/5 seeds (42, 123) where both exact-marginalization variants fail, suggesting single-sample gradient noise helps escape degenerate equilibria in anti-coordination.

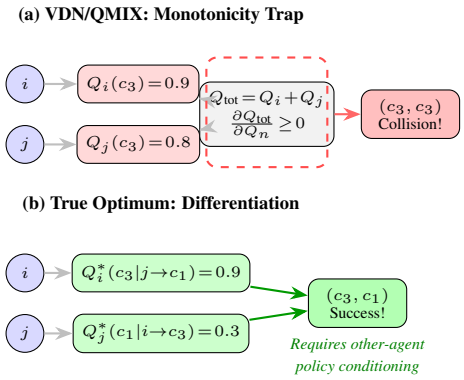

Figure 6. Anti-coordination failure mechanism. **(a)** VDN/QMIX monotonicity ($\partial Q_{\text{tot}}/\partial Q_n \geq 0$) forces agents to independently prefer the same channel, causing collisions. **(b)** The optimum requires Q^* conditioned on others' policies—inexpressible under additive/monotonic constraints. This generalizes to $N = 10$ agents (SU collision > 0.55; Table 1).

(42, 123) do not succeed for either Full variant, suggesting single-sample noise can *help* escape degenerate equilibria.

5. Discussion

Anti-coordination explains VD failure. Figure 6 illustrates the core structural problem. VDN's additive and QMIX's monotonic decompositions assume individually

optimal actions compose into a jointly optimal action. In anti-coordination this is violated: individual Q-functions drive all agents toward the same channel (SU collision $>$ 0.55). QPLEX removes monotonicity but still fails (0.434 vs. Random 1.103); see the toy analysis in Appendix D.

Value representation is the primary bottleneck. The Boltzmann experiment (Fig. 4) is a crisp diagnostic: if deterministic execution caused the failure, softmax sampling should help. Instead, all Boltzmann variants remain far below Random, because the Q-values themselves are degenerate under VDN/QMIX's structural constraints. The hyperparameter sweep (Table 3) confirms this is not a tuning artifact: the best VDN configuration (RMSProp, 0.293) is still $3.8\times$ below Random. We note that coordination-graph methods (DCG/DDFG (Böhmer et al., 2020)) and IGM-fixing approaches (QFIX (Liu et al., 2025)) represent promising alternative decompositions that could potentially overcome the structural limitations we identify; evaluating these is a priority for future work.

Exact counterfactual marginalization does not help. COMA-FULL's failure is surprising (Fig. 5). The COMA-Full-Joint variant (0.066), which conditions the critic on the full $N \times K$ joint action, performs *worse* than the agent-only variant (0.151), ruling out critic under-parameterization. We hypothesize that exact marginalization reduces gradient signal magnitude (advantage $\to 0$ under near-uniform policies), while single-sample noise provides stronger signals that occasionally escape degenerate basins, analogous to how SGD noise helps escape sharp minima.

Policy gradient methods succeed through stochasticity. IPPO/MAPPO match Random ($\sim$1.10 bps/Hz) by naturally maintaining stochastic policies. The entropy ablation confirms this is inherent to policy gradients, not entropy regularization (1.086–1.090 across $\beta_{\text{ent}} \in \{0.0, 0.01, 0.03\}$). The centralized critic provides no benefit (MAPPO: 1.113 vs. IPPO: 1.116), and Musical Chairs' success at $N = 50$ (0.360 bps/Hz, highest) confirms explicit anti-coordination is key.

Broader implications for MARL algorithm selection. Our results suggest that the dominant performance of value decomposition on cooperative benchmarks (e.g., StarCraft (Samvelyan et al., 2019)) does not transfer to anti-coordination domains. This has implications beyond DSA: any multi-agent problem where agents must differentiate—including distributed task allocation, multi-robot coverage, and load balancing—may exhibit similar VD failure. The success of stochastic policy gradient methods is notable because they achieve anti-coordination *without explicit coordination mechanisms*, purely through maintaining action entropy.

Limitations. Network sizes ($N \leq 50$) are small relative to 6G; mean-field MARL (Yang et al., 2018) should be tested. We omit coordination-graph methods (DCG (Böhmer et al., 2020), DDFG), IGM-fixing layers (QFIX (Liu et al., 2025)), auto-regressive PG baselines, communication-based MARL (CommNet, TarMAC), and QTRAN (Son et al., 2019)—these are important directions that could test whether higher-order factorization or sequential action selection resolves the anti-coordination failure. PU collisions are treated as a soft cost (λ); constrained-RL formulations with hard PU protection guarantees would broaden applicability. All evaluation is simulation-based; channel models use Rayleigh fading without shadowing or spatial correlation. Our negative results are specific to the strict anti-coordination regime of our DSA environments; value decomposition succeeds in related tasks with richer observations (Naderializadeh et al., 2021).

6. Conclusion

We evaluated sixteen MARL methods for DSA across fourteen experiments. Table 6 summarizes the key findings for each paradigm. Under the additive/monotonic decomposition constraints of VDN and QMIX, value decomposition dramatically underperforms: 2.5–$3\times$ below Random at $N = 10$, and below Random at $N = 50$. This failure persists across Boltzmann execution, hyperparameter sweeps (optimizer, GRU size, buffer), extended training, agent ID symmetry breaking, PU sensing, and idle actions—a total of twelve ablations, none of which resolves the gap. Full COMA with exact counterfactual marginalization—including a joint-action critic variant (0.066)—is worse than single-sample COMA-lite (0.493), providing evidence that gradient noise can help escape degenerate equilibria in anti-coordination domains. Policy gradient methods (IPPO, MAPPO) are the only learned methods to consistently match stochastic baselines across all network densities tested.

Practical implications. Our results have direct implications for MARL algorithm selection in wireless networks. For DSA and similar anti-coordination problems, practitioners should default to policy gradient methods (IPPO/MAPPO) rather than value decomposition, despite the latter's success in cooperative benchmarks like StarCraft (Samvelyan et al., 2019). The failure is not a tuning issue—it is structural, arising from the monotonicity constraints inherent to VDN and QMIX. When explicit anti-coordination protocols are feasible, bandit methods like Musical Chairs (Rosenski et al., 2016) provide the strongest performance at high density.

Future directions. Several promising directions could address the limitations identified in this study.

Coordination-graph decompositions (DCG (Böhmer et al., 2020), DDFG) and IGM-fixing layers (QFIX (Liu et al., 2025)) represent alternative factorization approaches that may overcome the structural constraints of additive and monotonic decomposition. Auto-regressive policy gradient methods, where agents select actions sequentially conditioned on predecessors, could provide the inter-agent conditioning that value decomposition lacks. Communication-augmented MARL (CommNet, TarMAC) may enable learned coordination without centralized critics. QTRAN (Son et al., 2019), which relaxes the IGM constraint entirely, is another important baseline we did not evaluate; its unrestricted joint Q-function may handle anti-coordination better than the constrained decompositions studied here, though at the cost of training stability. Attention-based credit assignment methods could also provide more flexible inter-agent value factorization. For scaling to realistic 6G densities ($N > 100$), mean-field MARL (Yang et al., 2018) offers a tractable approximation by modeling the aggregate effect of the population rather than individual agent interactions. On the methodological side, our COMA results raise the broader question of when variance reduction in policy gradient estimators is beneficial versus harmful—a question with implications for multi-agent anti-coordination settings more generally. The observation that single-sample estimation outperforms exact marginalization in our setting deserves further theoretical investigation. Finally, hardware testbed validation, integration with 3GPP-compliant protocol stacks, and evaluation under realistic propagation models with shadowing and multi-path fading are essential steps toward real-world deployment of learned DSA policies. We hope this systematic negative result for value decomposition in anti-coordination settings will guide future MARL research toward structure-aware algorithm selection and away from the assumption that methods dominant on cooperative benchmarks transfer universally. More broadly, our study demonstrates the value of comprehensive negative results: by exhaustively testing interventions (twelve ablations, two execution strategies, three decomposition architectures), we provide strong evidence that the failure is structural rather than incidental, saving the community from pursuing unproductive optimization directions for VDN/QMIX in anti-coordination domains.

Table 6. Summary of key findings across paradigms. TP = throughput (bps/Hz) at $N = 10$, $K = 8$. ✓ = matches or exceeds Random; × = fails.

Paradigm	TP	Anti-c.	Scales?	Finding
Non-learn. (Greedy)	1.94	✓	×	Collapses $N \geq 50$
Stoch. baselines	1.10	✓	✓	Diversity suffices
Bandits (MusCh.)	1.77	✓	✓	Best at $N = 50$
Indep. VB (IDQN)	0.60	×	×	Greedy convergence
VD ε-greedy	0.34–0.43	×	×	Monotonicity trap
VD Boltzmann	0.29–0.48	×	×	Q-values degenerate
COMA-lite	0.49	Bimod.	—	2/5 seeds succeed
COMA-Full	0.07–0.15	×	—	Noise helps
PG (IPPO)	1.10	✓	✓	Stochastic policy
Cent.-crit. (MAPPO)	1.10	✓	✓	No critic benefit

References

Böhmer, W., Kurin, V., and Whiteson, S. Deep coordination graphs. In *ICML*, 2020.

Boursier, E. and Perchet, V. SIC-MMAB: Synchronisation involves communication in multiplayer multi-armed bandits. In *NeurIPS*, 2019.

Claus, C. and Boutilier, C. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI*, pp. 746–752, 1998.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.

Hausknecht, M. and Stone, P. Deep recurrent Q-learning for partially observable MDPs. *AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*, 2015.

Hu, J., Jiang, S., Harding, S. A., Wu, H., and Liao, S.-w. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. In *arXiv preprint arXiv:2102.03479*, 2021.

Jain, R., Chiu, D.-M., and Hawe, W. R. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. *DEC Technical Report TR-301*, 1984.

Letaief, K. B., Chen, W., Shi, Y., Zhang, J., and Zhang, Y.-J. A. The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine*, 57(8):84–90, 2019.

Liu, S. et al. QFIX: Provably optimal Q-value decomposition with fixing layers for cooperative MARL. *arXiv preprint arXiv:2505.10484*, 2025.

Luong, N. C., Hoang, D. T., Gong, S., Niyato, D., Wang, P., Liang, Y.-C., and Kim, D. I. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials*, 21(4):3133–3174, 2019.

Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Naderializadeh, N., Sydir, J., Simsek, M., and Nikopour, H. Resource management in wireless networks via multi-agent deep reinforcement learning. *IEEE Transactions on Wireless Communications*, 20(6):3507–3523, 2021.

Naparstek, O. and Cohen, K. Deep multi-user reinforcement learning for distributed dynamic spectrum access. *IEEE Transactions on Wireless Communications*, 18(1): 310–323, 2019.

Papoudakis, G., Christianos, F., Schäfer, L., and Albrecht, S. V. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Rashid, T., Samvelyan, M., Schroeder de Witt, C., Farquhar, G., Foerster, J., and Whiteson, S. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*, 2018.

Rosenski, J., Shamir, O., and Szlak, L. Multi-player bandits – a musical chairs approach. In *ICML*, 2016.

Samsung Research. 6G: The next hyper-connected experience for all. White paper, Samsung Electronics, 2020.

Samvelyan, M., Rashid, T., Schroeder de Witt, C., Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C.-M., Torr, P. H. S., Foerster, J., and Whiteson, S. The StarCraft multi-agent challenge. In *AAMAS*, 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., and Yi, Y. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *ICML*, 2019.

Sunehag, P., Lever, G., Gruslys, A., et al. Value-decomposition networks for cooperative multi-agent learning. In *AAMAS*, 2018.

Tan, M. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *ICML*, pp. 330–337, 1993.

Tataria, H., Shafi, M., Molisch, A. F., Dohler, M., Sjöland, H., and Tufvesson, F. 6G wireless systems: Vision, requirements, challenges, insights, and opportunities. *Proceedings of the IEEE*, 109(7):1166–1199, 2021.

Wang, J., Ren, Z., Liu, T., Yu, Y., and Zhang, C. QPLEX: Duplex dueling multi-agent Q-learning. In *ICLR*, 2021.

Wang, S., Liu, H., Gomes, P. H., and Krishnamachari, B. Deep reinforcement learning for dynamic multichannel access in wireless networks. *IEEE Transactions on Cognitive Communications and Networking*, 4(2):352–365, 2018.

Wei, E. and Luke, S. Lenient learning in independent-learner stochastic cooperative games. In *Journal of Machine Learning Research*, volume 17, pp. 1–42, 2016.

Xu, Y., Yu, J., Headley, W. C., and Buehrer, R. M. Deep reinforcement learning for dynamic spectrum access in wireless networks. In *IEEE MILCOM*, pp. 207–212, 2018.

Xu, Y., Yu, J., and Buehrer, R. M. The application of deep reinforcement learning to distributed spectrum access in dynamic heterogeneous environments with partial observations. *IEEE Transactions on Cognitive Communications and Networking*, 6(3):1206–1218, 2020.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. Mean field multi-agent reinforcement learning. In *ICML*, 2018.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of PPO in cooperative, multi-agent games. In *NeurIPS*, 2022.

A. Hyperparameter Settings

Table 7. Environment and shared hyperparameters.

Parameter	Value
Episodes / Timesteps	200 / 120 (24K steps)
Observation window W	5
Mean SNR $\bar{\gamma}$ / Noise σ^2	10 / 0.1
PU p_{on} / p_{off}	0.3 / 0.5
Discount γ / GRU hidden	0.99 / 48
Seeds	42, 123, 456, 789, 1024

Table 8. Value-based method hyperparameters.

Parameter	Value
Learning rate	1×10^{-3}
Buffer / Batch (recurrent)	20K / 32
Target update	80 steps
ε: start / end / decay	1.0 / 0.05 / 0.997
Boltzmann τ: start / min / decay	1.0 / 0.1 / 0.995

Table 9. Policy gradient hyperparameters (IPPO, MAPPO, COMA).

Parameter	Value
LR (actor / critic)	3×10^{-4} / 1×10^{-3}
Clip ϵ / GAE λ	0.2 / 0.95
Entropy (IPPO / MAPPO)	0.02 / 0.03
PPO epochs / COMA-Full CF	3 / exact ($K=8$)

B. Full Metric Tables

C. Ablation Details

D. Toy Game Analysis

In a 2-agent, 2-channel game: $Q^*([1,2]) = Q^*([2,1]) = 1$, $Q^*([1,1]) = Q^*([2,2]) = 0$. Under VDN, if $Q_1(1) > Q_1(2)$, greedy action is $[1,1]$ — the worst joint action. QMIX's monotonicity prevents penalizing $[1,1]$ while rewarding $[1,2]$. QPLEX can represent this in principle but faces a chicken-and-egg problem: accurate Q-values need knowledge of the other agent's policy. This generalizes to $N=10$, $K=8$, explaining SU collision rates > 0.55.

E. Boltzmann Per-Seed Results

F. Full COMA Details

COMA-FULL uses: GRU actor (48 hidden), centralized critic, exact counterfactual enumeration ($K=8$ actions), 3 epochs per rollout, LR 3×10^{-4}, gradient clip 0.5. Two critic variants are evaluated:

- **COMA-Full**: critic input = global state + acting agent's one-hot action ($|s| + K$ dims). Output: per-agent Q-values (N dims).
- **COMA-Full-Joint**: critic input = global state + full joint action one-hot ($|s| + NK$ dims $= 2K + NK + NK$). Hidden layers: $128 \to 64$. This directly addresses the concern that critic under-parameterization limits performance.

COMA-Full-Joint's worse performance (0.066 vs. 0.151) suggests the richer critic is *harder* to train due to the high-dimensional input ($|s| + NK = 96$ dims), without providing better credit assignment.

G. Compute Budget

IPPO is $\sim$3$\times$ faster than VD: $\sim$360s vs. $\sim$1200s per seed (300 episodes), while achieving $\sim$3$\times$ higher throughput. COMA-Full adds K-fold critic overhead per agent per training step.

H. Code and Reproducibility

Code will be released upon acceptance.

Table 10. Scalability: per-agent throughput (bps/Hz) across density configurations.

Method	$N=5$	$N=10$	$N=20$	$N=50$
Random / Greedy	1.149 / 1.978	1.103 / 1.938	0.706 / 0.892	0.299 / 0.183
S-ALOHA / p-CSMA	1.203 / 1.188	1.114 / 1.119	0.709 / 0.709	0.301 / 0.301
IDQN / IDRQN / IPPO	0.777 / 0.336 / 1.143	0.604 / 0.269 / 1.101	0.268 / 0.165 / 0.706	0.167 / — / 0.301
VDN / QMIX / QPLEX	0.461 / 0.543 / 0.496	0.342 / 0.359 / 0.434	0.186 / 0.196 / 0.194	0.157 / 0.164 / 0.168
MAPPO / MusChairs	1.160 / —	1.104 / 1.772	0.707 / —	0.301 / 0.360

Table 11. Non-stationarity (PU shift at ep. 100).

Method	**Phase 1**	**Phase 2**	**Drop**
IDRQN	0.243	0.112	53.9%
IPPO / MAPPO	0.587 / 0.591	0.596 / 0.590	−1.5% / 0.3%
VDN / QPLEX	0.148 / 0.148	0.168 / 0.159	−13.4% / −7.5%

Table 12. Extended training. VD gets *worse* with more training.

Method	**200 ep**	**300 ep**	**350 ep**
VDN / QPLEX	0.342 / 0.434	0.238 / 0.261	0.220 / —
IPPO / MAPPO	1.101 / 1.104	1.105 / —	1.130 / 1.120

Table 13. Replay ratio (VDN / QMIX, TP bps/Hz).

	rr=1	**rr=2**	**rr=4**
VDN / QMIX	0.222 / 0.220	0.299 / 0.288	0.606 / 0.328

Table 14. Agent ID ablation (TP bps/Hz).

	No ID	**With ID**
VDN / QMIX / QPLEX	0.332 / 0.277 / 0.336	0.317 / 0.296 / 0.340

Table 15. PU sensing variant (TP bps/Hz).

	Base	**+PU Sense**
IPPO / MAPPO	1.101 / 1.104	1.102 / 1.105
VDN / QPLEX	0.342 / 0.434	0.292 / 0.405

Table 16. Idle action variant and penalty ablation.

	Idle Action		**Penalty λ (MAPPO)**			
	TP	**PU Col**	0.0	0.5	1.0	2.0
IPPO+idle	1.109	0.369	1.103	0.920	0.732	0.352
VDN+idle	0.300	0.328				

Table 17. Boltzmann VD per-seed TP (bps/Hz).

Seed	**VDN-B**	**QMIX-B**	**QPLEX-B**
42 / 123	0.305 / 0.325	0.532 / 0.476	0.344 / 0.294
456 / 789	0.349 / 0.290	0.443 / 0.468	0.162 / 0.320
1024	0.315	0.502	0.322
Mean	0.317 ± 0.027	0.484 ± 0.042	0.288 ± 0.091