

Multi-Agent Warehouse Logistics over WiFi Direct Ad Hoc Networks

Chien-Cheng Wu

Department of Computer Science and Engineering

National Sun Yat-sen University

Kaohsiung, Taiwan

ccwu@cse.nsysu.edu.tw

Abstract—Warehouse Automated Guided Vehicle (AGV) fleets typically rely on infrastructure-mode WiFi, which routes every inter-agent message through a central access point and thereby incurs double-hop latency, redundant energy consumption, and a single point of failure. We replace this backbone with a WiFi Direct peer-to-peer ad hoc overlay: clients uplink compact state beacons to a nearby group owner, which disseminates one aggregated broadcast per interval. This halves network-wide per-message energy, cuts device-side radio energy by 33–46% under transmit power control, reduces occupied beacon airtime by ≈ 57 –74%, and provides an execution substrate for a Centralised Training with Decentralised Execution (CTDE) Multi-Agent Reinforcement Learning (MARL) policy. Empirically, the simplest heuristics trail learned policies by over ten percentage points (pp), and the strongest heuristic matches them in mean but not in worst case: our CTDE-QMIX policy sustains mean pick rates of 77.0–80.0% (1.3–1.7 pp above the strongest heuristic, 1.3–3.2 pp above the co-trained VDN/MAPPO baselines), raises the worst-case floor under the most uncertain demand, and stays within 1 pp of its clean-channel mean under i.i.d. beacon loss up to 30% and handoff blackouts. Co-designing the wireless and learning layers yields AGV fleets that are energy-efficient, robust to infrastructure failure, and reliable under demand uncertainty.

Index Terms—WiFi Direct, Ad Hoc Network, Multi-Agent Reinforcement Learning, Energy Efficiency, Decentralised Execution

I. INTRODUCTION

Modern warehouses pair human pickers with autonomous mobile robots, placing order picking at the core of collaboration. Operational efficiency is measured by the *pick rate*, defined as items retrieved per time window, directly governing throughput and cost. Classical schedulers utilize hand-engineered heuristics [1] that perform well under stable demand but degrade as arrival patterns drift or fleets change, motivating Multi-Agent Reinforcement Learning (MARL) under the Centralised Training with Decentralised Execution (CTDE) paradigm. Decentralised execution, however, presupposes a low-latency, energy-efficient communication fabric that infrastructure-mode WiFi struggles to provide, since every peer-to-peer exchange must transit the shared access point (AP), doubling its wireless hop count.

The optimal substrate for decentralised warehouse control is WiFi Direct (WiFi P2P), providing peer-to-peer connections without access points: within a P2P group, links to the Group Owner (GO) are single-hop, and the GO's aggregated broad-

cast disseminates group state without AP, while spatial reuse holds across non-overlapping groups. The resulting dynamic overlay among AGVs and pickers simultaneously serves as the real-time command-and-telemetry channel and message-passing substrate for CTDE-MARL execution, each agent sharing partial observations only with k -nearest neighbours in a co-design where topology and learning reinforce one another.

This paper makes three contributions: (1) a WiFi Direct communication architecture with a corrected, relay-honest analysis predicting $\sim 50\%$ network-wide energy reduction, 33–46% device-side radio-energy reduction under transmit power control, and ≈ 57 –74% lower beacon airtime (Section IV); (2) a CTDE-MARL framework that exploits proximity-based WiFi Direct message passing at run-time, eliminating the centralised coordinator, with staleness-aware beacon encoding and beacon-dropout training (Section VI); and (3) an empirical evaluation of three heuristic baselines (FM, PDM, CTA), two MARL baselines (VDN, MAPPO), and the trained CTDE-QMIX policy across three order-arrival distributions, with a robustness study under beacon loss and GO-handoff blackouts (Section VII).

II. RELATED WORK

Heuristic AGV routing is extensively documented: Person-to-Goods (PTG) paradigms guide AGVs behind human pickers along Travelling-Salesman (TSP) trajectories [1]; zoning strategies such as Pick-Don't-Move restrict pickers to aisles while AGVs roam, inducing under-utilisation [2]; and Closest-Task Assignment (CTA) supports Goods-to-Person (GTP) operations by allocating storage racks to the closest AGV via A^* [3], [4]. From a wireless perspective, industrial AGV fleets commonly utilize infrastructure-mode WiFi, funneling all traffic through an access point, introducing a bottleneck and single point of failure, with AP-induced latency of 10–30 ms under load [5]. To address these limitations, device-to-device (D2D) communication via WiFi Direct has emerged as a promising alternative. Establishing connections via Group Owner (GO) negotiation achieves both lower round-trip latency and lower energy consumption at short range [7], [8]. To our knowledge, the present work is the first deployment of WiFi Direct as the substrate for MARL-controlled AGV fleets. On the learning side, value-decomposition methods factor a joint action-value into per-agent utilities: VDN [12] additively, and

QMIX [9] through a monotonic mixing function that keeps the greedy joint action recoverable by local argmax. MAPPO [10] showed PPO with a shared centralised critic to be competitive with lower hyperparameter sensitivity. Communication-aware MARL [11] models the inter-agent channel explicitly, finding that local-neighbourhood messaging often matches global communication at far lower bandwidth. MARL has also been applied to warehouse pick-up-and-delivery directly: shared-experience actor-critic on the multi-robot warehouse (RWARE) benchmark [13], and hierarchical MARL scaled to fleets with robotic and human co-workers [14]. These works assume an idealised or unmodelled channel; our contribution is orthogonal: we ground the neighbourhood channel in a concrete WiFi Direct design, quantify its energy and airtime costs, and train against its failure modes (beacon loss, GO handoff).

III. SYSTEM MODEL

A. Warehouse Environment and Agents

The environment is a grid warehouse with shelves (non-traversable cells holding items), delivery stations, charging stations, and pathways; every cell lies within 2.4/5 GHz WiFi Direct range of at least one peer. Battery-powered AGVs transport items and empty storage media, each carrying an IEEE 802.11 WiFi Direct radio and an edge-compute unit running the decentralised MARL actor; human pickers retrieve items and carry WiFi Direct handheld terminals for task assignment and acknowledgement. High-level actions are translated to atomic grid moves (N, S, E, W, Wait) at each time-step; AGVs approaching the same cell resolve the conflict by mission-criticality priority via distributed token passing over WiFi Direct (Section IV-E).

B. Order Arrival Distributions

Robustness is evaluated under three arrival models: Normal (μ, σ) for stable, forecast-driven demand; Poisson (λ) for bursty e-commerce arrivals; and Uniform over $[a, b]$ for maximally uncertain demand.

C. Wireless Channel Model

We model the warehouse channel as a path-loss environment with log-normal shadowing, $PL(d) = PL(d_0) + 10n \log_{10}(d/d_0) + X_\sigma$, where $d_0 = 1$ m, $n \approx 2.2\text{--}3.0$ is the path-loss exponent (calibrated for cluttered warehouse interiors), and $X_\sigma \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 6$ dB captures shelving-induced shadowing. A link is active when received signal strength exceeds -70 dBm (reliable WiFi Direct throughput ≥ 20 Mbps); at a typical 15 dBm transmit power the effective radius is $r_c = 15\text{--}20$ m (two to three aisles). This model captures only large-scale effects; links see small-scale fading (Rician in line-of-sight aisles, Rayleigh otherwise) and transient blockage from passing AGVs and bodies. Rather than simulating the fast channel, we abstract these dynamics into a per-link, per-beacon outage probability p (i.i.d.; a two-state Markov variant models correlated fades) and GO-handoff blackouts suppressing an agent's beacons for $\lceil 250 \text{ ms}/T_b \rceil = 3$

intervals (Section IV-A); Section VII-E sweeps both knobs, validating the abstraction empirically.

IV. WiFi DIRECT AD HOC PROTOCOL DESIGN

A. Group Owner Negotiation and Dynamic Topology

WiFi Direct organises devices into P2P groups under the management of a Group Owner (GO). Specifically, following the specification's Intent Value negotiation [6], an AGV with battery $> \tau_{\text{bat}}$ and CPU load $< \tau_{\text{cpu}}$ advertises high intent to become GO; peers within range r_c associate as clients (up to $G_{\text{max}} = 8$ per GO), and the GO broadcasts aggregate state every $T_b = 100$ ms. This dynamic topology naturally implements the k -nearest-neighbour graph required for MARL observation sharing (Section VI-B). Since standard WiFi Direct group formation takes several seconds [6], fleet devices are pre-provisioned as persistent-group peers (cached credentials, fixed channel plan, static addressing); an AGV moving out of range rejoins a neighbouring group via the two-frame Invitation exchange and reassociation, within a provisioned 250 ms handoff budget. Section VI-B encodes such gaps in the policy input and Section VII-E quantifies their impact.

B. Message Format and Proximity-Based Observation Sharing

Each agent i periodically broadcasts a structured state beacon: spatial coordinates and heading (4 B), task type and destination (4 B), battery and load status (2 B), queue length at the nearest station (2 B), and routing intent (2 B), constituting a $B_{\text{obs}} = 14$ B payload plus a 28 B P2P header H , i.e. an on-air packet of $B_{\text{msg}} = B_{\text{obs}} + H = 42$ bytes, or 1.12 kb/s payload (3.36 kb/s on-air) per AGV at $T_b = 100$ ms, well within WiFi Direct capacity.

C. Theoretical Analysis: Energy

We adopt representative per-byte costs $E_{\text{tx}} = 0.46 \mu\text{J/B}$ and $E_{\text{rx}} = 0.18 \mu\text{J/B}$ for an IEEE 802.11n transceiver at access-point range, consistent in magnitude with measured 802.11 per-frame energy anatomies [15]. In infrastructure mode, a message of B bytes traverses two hops ($i \rightarrow \text{AP} \rightarrow j$), meaning the *network-wide* consumption, combining device and AP, equals $2B(E_{\text{tx}} + E_{\text{rx}})$, versus $B(E_{\text{tx}} + E_{\text{rx}})$ for a single P2P hop: $\eta_{\text{net}} = 50\%$. This represents an *infrastructure-energy* perspective: since the eliminated hop's energy is dissipated at the mains-powered AP, at fixed transmit power the endpoints' *battery-side* energy is identical across both modes. The physical device-side efficiency stems from transmit power control (TPC) over the shorter P2P link: using a first-order radio model $E_{\text{tx}}(B, d) = B(E_{\text{elec}} + \varepsilon_{\text{amp}} d^n)$ calibrated so that $E_{\text{elec}} = \varepsilon_{\text{amp}} d_{\text{AP}}^n = 0.23 \mu\text{J/B}$ at $d_{\text{AP}} = 40$ m (splitting the transceiver figure evenly between electronics and amplifier), a $d_{\text{P2P}} = 15$ m link at $n = 2.6$ requires only $\varepsilon_{\text{amp}} d_{\text{P2P}}^n \approx 0.018 \mu\text{J/B}$, yielding an 11 dB power drop. This results in $E_{\text{tx}} \approx 0.248 \mu\text{J/B}$, giving 46% lower transmit energy and 33% lower total device energy per delivered message. With $N = 20$ AGVs at $f = 10$ beacons/s over an 8 h shift, this equals ≈ 2.6 J per vehicle (≈ 51 J fleet-wide). Beacon energy is negligible against a multi-MJ traction battery; we make no battery-autonomy claim from beacons

alone. The overlay's primary significance lies in the airtime reduction of Section IV-D, removing the central AP as a single point of failure, and applying the 33–46% TPC saving to bulk payload traffic, alongside halved network-wide per-message energy, reaching $\approx 67\%$ once AP electronics and P2P TPC are counted (a 50–67% network-wide range).

D. Theoretical Analysis: Bandwidth

Within a P2P group, client $\leftrightarrow$ client unicast is relayed through the GO (which acts as the group's access point), so single-hop delivery holds only for GO-involved links. Beacon dissemination therefore uses the aggregation of Section IV-A: each of the $G - 1$ clients in a group of size G uplinks its B_{msg} -byte beacon and the GO broadcasts one aggregated $H + G B_{\text{obs}}$ -byte frame per interval, reaching all members within T_b ; the occupied airtime per agent per interval is $A_{\text{P2P}} = [(G-1) B_{\text{msg}} + H]/G + B_{\text{obs}}$, versus $A_{\text{infra}} = (1+k) B_{\text{msg}}$ for infrastructure's one uplink plus k unicast downlinks. With $G = G_{\text{max}} = 8$ this gives $A_{\text{P2P}} = 54.25 \text{ B}$ per agent per interval (10,850 B/s at $N = 20$, $f = 10$), against 25,200 ($k=2$) and 42,000 B/s ($k=4$) in infrastructure mode: an airtime reduction of **56.9%** ($k=2$, **conservative**) and **74.2%** ($k=4$). At the simulated scale ($N = 5$, one group of $G = 5$, mean $k \approx 2$) the same model predicts 57.8%. We accordingly claim $\approx 57\text{--}74\%$; the infrastructure baseline is unicast fan-out, and Section VII-D compares these predictions against simulation, where intermittent isolation ($k = 0$) lowers the realised average.

E. Distributed Conflict Resolution via Token Passing

To resolve cell conflicts among $m \geq 2$ AGVs without a coordinator, a lightweight token-passing protocol is employed: each claimant broadcasts CLAIM(cell_id, src_cell, priority, wait_age, id), where priority denotes mission criticality (delivery > picking > returning) and wait_age tracks delay steps. A strict total order (priority + wait_age, then id) selects a unique winner; others WAIT and increment wait_age to prevent starvation. The overhead is m messages. src_cell detects swap conflicts (physical exchange), whereupon the lower-ordered agent retreats to a side cell (or both hold if boxed in) so the winner passes. A blocked agent can issue a *push request* compelling a stationary lower-priority AGV to yield (two extra messages); pickers yield after three requests. Because the claim order is total and obstructions are mobile or displaceable, conflicts cannot livelock; empirical results confirm zero deadlocks over 1,440 evaluation episodes.

V. HEURISTIC ALGORITHM DESIGN

We evaluate three heuristics spanning both operating paradigms. **Follow Me (FM)**, under the PTG paradigm, assigns multiple AGVs to follow each human picker, with a TSP over assigned items determining the retrieval sequence; it minimises picker idle time but can inflate AGV travel. **Pick, Don't Move (PDM)**, inspired by [2], partitions the warehouse into fixed picker zones where AGVs navigate along TSP paths,

each picker serving arriving AGVs by proximity. **Closest Task Assignment (CTA)**, adapted from [3] for the GTP paradigm, assigns each storage medium to the nearest available AGV using A* [4], with pickers processing AGVs in FIFO order at their stations. All three cycle each AGV through Picking, Delivering, and Returning phases, with A* paths respecting shelf obstacles and re-routing around peers. All controllers, whether heuristic or learned, execute through the identical conflict-resolution layer of Section IV-E, so throughput differences isolate task allocation and routing policy.

VI. MULTI-AGENT REINFORCEMENT LEARNING ALGORITHM

A. Problem Formulation as a Cooperative Markov Game

The warehouse is modelled as a cooperative multi-agent Markov game $\langle \mathcal{A}, S, \{O_i\}, \{A_i\}, P, R \rangle$ over N_{AGV} AGVs and N_{Picker} pickers. The global state S holds agent positions and statuses, shelf occupancy, station capacity, the request queue, and the arrival-distribution context; each partial observation O_i contains the agent's own position, mission, and battery plus the beacons from its current P2P group (Section VI-B). The primitive action sets A_i are {N, S, E, W, Wait, Pick, Drop, GoCharge} for AGVs and {Move, Pick, Wait} for pickers, with joint action $A_t = \{a_{i,t}\}_{i \in \mathcal{A}}$. High-level control is cast as temporally extended task-assignment options: an idle AGV commits to one of its five nearest unassigned orders or defers, busy agents are masked to continue their current option, and option execution (routing, pick, drop, charge) runs on the identical A*/conflict-resolution stack as the heuristics, so the layer below task assignment is shared by all controllers. The shared global reward incentivises high pick rate:

$$R_t = w_1 \Delta \text{Picks}_t - w_2 Q_t - w_3 C_t - w_4 \Delta E_t, \quad (1)$$

where ΔPicks_t counts items picked, Q_t is the queue length, C_t the conflict events, and ΔE_t the beacon energy in step t ; the last term encourages energy-frugal communication.

B. WiFi Direct-Aware Observation Sharing

Each agent's observation is augmented with beacons from its current P2P group, $O_i^+ = [o_i^{\text{local}}, \{b_j : j \in \mathcal{N}_i(t)\}]$, where o_i^{local} denotes local sensing and $\mathcal{N}_i(t)$ its group at time t . This enforces partial observability under proximity-constrained communication, ensuring no agent ever requires information beyond its immediate radio range; since $|\mathcal{N}_i(t)| \leq G_{\text{max}}$, O_i^+ is fixed-dimensional. Because beacons are frequently lost to fading or suppressed during handoffs (Section III-C), each neighbour slot also carries a staleness feature Δt_j (intervals since j 's last beacon) and a validity mask; a missing beacon retains its prior value for up to $\Delta t_{\text{max}} = 5$ intervals, then is zero-masked. During training, we randomise the per-link packet loss probability $p \sim \mathcal{U}[0, p_{\text{max}}]$ (*communication domain randomisation*, mirroring Section VI-E's demand-side randomisation), thereby enabling the policy to operate under degraded connectivity.

C. CTDE with QMIX

During centralised training [9], a mixing network with access to the global state S_t learns a joint value $Q_{\text{tot}}(S_t, A_t) = f_{\text{mix}}(\{Q_i(O_i^+, a_i)\})$ under a monotonicity constraint that keeps the decentralised $\arg \max$ consistent; during execution, each agent selects $\hat{a}_i = \arg \max_a Q_i(O_i^+, a)$ using local observations and beacons, without querying a central server. QMIX matches this specific substrate by design rather than as an arbitrary choice. The shared reward (1) is cooperative and highly compatible with a monotonic factorisation; off-policy replay yields sample efficiency for small fleet sizes; and greedy decentralised execution aligns with beacon-limited observability, requiring no communication beyond beacons; Section VII-C evaluates this against VDN (additive mixing) and MAPPO (on-policy centralised critic) under identical observations, rewards, and budgets.

D. Policy Network Architecture

Each agent's Q-network takes the flattened O_i^+ vector ($|\text{local obs}| + G_{\text{max}} \times |\text{beacon}|$ features plus per-slot staleness and validity), processes it through a fully-connected layer and a 64-unit GRU (Table II), and outputs Q-values via a linear layer over the option set; the recurrent state bridges short beacon gaps such as the 3-interval handoff blackout. The mixing network ingests the global state (training only) and outputs Q_{tot} through one hidden layer with absolute-value activations enforcing monotonicity.

E. Training and Domain Randomisation

Training is performed offline on a server with full state access. Each training episode samples one of {Normal, Poisson, Uniform} with equal probability, forcing distribution-agnostic policies; the communication-side randomisation of Section VI-B is applied simultaneously. Trained actors are then deployed onto each AGV's compute unit; no server interaction occurs during operation.

VII. NUMERICAL EXPERIMENTS

A. Simulation Setup

All experiments use a custom Python warehouse simulator, configured as in Table II. Pickers execute the same scripted policies as in the matching heuristic paradigm, so differences isolate the AGV policy; learned picker policies are left to future work.

B. MARL Training Configuration

All three learned methods (VDN, MAPPO, CTDE-QMIX) are trained with identical observations (Section VI-B), reward (1), environment, and 2,000-episode budget under the domain randomisation of Section VI-E; per-method learning rates come from a small sweep around the shared default in Table II; each method trains with 3 seeds and we report pooled episode statistics.

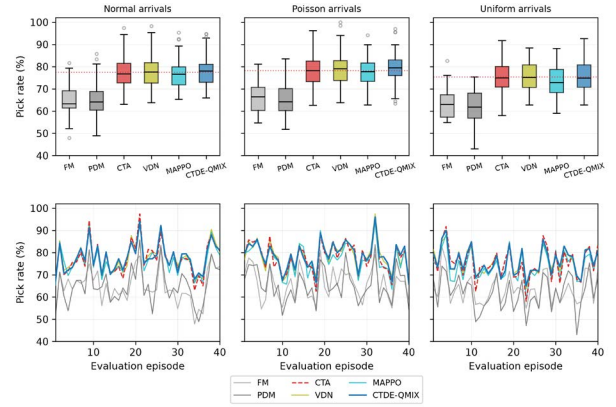


Fig. 1. Pick-rate distributions (top; dotted: CTA mean) and episode trajectories (bottom), 40 evaluation episodes per arrival distribution.

C. Comparative Results

Table I reports pick-rate statistics over 40 evaluation episodes per distribution for the three heuristics, the two MARL baselines, and CTDE-QMIX (mean $\pm$ std and minimum; bold: best per column). Fig. 1 shows the distributions and episode-level trajectories.

Pick-rate improvement. CTDE-QMIX gains +1.3 to +1.7 pp over the strongest heuristic (CTA) in every distribution, and 1.3–1.6 pp over the strongest learned baseline (VDN), locating the benefit in centralised monotonic value mixing.

Robustness and convergence. Episode-level minima are comparable across CTA and the learned methods (58.3–64.4%), but under the Uniform regime CTDE-QMIX raises the worst-case floor to 62.7% against CTA's 58.3% (+4.4 pp), and its global minimum over all 360 evaluation episodes (62.7%) is the highest of any method (VDN 59.8%, CTA/MAPPO 58.3%, FM 47.9%, PDM 43.0%): reduced tail risk where demand is hardest to forecast. Fig. 1 shows the learned policies tracking the CTA trajectory episode-by-episode (shared evaluation seeds), CTDE-QMIX posting the highest mean in every distribution while FM and PDM trail by 12–14 points.

D. Communication Overhead and Energy

Following Sections IV-C–IV-D, we evaluate every uplink, aggregated broadcast, and CLAIM/yield transmission against a simulated infrastructure baseline replaying the identical beacon pattern through a $d_{\text{AP}} = 40$ m AP with unicast fan-out to the *measured* neighbour count k . Averaged over the 1,440 evaluation episodes of the Table I campaign ($N = 5$), the overlay reduced occupied beacon airtime by 32% (analytically 57.8% for persistent $k = 2$; measured is lower as agents are intermittently isolated, $k = 0$, where the modes coincide), device-side radio energy by 31% (analytical: 33%), and network-wide energy including the AP by 62% (analytical: 50–67%); per-controller averages lie within 2 pp. These savings compromise no control quality: CTDE-QMIX actors rely solely on beacons from their local P2P group but deliver the gains of Table I, proving that locality-constrained information suffices.

TABLE I

PICK RATE (% , MEAN $\pm$ STD / MIN). HEURISTICS: 40 EVALUATION EPISODES PER DISTRIBUTION; LEARNED METHODS: SAME 40 EPISODES POOLED OVER 3 SEEDS (120 SCORES). BOLD: BEST PER COLUMN.

Method	Normal ($\mu=10, \sigma=3$)		Poisson ($\lambda=10$)		Uniform [5, 15]	
	Mean (%)	Min (%)	Mean (%)	Min (%)	Mean (%)	Min (%)
FM (PTG heuristic)	64.9 $\pm$ 7.7	47.9	66.1 $\pm$ 7.2	54.7	63.6 $\pm$ 6.8	54.8
PDM (zone heuristic)	65.2 $\pm$ 8.4	48.9	65.5 $\pm$ 7.6	51.8	61.8 $\pm$ 7.8	43.0
CTA (GTP heuristic)	77.8 $\pm$ 7.5	63.3	78.5 $\pm$ 6.9	62.9	75.7 $\pm$ 7.1	58.3
VDN	77.9 $\pm$ 7.3	63.8	78.5 $\pm$ 6.7	61.8	75.7 $\pm$ 6.1	59.8
MAPPO	76.3 $\pm$ 6.7	61.7	77.2 $\pm$ 6.4	61.8	74.5 $\pm$ 6.4	58.3
CTDE-QMIX	79.5$\pm$7.6	64.4	80.0$\pm$6.8	63.7	77.0$\pm$6.7	62.7

TABLE II

SIMULATION PARAMETERS (TOP) AND SHARED CTDE HYPERPARAMETERS (BOTTOM); REPLAY, TARGET-UPDATE, AND ϵ ROWS APPLY TO THE VALUE-BASED METHODS—MAPPO IS ON-POLICY

Warehouse grid / shelves / stations	20 $\times$ 20 cells / 120 / 4
AGVs / pickers	5 / 3
Eval. episodes / steps per episode	40 / 500
Arrivals (items/ T_w , $T_w=65$ steps)	N(10, 3), Pois(10), U[5, 15]
Comm. r_c / beacon T_b / $G_{\max}$	15 m (3 cells) / 100 ms / 8
Beacon loss p (eval. sweep)	{0, 5, 10, 20, 30}%
Handoff λ_h (eval.; 3- T_b blackout)	{0, 1, 2}/agent/100 steps
Learning rate α / discount γ	5×10^{-4} / 0.99
Replay / mini-batch	50,000 trans. (100 ep) / 8 ep, BPTT
Target network update	every 200 gradient steps
GRU / mixing hidden size	64 / 32
Reward weights (w_1, w_2, w_3, w_4)	(1.0, 0.05, 0.5, 0.01)
Training episodes / seeds	2,000 / 3
ϵ -greedy anneal	1.0 $\rightarrow$ 0.05 over 600 episodes
Train-time loss $p_{\max}$	0.2 ($p \sim \mathcal{U}[0, p_{\max}]$, Sec. VI-B)

TABLE III

MEAN PICK RATE (%) VS. BEACON-LOSS PROBABILITY p (I.I.D.), POOLED OVER THE THREE ARRIVAL DISTRIBUTIONS.

Method	$p=0$	0.05	0.10	0.20	0.30
CTA	77.2 $\pm$ 6.8	77.4 $\pm$ 7.0	77.5 $\pm$ 6.6	76.0 $\pm$ 6.0	78.1 $\pm$ 7.6
CTDE-QMIX	77.8 $\pm$ 6.3	76.9 $\pm$ 6.0	78.3 $\pm$ 5.9	78.2 $\pm$ 6.2	77.3 $\pm$ 5.1

E. Robustness to Link Dynamics

We validate the robustness of Section III-C by sweeping the per-link beacon-loss probability p (i.i.d.) and the GO-handoff rate λ_h for CTA and CTDE-QMIX; Table III details this loss sweep. Neither controller exhibits performance degradation across the swept spectrum: CTA, utilizing beacons solely for conflict management, fluctuates non-monotonically within 2.1 pp (76.0–78.1%), while CTDE-QMIX stays within 0.9 pp of its baseline clean-channel mean at each loss level, ending 0.5 pp lower at $p = 0.3$ (77.8% $\rightarrow$ 77.3%); such minor differences lie well inside one episode-level standard deviation. The staleness-aware pipeline in Section VI-B (incorporating validity masking, last-value hold, and recurrent tracking for short gaps) covers sporadic losses before they reach the value network, allowing decentralized execution to survive 30% i.i.d. loss without visible throughput penalties. The GO-handoff evaluation confirms this: at $\lambda_h = 1$ CTA drops by ≈ 1.0 pp and CTDE-QMIX by ≈ 0.4 pp; both recover at $\lambda_h = 2$.

VIII. CONCLUSION

In this work, we proposed a WiFi Direct ad hoc overlay serving as the wireless substrate for MARL-controlled AGV fleets: this overlay eliminates the AP’s double-hop routing and energy costs, matching local-neighborhood CTDE execution

constraints while avoiding any runtime coordinator, and analytically providing 50% lower network-wide per-message energy, 33–46% lower device-side radio energy under transmit power control, and ≈ 57 –74% lower beacon airtime. Empirically, the CTDE-QMIX policy yields average pick rates of 79.5%, 80.0%, and 77.0% under Normal, Poisson, and Uniform demand; this provides +1.3 to +1.7 pp gains over the strongest heuristic (CTA) and 1.3–1.6 pp over the strongest CTDE baseline (VDN), while ensuring a worst-case evaluation episode (62.7%) exceeding every other method’s (VDN: 59.8%, CTA: 58.3%), deviating by at most 0.9 pp in mean under 30% beacon loss. Ongoing research involves dynamic GO selection during battery depletion, investigating the channel congestion interplay with MARL, optimizing picker policies, and constructing physical AGV testbeds.

REFERENCES

- [1] T. T. Mac, C. Copot, D. T. Tran, and R. D. Keyser, “Heuristic approaches in robot path planning: A survey,” *Robot. Auton. Syst.*, vol. 86, pp. 13–28, 2016.
- [2] K. Azadeh *et al.*, “Zoning strategies for human–robot collaborative picking,” *Decis. Sci.*, 2023.
- [3] D. Foad *et al.*, “A systematic literature review of A* pathfinding,” *Procedia Comput. Sci.*, vol. 179, pp. 507–514, 2021.
- [4] S. Biswas, S. G. Anavatti, and M. A. Garratt, “Nearest neighbour based task allocation with multi-agent path planning in dynamic environments,” in *Proc. ICAMIMIA*, 2017, pp. 181–186.
- [5] A. Fellan *et al.*, “Enabling communication technologies for automated unmanned vehicles in industry 4.0,” in *Proc. ICTC*, 2018, pp. 171–176.
- [6] D. Camps-Mur, A. Garcia-Saavedra, and P. Serrano, “Device-to-device communications with Wi-Fi Direct: Overview and experimentation,” *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 96–104, 2013.
- [7] A. Laha *et al.*, “An energy efficient routing protocol for device-to-device based multihop smartphone networks,” in *Proc. IEEE ICC*, 2015, pp. 5448–5453.
- [8] A. Asadi, Q. Wang, and V. Mancuso, “A survey on device-to-device communication in cellular networks,” *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, 2014.
- [9] T. Rashid *et al.*, “QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning,” in *Proc. ICML*, 2018.
- [10] C. Yu *et al.*, “The surprising effectiveness of PPO in cooperative multi-agent games,” in *Proc. NeurIPS Datasets and Benchmarks*, 2022.
- [11] S. Sukhbaatar, A. Szlam, and R. Fergus, “Learning multiagent communication with backpropagation,” in *Proc. NeurIPS*, 2016, pp. 2252–2260.
- [12] P. Sunehag *et al.*, “Value-decomposition networks for cooperative multi-agent learning based on team reward,” in *Proc. AAMAS*, 2018, pp. 2085–2087.
- [13] F. Christianos, L. Schäfer, and S. V. Albrecht, “Shared experience actor-critic for multi-agent reinforcement learning,” in *Proc. NeurIPS*, 2020, pp. 10707–10717.
- [14] A. Krnjaic *et al.*, “Scalable multi-agent reinforcement learning for warehouse logistics with robotic and human co-workers,” in *Proc. IEEE/RSJ IROS*, 2024, pp. 677–684.
- [15] A. Garcia-Saavedra, P. Serrano, A. Banchs, and G. Bianchi, “Energy consumption anatomy of 802.11 devices and its implication on modeling and design,” in *Proc. ACM CoNEXT*, 2012.