COMRAD: A Benchmark for Embodied Cooperative Multi-Agent Reinforcement Learning

author names withheld

Under Review for NExT-Game 2026

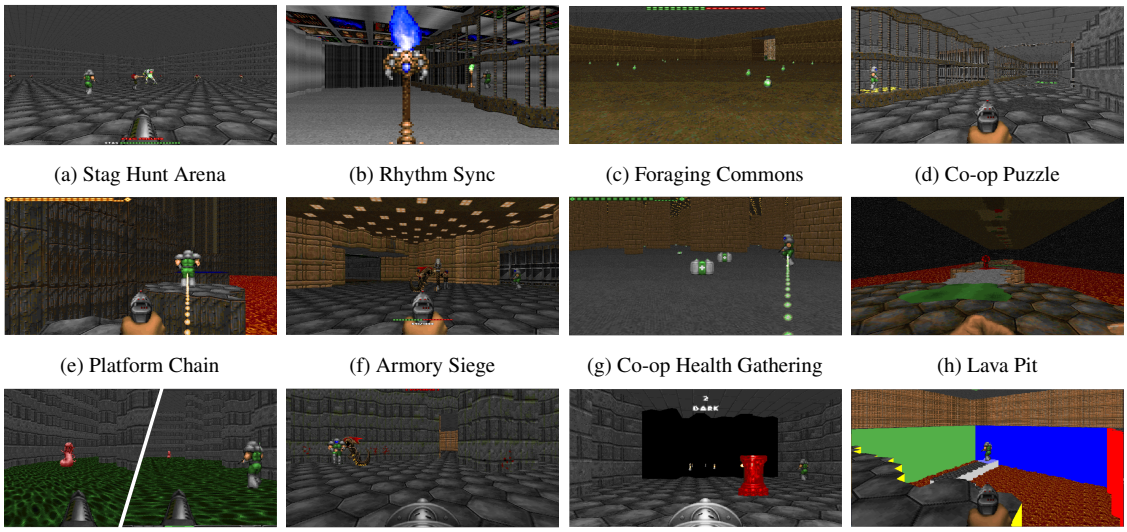
Abstract

Benchmarks are central to the development and evaluation of multi-agent reinforcement learning (MARL) algorithms. As the cooperative MARL community has grown, two categories of evaluation environments have proven indispensable: low-dimensional feature-vector benchmarks that isolate algorithmic behavior in compact state spaces, and two-dimensional pixel-based benchmarks that rely on overhead visual observations. However, to bridge the gap to real-world embodied settings such as robotics and autonomous navigation, algorithms must handle high-dimensional, 3D egocentric visual complexity. While various 3D environments have been explored for vision-based RL, no existing platform simultaneously provides a standardized, purely cooperative multi-agent benchmark with 3D first-person observations and high-throughput simulation. To address this gap, we introduce **COMRAD**, a **CO**operative **M**ulti-Agent **R**einforcement Le**A**rning benchmark suite in **D**oom, featuring a diverse set of challenging scenarios spanning role asymmetry, temporal synchronization, and spatial navigation. To introduce within-scenario variability, we develop **DoomGen**, a procedural map generator that produces diverse layout configurations for every scenario. We integrate COMRAD with Sample Factory, a high-throughput asynchronous RL framework, and implement seven MARL baselines on top of it, reaching $\sim$25K frames per second during training. Our experiments show that COMRAD poses significant challenges for current CTDE methods, establishing visual cooperative MARL as an important open frontier.

1. Introduction

The development of reinforcement learning (RL) has been profoundly shaped by its benchmarks. In single-agent RL, the Arcade Learning Environment, OpenAI Gym, and MuJoCo enabled systematic comparison from pixels and continuous control, respectively [2, 3, 20]. In cooperative multi-agent reinforcement learning (MARL), SMAC and SMACv2 played a similar role for Centralized Training with Decentralized Execution (CTDE), giving the community a common substrate on which methods such as QMIX, MAPPO, and HAPPO could be compared [6, 8, 17, 18, 23]. Well-designed benchmarks, therefore, do not merely measure progress; they define the research frontier, frame the community's shared challenges, and create the conditions for reproducible, comparable science.

Yet benchmark coverage for strategic multi-agent learning from first-person vision remains thin. Most standard CTDE suites use low-dimensional state features or top-down observations, while existing visual environments either emphasize social dilemmas from overhead views or broader embodied tasks such as language grounding and household planning rather than reproducible cooperative control from pixels [1, 4, 9–16, 21]. This leaves a gap for studying learning dynamics in partially observed cooperative games where synchronization, signaling, and role specialization must arise from egocentric local interaction.



(a) Stag Hunt Arena (b) Rhythm Sync (c) Foraging Commons (d) Co-op Puzzle

(e) Platform Chain (f) Armory Siege (g) Co-op Health Gathering (h) Lava Pit

(i) Smart/Dumb Enemies (j) Ammo Carrier (k) Stealth Labyrinth (l) Lava Maze

Figure 1: In-game screenshots from all 13 COMRAD scenarios. Each scenario is presented in a 3D first-person egocentric view, requiring agents to coordinate solely based on their local visual observations.

We address this gap with **COMRAD**, a cooperative MARL benchmark suite built on ViZDoom [7]. COMRAD targets a controlled slice of strategic interaction: fully cooperative partially observed games in which agents must coordinate from egocentric RGB observations while remaining practical enough for multi-seed CTDE evaluation. The suite contains 13 scenarios spanning temporal commitment, resource stewardship, role asymmetry, grounded signaling, and combat coordination. Agents act in factorized discrete action spaces spanning navigation, interaction, and combat primitives, which keeps the benchmark focused on strategic coordination rather than low-level continuous control.

COMRAD makes three contributions. First, it introduces 13 embodied cooperative games together with **DoomGen**, a procedural layout generator that reduces map memorization and enables controlled generalization studies. Second, it provides a unified high-throughput evaluation stack covering seven CTDE baseline families: IPPO, MAPPO, HAPPO, IDQN, VDN, QMIX, and QPLEX. Third, it exposes a structured failure landscape at a common 100M-step frontier: no method family dominates, grounded signaling remains unsolved, and stronger centralized coupling is not uniformly beneficial in asymmetric coordination tasks.

2. COMRAD

COMRAD is a fully cooperative partially observed benchmark built on ViZDoom [7]. Each scenario is packaged as a self-contained WAD plus ACS script controlling geometry, game-tic logic, and objective bookkeeping. Agents connect to a synchronized multiplayer instance, receive egocentric first-person observations, and act through ViZDoom's button interface, while a lightweight Python layer handles preprocessing, reward shaping, and integration with the training stack.

Table 1: Comparison of representative cooperative MARL benchmarks. COMRAD uniquely combines 3D egocentric observations, purely cooperative tasks, procedural layout generation, and integrated CTDE baselines in a single open-source suite (OSS). A standard CTDE baseline trained for 100M environment steps completes in approximately one hour on a single GPU.

Benchmark	**3D/Ego.**	**Vision**	**Coop.**	**ProcGen**	**CTDE Baselines**	**OSS**	N **agents**	**Scenarios**	**Year**
SMAC / SMACv2	✗	✗	✓	✗/✓	✓	✓	2–27	14 / ∞	2019/23
Google Research Football	✗	Opt.	✓	✗	✗	✓	2–11	11+	2019
Overcooked AI	✗	✗	✓	✗	✗	✓	2	5	2019
Melting Pot 2.0	✗	✓	Mixed	✗	✗	✓	2–16	256	2023
JaxMARL	✗	✗	✓	✗	✓	✓	2–10	8+	2024
MEAL	✗	✓	✓	✗	✓	✓	2–4	4	2025
XLand	✓	✓	Mixed	✓	✗	✗	2–16	∞	2021
Watch&Help	✓	✓	✓	✗	✗	✓	2	5	2020
TeamCraft	✓	✓	✓	✗	✗	✓	2	10	2024
COMRAD (Ours)	✓	✓	✓	✓	✓	✓	2–8	**13** / ∞	2026

2.1. Environments and Scenarios

COMRAD provides 13 scenarios spanning several recurring strategic demands: commitment under partial observability, long-horizon resource stewardship, persistent role asymmetry, and grounded low-bandwidth signaling. Table 2 summarizes the benchmark suite, while Appendix **??** gives the full mechanics and reward definitions.

2.2. Observation and Action Spaces

Observations. COMRAD agents act from egocentric RGB observations resized to 128×72 pixels before entering the policy network. This resolution provides sufficient visual fidelity for agents to distinguish objects, teammates, and environmental features while remaining computationally tractable. Several scenarios additionally expose a local measurement vector (*Vision+Vec*) that contains scenario-specific information, such as health, ammo, progress counters, or positional hints, to accelerate learning. These measurements are strictly local to the observing agent and contain no direct information about teammates' states or positions.

Actions. ViZDoom's native button interface is exposed as a factorized discrete action space. Each scenario defines a Cartesian product over task-appropriate primitives spanning movement, rotation, interaction, combat, and weapon selection. This yields a single structured action per decision step. Action space size $|\mathcal{A}|$ ranges from 18 to 162 across the 13 scenarios. This design decouples cooperative strategy learning from low-level continuous control, making algorithmic differences in coordination the primary performance signal.

Rewards. Reward shaping is scenario-specific, but the reporting principle is uniform. Each task defines a benchmark score J that captures the cooperative outcome of interest, such as survival time, checkpoint progress, completed stages, or kills. We report J throughout because it is more comparable across algorithm families than raw shaped return; full reward definitions are deferred to Appendix **??**.

Table 2: Overview of the 13 COMRAD cooperative scenarios and their key properties. *Diff.* groups scenarios by difficulty tier: Easy (1), Medium (3), Difficult (5), and Very Difficult (7). N denotes supported team sizes; all scenarios support at least 2 agents. *Vis+Vec* indicates that egocentric RGB frames are supplemented with a strictly local measurement vector. $|\mathcal{A}|$ is the per-agent factorized discrete action space size. *Objective* is the task-level benchmark score reported throughout the paper.

Diff.	Scenario	N	Obs.	$\lvert\mathcal{A}\rvert$	Objective	Coordination Challenge
1	Stag Hunt Arena	2–4	Vision	54	Maximize stag kills	Agents must resolve the hare-versus-stag coordination dilemma under partial observability.
3	Rhythm Sync	2	Vis+Vec	54	Synchronize strict timings	Separated agents must activate switches in strict temporal windows without communication.
	Foraging Commons	2–4	Vis+Vec	27	Maximize survival time	Balance individual harvesting against replenishing a shared, limited resource pool.
	Co-op Puzzle	2	Vision	18	Completed puzzle pairs	Agents sequentially unlock gates for one another in a strict turn-based progression.
4	Platform Chain	2	Vision	54	Joint checkpoint progress	A physical tether constrains agents to maintain synchronized spatial proximity.
	Armory Siege	2–8	Vis+Vec	162	Maximize survival & kills	A team defends a core while rotating members to collect distant resources.
	Coop Health Gathering	2	Vision	27	Maximize survival time	Agents navigate a decaying health field together to maximize overlapping collection.
5	Lava Pit	2	Vision	36	Joint traversal	Agents alternate holding pressure plates to support each other's crossing over hazards.
	Smart Enemies	2–4	Vision	54	Maximize enemies kills	Enemies accelerate under concentrated fire, forcing agents to maintain dispersed positions.
	Dumb Enemies	2–4	Vision	54	Maximize enemies kills	Enemies flee solo agents but slow near groups, requiring cooperative pursuit patterns.
6	Ammo Carrier	2	Vis+Vec	36	Maximize survival time	A mobile carrier resupplies a stationary defender to maintain continuous defensive fire.
7	Stealth Labyrinth	2	Vision	54	Activate isolated relays	A torch-gunner pair survives shooting enemies under extreme asymmetric observability.
	Lava Maze	2	Vis+Vec	90	Maximize levels cleared	Form emergent communication between an active navigator and a remote information-providing signaler.

3. Experiments

3.1. Baseline Algorithms

We evaluate seven CTDE algorithms spanning the two dominant cooperative MARL families. The on-policy actor-critic baselines are IPPO, MAPPO [23], and HAPPO [8]. The off-policy value-decomposition baselines are IDQN, VDN [19], QMIX [17], and QPLEX [22]. All are adapted to the same asynchronous training stack and evaluated under a shared protocol.

3.2. Baseline Results

We evaluate eight baseline configurations from seven CTDE families under a common 100M-step budget. Table 3 reports the final benchmark scores alongside the theoretical performance ceiling for each scenario. The central result is that no single family dominates, and COMRAD remains far from saturated.

Table 3: Full baseline matrix at the common 100M-step frontier. Entries report mean $\pm$ 95% CI half-width over five independent runs with seeds $\{42, 68, 81, 97, 154\}$. **Ceiling** is the theoretical maximum score achievable under optimal play for each scenario.

Scenario	IPPO	MAPPO	HAPPO	IDQN	VDN	QMIX	QPLEX-D	QPLEX-Q	**Ceiling**
Stag Hunt Arena	17.55 ± 0.22	**25.66 ± 0.32**	2.58 ± 0.10	14.33 ± 0.18	17.96 ± 0.39	0.28 ± 0.06	1.49 ± 0.10	3.42 ± 0.20	36
Rhythm Sync	0.59 ± 0.05	**0.65 ± 0.05**	0.22 ± 0.02	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	1.00
Foraging Commons	2355.64 ± 30.33	2510.93 ± 31.18	2199.89 ± 39.31	2190.70 ± 32.66	2327.13 ± 42.84	**2544.89 ± 46.77**	1905.26 ± 23.66	2345.75 ± 39.38	21000
Co-op Puzzle	3.03 ± 0.07	**3.88 ± 0.09**	2.23 ± 0.14	0.03 ± 0.01	0.07 ± 0.02	0.00 ± 0.01	0.00 ± 0.01	0.07 ± 0.03	5
Platform Chain	2.00 ± 0.02	**4.78 ± 0.19**	4.67 ± 0.18	0.68 ± 0.06	1.12 ± 0.08	0.97 ± 0.02	3.40 ± 0.22	0.97 ± 0.03	47
Armory Siege	16.71 ± 2.90	19.87 ± 2.32	-0.71 ± 1.07	10.97 ± 0.98	18.32 ± 1.46	**23.12 ± 3.83**	1.19 ± 0.89	14.13 ± 2.47	100
Co-op Health Gathering	**1104.90 ± 117.30**	553.96 ± 27.82	197.74 ± 7.98	164.20 ± 2.12	573.83 ± 27.28	166.43 ± 2.07	193.95 ± 5.10	267.78 ± 18.39	8400
Lava Pit	0.92 ± 0.01	**1.00 ± 0.01**	**1.00 ± 0.01**	0.92 ± 0.01	0.80 ± 0.05	0.97 ± 0.03	0.00 ± 0.01	**1.00 ± 0.01**	11
Smart Enemies	11.56 ± 0.86	**14.23 ± 0.65**	6.85 ± 0.38	3.30 ± 0.23	4.89 ± 0.38	0.64 ± 0.08	0.00 ± 0.01	0.65 ± 0.07	84
Dumb Enemies	**12.44 ± 1.93**	2.29 ± 0.20	1.03 ± 0.13	1.07 ± 0.11	2.76 ± 0.22	1.10 ± 0.15	0.00 ± 0.01	1.80 ± 0.16	90
Ammo Carrier	**7052.93 ± 483.17**	1362.73 ± 69.09	1340.99 ± 58.07	3836.13 ± 384.96	2542.51 ± 219.25	2379.46 ± 139.55	1530.98 ± 59.06	2080.97 ± 192.02	8400
Stealth Labyrinth	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	**0.21 ± 0.01**	0.00 ± 0.01	0.00 ± 0.01	1.00
Lava Maze	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	6

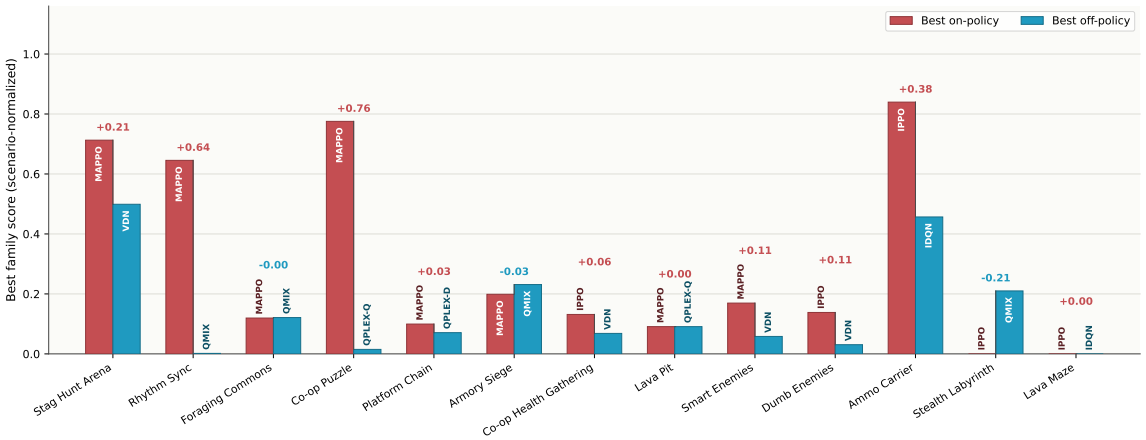


Figure 2: Scenario-normalized gap between the strongest on-policy and off-policy baseline. Each scenario is normalized by its natural score range before comparison.

Performance differences are structured by the type of strategic dependency each task induces. On-policy actor-critic methods are strongest on commitment-heavy synchronization problems such as *Stag Hunt Arena*, *Rhythm Sync*, and *Co-op Puzzle*, with MAPPO the most reliable overall baseline. Off-policy value-decomposition methods are stronger on long-horizon stewardship and defense, with QMIX leading *Foraging Commons* and *Armory Siege*.

The most informative result is not a single winner but a recurring failure landscape. *Lava Maze* remains unsolved by every method, and only QMIX achieves a non-zero score on *Stealth Labyrinth*. These tasks combine severe asymmetric information with grounded behavioral coordination, suggesting that embodied signaling and role-specialized partner modeling remain open problems. A second failure mode appears in *Ammo Carrier*, where decentralized IPPO (**7052.93**) substantially outperforms centralized baselines such as MAPPO (**1362.73**), indicating that stronger joint coupling is not universally beneficial in asymmetric logistics games. Appendix A provides a trajectory-level case study and a fuller capability-oriented interpretation.

3.3. Case Study: Asymmetric Logistics

Ammo Carrier is especially informative because it is not merely hard; it reverses the expected ranking between decentralized and centralized methods. The task assigns permanently differentiated roles: a stationary defender must hold the hub under enemy pressure, while a mobile runner sustains a repeated depot-to-hub resupply loop. IPPO's large margin over MAPPO and HAPPO suggests that stronger joint coupling can interfere with stable role-specialized behavior when one agent must maintain an independent long-horizon routine. We therefore treat *Ammo Carrier* as evidence that centralized training is not uniformly helpful in cooperative games with delayed, role-specific credit assignment.

4. Discussion and Limitations

COMRAD is designed to shift benchmark pressure away from low-dimensional cooperative control toward embodied partial-information games. The main empirical message is that this shift does not simply make every task uniformly harder. Instead, it reveals structured mismatches between standard CTDE inductive biases and different coordination mechanisms: commitment-heavy tasks favor on-policy methods, long-horizon stewardship favors value decomposition, and signaling plus severe role asymmetry remain weak across the board.

The benchmark also makes deliberate scope choices. It studies fully cooperative games rather than mixed-motive incentives, so it isolates learning dynamics inside a fixed joint objective rather than bargaining or mechanism design. Its benchmark scores are scenario-specific, which improves task fidelity but requires caution in cross-scenario aggregation. Finally, the procedural curriculum tooling is infrastructure for controlled experimentation, not a claim that curriculum learning is solved. We therefore view COMRAD as a benchmark for strategic cooperative learning under embodied partial information, and as a foundation for future extensions toward broader game-theoretic settings.

5. Conclusion

We introduced COMRAD, a benchmark for embodied cooperative MARL that pairs 13 first-person ViZDoom scenarios with procedural generation and seven integrated CTDE baselines. Across five-seed experiments at a common 100M-step frontier, COMRAD reveals a structured failure landscape rather than a single dominant method family: commitment-heavy synchronization favors on-policy methods, long-horizon stewardship favors value decomposition, and grounded signaling plus severe role asymmetry remain largely unsolved. We hope COMRAD provides a useful evaluation target for future work on learning dynamics and benchmark design in strategic multi-agent settings.

Impact Statement

COMRAD is intended as a research benchmark for cooperative multi-agent reinforcement learning from embodied visual input. The benchmark may help improve coordination, robustness, and data efficiency in multi-agent systems with applications in theoretical cooperative studies, assistive automation, and distributed control, but the same capabilities could also be adapted to safety-critical or adversarial settings. We therefore view the benchmark as infrastructure for controlled scientific evaluation rather than a claim of readiness for real-world deployment, and we encourage downstream work to pair performance improvements with safety analysis and application-specific oversight.

References

[1] John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, DJ Strouse, Michael B Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs, and Joel Z Leibo. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2022.

[2] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[4] Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, et al. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. *arXiv preprint arXiv:2411.00081*, 2024.

[5] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, K. Larson, and Thore Graepel. Open problems in cooperative ai. *ArXiv*, abs/2012.08630, 2020. URL https://api.semanticscholar.org/CorpusID:229220772.

[6] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Nicolaus Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=5OjLGiJW3u.

[7] Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. Vizdoom: A Doom-based AI research platform for visual reinforcement learning. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8, 2016. doi: 10.1109/CIG.2016.7860433.

[8] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[9] Qian Long, Zhi Li, Ran Gong, Ying Nian Wu, Demetri Terzopoulos, and Xiaofeng Gao. Teamcraft: A benchmark for multi-modal multi-agent systems in minecraft. *arXiv preprint arXiv:2412.05255*, 2024.

[10] Siddharth Nayak, Adelmo Morrison Orozco, Marina Ten Have, Vittal Thirumalai, Jackson Zhang, Darren Chen, Aditya Kapoor, Eric Robinson, Karthik Gopalakrishnan, Brian Ichter, James Harrison, Anuj Mahajan, and Hamsa Balakrishnan. MAP-THOR: Benchmarking long-horizon multi-agent planning frameworks in partially observable environments. In *Multimodal Foundation Model meets Embodied AI Workshop @ ICML2024*, 2024. URL https://openreview.net/forum?id=ZygZN5egzy.

[11] Siddharth Nayak, Adelmo Morrison Orozco, Marina Ten Have, Vittal Thirumalai, Jackson Zhang, Darren Chen, Aditya Kapoor, Eric Robinson, Karthik Gopalakrishnan, James Harrison, Brian Ichter, Anuj Mahajan, and Hamsa Balakrishnan. Llamar: Long-horizon planning for multi-agent robots in partially observable environments, 2025.

[12] Bassel Al Omari, Michael Matthews, Alexander Rutherford, and Jakob Nicolaus Foerster. Multi-agent craftax: Benchmarking open-ended multi-agent reinforcement learning at the hyperscale, 2025. URL https://arxiv.org/abs/2511.04904.

[13] Open Ended Learning Team, Adam Stooke, Feryal Behbahani, Wojciech M. Czarnecki, Marta Garnelo, Arthur Gretton, Arthur Guez, Jonathan Hamel, et al. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

[14] George Papadopoulos, Andreas Kontogiannis, Foteini Papadopoulou, Chaido Poulianou, Ioannis Koumentis, and George Vouros. An extended benchmarking of multi-agent reinforcement learning algorithms in complex fully cooperative tasks. *arXiv preprint arXiv:2502.04773*, 2025.

[15] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021.

[16] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020.

[17] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 4295–4304, 2018.

[18] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft Multi-Agent Challenge. *CoRR*, abs/1902.04043, 2019.

[19] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 2085–2087, 2018.

[20] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[21] Tristan Tomilin, Luka van den Boogaard, Samuel Garcin, Bram Grooten, Meng Fang, and Mykola Pechenizkiy. Meal: A benchmark for continual multi-agent reinforcement learning. In *Proceedings of the Datasets & Benchmarks Track at NeurIPS 2025*, 2025. URL https://arxiv.org/abs/2506.14990.

[22] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex dueling multi-agent q-learning. In *International Conference on Learning Representations (ICLR)*, 2021.

[23] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24611–24624, 2022.

Appendix A. Strategic Failure Analysis

Ammo Carrier produces the largest ranking reversal in the benchmark: decentralized IPPO substantially outperforms every centralized baseline in Table 3. The task assigns permanently differentiated roles. A stationary defender must hold the hub under enemy pressure, while a mobile runner must sustain a repeated depot-to-hub resupply loop. This is a plausible case where stronger centralized coupling can interfere with stable role-specialized behavior when one agent must maintain an independent long-horizon routine with delayed, role-specific payoff.

Following Dafoe et al. [5], the broader pattern can be interpreted along three recurring capability axes. *Stealth Labyrinth* and *Ammo Carrier* stress embodied partner understanding under durable asymmetry. *Lava Maze* isolates grounded low-bandwidth communication and remains unsolved at 100M steps. *Rhythm Sync* and *Platform Chain* stress temporal commitment without direct access to a partner's internal state. We treat these as interpretive summaries rather than causal decompositions.

A.1. Behavioral Analysis: When Centralized Training Hurts

Aggregate benchmark scores reveal *which* algorithms perform best, but not *why* coordination succeeds or fails in a given scenario. We therefore study *Ammo Carrier*, the fixed-role asymmetric defense-and-resupply scenario that produces the largest reversal in the benchmark: decentralized IPPO substantially outperforms every CTDE baseline in Table 3.

Task structure. *Ammo Carrier* assigns agents to permanently differentiated roles: a stationary defender must hold the hub under enemy pressure, while a mobile runner must sustain a repeated depot-to-hub resupply loop. The core challenge is not fine-grained handoff timing, but whether the runner learns a stable long-horizon logistics routine whose payoff is delayed and role-specific.

Behavioral signature of failure. To visualize this qualitatively, Figure 5 compares occupancy density heatmaps from representative rollout trajectories. IPPO concentrates movement along a recurring depot-to-hub corridor, consistent with a learned resupply cycle. HAPPO, by contrast, shows a much less regular pattern: the runner does not consistently maintain the loop, periodically abandoning the logistics routine to move toward the defender or revisiting depots in irregular patterns. The result is episodic ammo starvation and substantially lower survival times. This pattern is consistent with a breakdown in persistent role specialization, but is not by itself diagnostic of which component of the training setup causes that breakdown.

Interpretation. A reasonable explanation is that strict role asymmetry changes what effective coordination looks like under optimization. In tightly coupled joint tasks, centralized training and shared optimization pressure can help agents align on moment-to-moment joint behavior. In *Ammo Carrier*, however, the optimal joint policy also requires one agent to sustain an independent long-horizon logistics loop whose payoff is delayed and role-specific. Centralized updates may therefore introduce counterproductive coordination pressure that disrupts this independence: the runner's gradient updates are influenced by the defender's value signal in a way that interferes with the sustained logistics commitment the scenario demands. A decentralized update, by contrast, allows the runner to optimize its loop policy without cross-agent interference. We treat this as an informed interpretation rather than a causal proof; critic-scope, reward-sharing, and role-factorization ablations would be needed to isolate the effect.

Even so, this failure mode exposes a recurring Cooperative AI challenge: CTDE methods optimized for tightly coupled joint credit assignment may be poorly suited to *role-specialized*

understanding/trust settings, where effective cooperation depends less on tighter joint coupling than on learning stable complementary behavioral contracts. Making this failure class behaviorally visible, rather than merely observable as lower return, is one of the contributions of COMRAD, and we connect it to the broader Cooperative AI taxonomy in Section B.

Appendix B. Cooperative AI Capability Analysis

Following Dafoe et al. [5], we use COMRAD as an interpretive readout over three directly tested Cooperative AI capabilities: *understanding*, *communication*, and *commitment*, plus a narrower lens on *institutions*. The benchmark does not map one-to-one onto these bins; most tasks couple several capabilities at once, and the institutional lens is necessarily indirect. The purpose of this section is not to claim causal decomposition, but to map observed benchmark failures onto capability classes and identify where current methods systematically break down. Detailed scenario descriptions and demonstrations are provided in Appendix **??**.

Understanding. In the taxonomy of Dafoe et al. [5], understanding includes not only predicting the physical consequences of one's actions, but also anticipating other agents' behavior and the private factors that drive it. COMRAD probes this capability in an embodied, first-person form. *Stag Hunt Arena* and *Lava Pit* require partner inference under uncertainty, while *Ammo Carrier* and *Stealth Labyrinth* sharpen the demand through durable role asymmetry. The empirical pattern is consistent: strong baselines cope when partner behavior is structured and locally legible, but they struggle badly when partner behavior must be inferred under extreme asymmetry and persistent occlusion, as in *Stealth Labyrinth*. We therefore read these failures primarily as limits in embodied partner modeling.

Communication. For Dafoe et al. [5], communication is the explicit exchange of information that improves coordination, with common ground and channel constraints playing central roles. COMRAD intentionally removes an explicit message channel, so the relevant question is narrower: can agents improvise communication through behavior or exploit weak environment-mediated cues? *Lava Maze* is the cleanest test: one agent has privileged route information and must convey it through constrained color-coded actions to a remote navigator. Every baseline remains at zero at 100M steps, which is the strongest evidence in the paper that low-bandwidth emergent communication from pixels remains unsolved for current CTDE methods. *Smart Enemies* and *Dumb Enemies* are easier because the signal is not produced by a partner but by the shared enemy dynamics induced by team positioning. The gap between those tasks suggests that reading a shared external signal is substantially easier than inventing and grounding a new one. Thus, some environment-mediated common ground is learnable, but partner-originated communication remains essentially unsolved.

Commitment. Commitment is the ability to make cooperative behavior credible, often through commitment devices or protocols that prevent profitable deviation. *Rhythm Sync* is the clearest example: spatially separated agents must repeatedly act within narrow time windows, so coordination depends on sustaining an internalized joint routine. Every off-policy baseline remains at or near zero, whereas MAPPO reaches **0.65**, suggesting that temporal commitment without direct mutual observation is a real bottleneck for current value-decomposition methods. *Stag Hunt Arena* reinforces the same message: MAPPO leads, but the task remains far from saturated, so even strong on-policy methods do not robustly maintain the payoff-dominant convention. *Platform Chain* is especially informative because the chain acts as a hardware-like commitment device in the sense discussed by Dafoe et al. [5]: the environment constrains unilateral deviation, yet substantial coordination

difficulty remains. Taken together, current baselines do better when commitment is scaffolded by environmental structure, such as a chain or repeated rhythm, than when it must be maintained primarily through learned convention over long horizons.

Institutions. Institutions, by Dafoe et al. [5], are the rules, bargains, and incentive structures that shape the cooperative game itself. COMRAD does not test institution formation, bargaining, or mechanism design directly. Instead, *Foraging Commons* and *Armory Siege* place agents inside fixed resource-allocation rules that reward stewardship over myopic greed. QMIX's strong performance on both suggests that value decomposition can make partial progress on long-horizon collective credit assignment when the rule structure is fixed and legible. These scenarios, therefore, study behavior *within* institutional scaffolds rather than institutions in the fuller Cooperative-AI sense; this is a deliberate scope boundary.

What COMRAD presently shows. The resulting picture spans all four Cooperative AI capabilities. Understanding under severe role asymmetry, communication through grounded behavioral signals, and commitment without structural support remain substantially weak. Institutions show a distinct pattern: value-decomposition methods can partially handle long-horizon collective credit assignment under fixed rule structures, making this the one capability in which current baselines show consistent progress. A recurring failure mode across the first three capabilities is what we informally call *role-asymmetric trust*: in the taxonomy of Dafoe et al. [5], this is better understood as a composite demand on understanding and commitment, sometimes exacerbated by the absence of communication, rather than as a fifth standalone capability. Taken together, COMRAD distinguishes which cooperative capabilities are tractable in the egocentric visual setting from those that remain open challenges. The specific failure modes exposed by this benchmark provide clear directions for future algorithm and training-objective design, which we discuss in Section 4.

[t]0.49

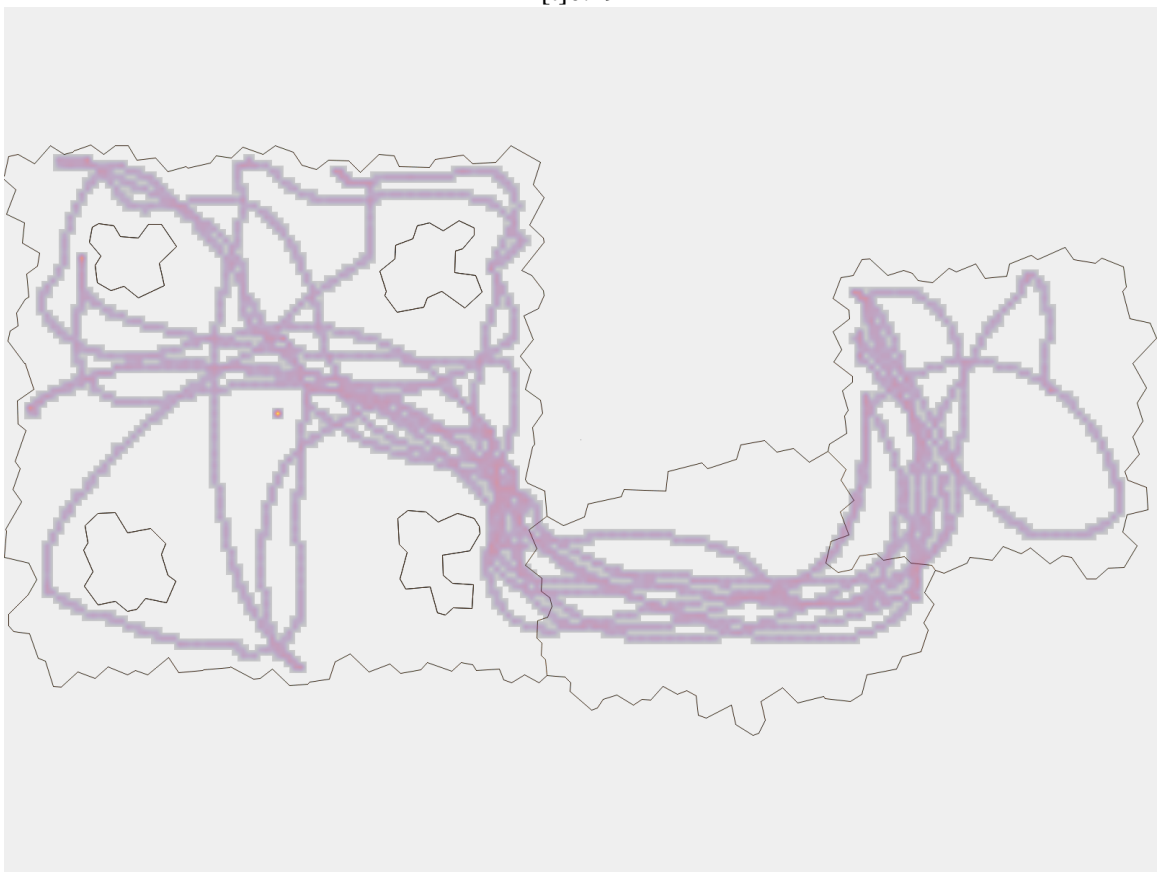


Figure 3: IPPO. The runner repeatedly traverses the corridor between the depot region and the defender hub, consistent with a learned resupply loop.

[t]0.49