Learned Coordination Conventions in Cooperative MARL: Measuring the Translation Gap Between Theory-Informed Roles and Learned Routing

author names withheld

Under Review for NExT-Game 2026

Keywords: multi-agent reinforcement learning, role-conditioned attention, game-theoretic learning, equilibrium selection, strategic AI

Abstract

Role-semantic assignments supply priors over how heterogeneous agents may coordinate; cooperative MARL agents instead settle on conventions through decentralized, non-stationary learning, with no guarantee the resulting structure matches those priors – an equilibrium selection question that theory alone cannot resolve. We ask whether the convention selected by training is legible from the policy architecture, and how it relates to a theory-informed role-semantic prior, using a diagnostic that combines a role-routing matrix, formation sensitivity ($\Delta_{\max}$), and gradient/occlusion attribution across three-role MiniGrid and SMACv2 (Terran), five MAPPO conditions, and a 3v3–9v9 scaling study.

Label-conditioned attention produces substantially more concentrated and role-specific routing than flat MLP baselines (entry std 0.246 vs. 0.055); the signature is stable under scaling, transfers zero-shot from 3v3 to 9v9 above the from-scratch baseline (0.224 vs. 0.110), and is invariant to ally-slot padding. Removing role labels under shared attention costs 14.0 pp. Under 5-seed re-evaluation, three of four theory-informed predictions hold and one is tied; the earlier 3-seed Strike→Vanguard finding was a small-n artifact, showing the diagnostic is itself seed-sensitive. We read this as evidence that an architecture-exposed routing channel makes aspects of the learned convention more directly measurable – contributing an empirical evaluation framework for coordination structure in cooperative MARL, not a new solution concept.

1. Introduction

A central question for game-theoretic learning is no longer only *what equilibrium should exist*, but *what coordination convention a learning system actually settles on* – an equilibrium selection question that theory alone cannot resolve. Classical role-based reasoning supplies strategic priors under explicit rationality assumptions; cooperative MARL agents instead select a convention through decentralized gradient updates in non-stationary environments, with no a priori reason the resulting role-conditioned structure should align with those priors – what we term a translation gap between theory-informed expectations and learned coordination structure.

We treat this as a question of coordination legibility: whether the role-conditioned interaction structure of a trained policy is measurable from its architecture, and how it relates to a theory-informed role-semantic prior. When role labels are injected into both `self_tok` and `ally_tok`, cross-attention exposes role-pair patterns to direct inspection. Our central claim is that this pathway yields a more concentrated, role-specific signature than flat MLP baselines and is robust to scaling, cross-scale transfer, and padding – a measurable architectural property of the learned convention, not a claim that attention causally explains behavior or that the convention is an equilibrium.

Contributions: (i) an empirical evaluation framework (role-routing matrix, $\Delta_{\max}$, and gradient/occlusion-based routing analysis) for measuring learned coordination conventions from trained cooperative MARL policies; (ii) evidence on three-role MiniGrid and SMACv2 (Terran) at 3v3–9v9 that label-conditioned attention yields more concentrated routing, stronger cross-scale zero-shot transfer, and padding-invariant behavior than MLP baselines; (iii) a 5-seed re-evaluation showing partial alignment between learned conventions and designer-specified priors, while exposing where small-n seed noise manufactures apparent strategic divergence.

2. Related Work

ROMA [8] learns latent role representations; we take the complementary view, analyzing how architectures exploit pre-defined role labels. Zambaldi et al. [11] use self-attention as a relational inductive bias for single-agent RL; entity-centric observations [1] provide the input format. Cooperative optimizers (QMIX [7], MAPPO [10]) lack a structural path for role-pair routing, and prior attention-based MARL [5, 6] targets inter-agent communication rather than role-conditioned entity selection. RODE [9] and UPDeT [4] target performance and transfer; Other-Play [2] and Off-Belief Learning [3] address equilibrium selection by shaping which convention is learned. We instead hold the algorithm fixed and ask whether the selected convention becomes *inspectable*, and how it relates to a role-semantic designer prior.

3. Mechanism: Role-Label-Conditioned Routing

Routing path. In an MLP, role labels are concatenated with other features and "which entity to condition on given my role" is implicit in weight products – a weaker structural prior, not an impossibility. In our architecture, role labels appear in both `self_tok[4:7]` (encoded into the Query) and each `ally_tok` role-onehot block (encoded into Key/Value), so the cross-attention score $\text{softmax}((W_Q\mathbf{x}_i)^\top(W_K\mathbf{x}_j)/\sqrt{d_k})$ becomes a direct function of the role-pair (r_i, r_j). Cross-attention is applied per entity group (ally, enemy, zone) with the self-embedding as query; our routing analysis isolates the ally branch. An Intra-Set Self-Attention block enriches each ally token before the cross-attention Key. We treat these weights as architecture-exposed coordination patterns; attribution and occlusion analyses (§5) provide complementary, less interpretation-laden views.

Why these predictions? Under the designer-specified role semantics, Strike (ranged damage) should benefit most from Support, Vanguard (frontline) should react most to formation changes, and Support should distribute attention more uniformly while tracking team state.

Testable predictions. (1) attention produces more concentrated role routing than MLP – routing-matrix entry std at least $2\times$ the MLP gradient-attribution std **[PASS,** $4.5\times$**:** 0.246 **vs.** 0.055**, 5 seeds]**; (2) under shared-attention, zeroing role labels degrades performance by >5 pp **[PASS,** -14.0 **pp; bundles label removal with policy sharing – upper-bounds rather than isolates the label effect]**; (3) $\Delta_{\max}$ is highest for Vanguard **[TIED: Support** 0.074**, Vanguard** 0.066**, gap** $<$ **either seed std; qualitatively recovers "non-DPS roles are formation-sensitive"]**; (4) Strike attends preferentially to Support **[PASS in** $4/5$ **seeds at both 3v3 and 9v9; the 3-seed Strike**$\rightarrow$**Vanguard result was a small-**n **artifact]**.

4. Method

Environments. **Domain 1** is a custom 15×15 MiniGrid task with three role-specialized agents (Strike, Vanguard, Support) cooperating to capture and hold zones under role-specific reward shaping. **Domain 2** is a custom SMACv2 scenario using three Terran unit types as role proxies (Marine = Strike, Marauder = Vanguard, Medivac = Support). SMACv2 does not natively expose role labels;

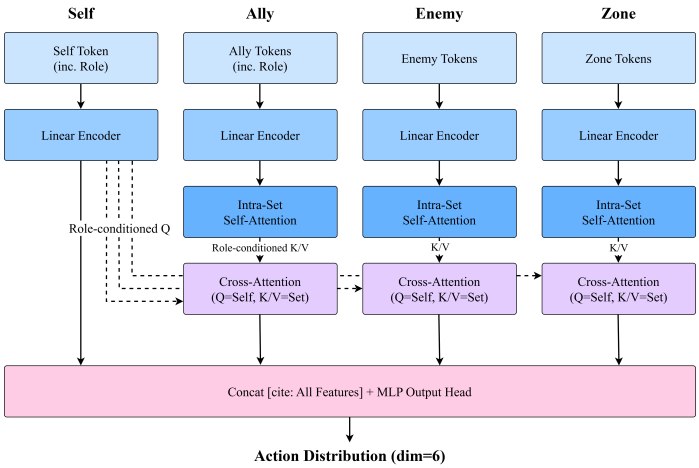

Figure 1: SLIC architecture (Cond. 1). Per-set encoders feed Intra-Set Self-Attention, then Cross-Attention with the self-embedding as query. The role-conditioned routing path is highlighted with thicker dashed arrows.

we inject a 3-dim unit-type one-hot into both the agent's own token and each ally token, preserving the structural condition required for role-conditioned Q/K routing.

Entity-centric observation. Agents receive four named token groups – `self_tok`, `ally_tok`, `enemy_tok`, `zone_tok` – plus validity masks. The role one-hot (3 dims) appears in both `self_tok` and each `ally_tok` slot; this placement is load-bearing for the routing mechanism. The 3v3 SMACv2 observation is 65-dim (182-dim padded for the scaling study).

Five ablation conditions. The conditions span two axes: *architecture* (attention vs. MLP) and *role-label + policy sharing*. **C1 Full**: per-set encoders (dim 64) + Intra-Set Self-Attn (residual+LN) + Cross-Attn with self as query; per-role policies, labels present. **C2**: C1 minus Intra-Set blocks. **C3 Role+MLP**: flat MLP, per-role, labels in input. **C4 No-Role (shared attention)**: C1 architecture, role one-hots zeroed, single shared policy. **C5 Flat MLP**: C3 with labels zeroed. C1 vs. C4 bundles label removal with a per-role→shared-policy change (upper bound on label effect); C3 vs. C5 isolates label-in-input for MLP ($+1.6$ pp, within noise); C4 vs. C5 confounds both axes plus a $3\times$ data-per-parameter advantage from sharing. All attention applies `key_padding_mask` with slot-0 force-unmask safety. Training uses MAPPO [10] with identical hyperparameters (2M steps; 5 seeds for all 3v3 conditions and 9v9 C1/C3, 3 seeds for 6v6 C1/C3; GAE λ=0.95, γ=0.99, clip 0.2, entropy 0.01, Adam 3×10^{-4}, batch 512); reward mix $r = 0.8r_{\text{ind}} + 0.2\bar{r}_{\text{team}}$.

5. Experiments

5.1. Track 1: Performance Comparison (SMACv2 3v3)

Three observations matter. Attention helps: C1 outperforms C3 by 31.4 pp. Intra-set context helps: C2 trails C1 by 13.8 pp. Removing labels under shared-attention is associated with a 14.0 pp drop (C4 vs. C1) – but this contrast also swaps per-role for shared policies, so it upper-bounds rather than isolates the label effect. C3 and C5 are statistically indistinguishable (0.480 vs. 0.464): adding labels to an MLP does not by itself produce useful role-conditioned routing. Expanding all 3v3 conditions to 5 seeds preserves the sign of every contrast.

Table 1: Track 1 win rate on SMACv2 3v3 (mean±std over 5 seeds of the last-10 evaluation average). **Bold** marks the highest mean.

	C1: Full	C2: No Self-Attn	C3: Role+MLP	C4: No Role (shared)	C5: Flat MLP
Win rate	$\mathbf{0.794 \pm 0.033}$	0.656 ± 0.104	0.480 ± 0.155	0.654 ± 0.087	0.464 ± 0.049
n seeds	5	5	5	5	5

Table 2: Scaling results (mean±std). 3v3 and 9v9 now use 5 seeds for both C1 and C3; 6v6 remains 3 seeds.

Scale	C1 Last-10	C3 Last-10	C1 AUC	C3 AUC
3v3	$\mathbf{0.794 \pm 0.033}$	0.480 ± 0.155	$\mathbf{0.678 \pm 0.019}$	0.328 ± 0.059
6v6	$\mathbf{0.380 \pm 0.098}$	0.363 ± 0.144	$\mathbf{0.379 \pm 0.064}$	0.311 ± 0.098
9v9	$\mathbf{0.110 \pm 0.090}$	0.060 ± 0.032	$\mathbf{0.119 \pm 0.067}$	0.089 ± 0.052

5.2. Scaling Study: 3v3, 6v6, 9v9

We test whether the C1 vs. C3 contrast survives larger team sizes. We train balanced 3v3, 6v6, and 9v9 variants with equal Strike/Vanguard/Support counts; enemy composition scales in parallel while preserving the asymmetric design (no enemy Support unit). All scaling runs use identical hyperparameters and a centralized critic over concatenated global observations.

In 3v3, C1 is clearly better than C3. In 6v6/9v9 the full model remains ahead on AUC and late-training averages, but margins shrink and seed variance grows – the 6v6 (0.380 vs. 0.363) and 9v9 (0.110 vs. 0.060) gaps are within or close to seed std and we do not claim significance. The performance claim is therefore modest: attention improves small-team performance and shows a noisy directional advantage at scale. The structural claim is stronger: role-conditioned routing persists at 9v9 (read from off-diagonal contrasts as magnitudes compress with more slots), cross-scale transfer remains positive, and the model is invariant to padded slots.

Cross-scale zero-shot transfer. C1's policy is invariant to ally count by construction. A 3v3-trained C1 checkpoint evaluated zero-shot on 9v9 (50 episodes per seed, 5 seeds) reaches 0.224 ± 0.088 – above from-scratch 9v9 C1 (0.110 ± 0.090) and far above 9v9 C3 (0.060 ± 0.032). C3 transfers poorly throughout (~ 0.084), consistent with MLP weights memorizing a fixed slot pattern.

Padding-confound ablation. A potential confound for the 3v3 gap is that C3 must process zero-padded ally slots that C1's mask blocks. Across fill strategies (Table 4), C1 is bit-exact invariant for data-driven fills and shifts negligibly under noise (KL ≤ 0.021); C3 responds non-monotonically. The best C3 fill recovers $+10.7$ pp of the 31.4 pp gap, bounding the padding contribution above by $\sim 34\%$ and leaving the majority as architectural residual.

5.3. Track 2: Role-to-Role Routing

Formation sensitivity and routing matrix. Two fixed scenarios – *Clustered* (allies near one landmark) and *Spread* (each ally at a distinct landmark) – define $\Delta_{\max}$ per role as the max difference in ally self-attention weight. C1 with 5 seeds yields $\Delta_{\max} = 0.024 \pm 0.017, 0.066 \pm 0.039, 0.074 \pm 0.029$ for Strike, Vanguard, Support: Support and Vanguard are statistically tied at the top, both well above Strike. Prediction 3 (Vanguard strictly highest) is therefore not strictly supported in mean ordering ($\Delta_{\text{Support}} > \Delta_{\text{Vanguard}}$ by 0.008, smaller than either std), but the qualitative design intent – the two non-DPS roles are formation-sensitive while Strike is not – is recovered. The routing

Table 3: Zero-shot transfer of 3v3-trained checkpoints, evaluated without retraining (50 fresh episodes per seed). Win rate is mean±std over 5 seeds. The 3v3 row (0.796 ± 0.054) differs slightly from Table 1 (0.794 ± 0.033) because Table 1 reports the last-10 evaluation average from training, whereas this table uses 50 fresh post-training rollouts.

Target	Cond. 1	Cond. 3	Gap
3v3	$\mathbf{0.796\pm0.054}$	0.468 ± 0.153	32.8 pp
6v6	$\mathbf{0.212\pm0.046}$	0.084 ± 0.033	12.8 pp
9v9	$\mathbf{0.224\pm0.088}$	0.084 ± 0.043	14.0 pp

Table 4: Padding-confound ablation at 3v3 (5 seeds, 30 episodes per seed). KL is divergence from the zero-fill baseline.

Fill strategy	C3 KL	C3 win rate
zeros	0.000	0.460 ± 0.138
noise σ=0.05	0.007	0.500 ± 0.078
noise σ=0.20	0.136	$\mathbf{0.567\pm0.127}$
mean valid ally	0.096	0.407 ± 0.171
copy valid ally	0.101	0.453 ± 0.209

Table 5: Cond. 1 role-routing matrix (5 seeds). Entry (i, j) is mean cross-attention weight from role i to ally slots occupied by role j, ± across-seed std. Diagonals are zero by construction.

Agent \ Ally	Strike	Vanguard	Support
Strike	—	0.473 ± 0.072	$\mathbf{0.527\pm0.072}$
Vanguard	$\mathbf{0.565\pm0.203}$	—	0.435 ± 0.203
Support	$\mathbf{0.633\pm0.200}$	0.367 ± 0.200	—

matrix (Table 5) extracts the self-query's cross-attention weight per ally slot, tagged by occupant role; slot-ordering bias is ruled out by reversal.

The matrix is concentrated, role-specific, and stably off-diagonal. Strike's largest routing weight is on Support, consistent with prediction 4 in $4/5$ seeds at both 3v3 and 9v9; the earlier 3-seed run flipped on a single seed under retraining-equivalent stochasticity. Vanguard and Support both route primarily toward Strike, an architecture-exposed pattern consistent with the two non-DPS roles attending to the engagement-driver.

MLP gradient attribution. For Cond. 3 we compute input gradients of the maximum action logit, separately measuring mean magnitude on spatial features and role-onehot dims. The Spatial/Role ratio is $5.8\times$/$2.3\times$/$3.7\times$ (Strike/Vanguard/Support): the MLP registers some role signal but spatial features dominate. The routing-matrix entry std is 0.246 (Cond. 1) versus 0.055 for the gradient-based pseudo-routing (Cond. 3), a $4.5\times$ concentration gap. We use std as a scale-comparable concentration proxy (zero under uniform routing, larger when a few role-pair entries dominate); row-entropy and KL-from-uniform are reasonable alternatives, and the magnitude of the gap makes the qualitative ranking robust to the choice.

Unified routing metric: occlusion. Attention weights and MLP gradients are not strictly apples-to-apples, so we additionally measure routing via *occlusion sensitivity* – zeroing each ally token (mask bit kept at 1) and measuring policy KL from baseline – which is well-defined for both architectures. On C1 at 3v3, occlusion correlates with attention at row-normalized Pearson r=0.976 (raw r=0.655). On C3 at 9v9, occlusion produces erratic magnitudes one to two orders larger than C1 (max-row KL ≈ 5.6 vs. ≈ 0.08), with the dominant slot varying across seeds rather than being role-conditioned: the MLP is highly sensitive to slot contents but not in a role-pair-structured way.

Domain 1 replication (exploratory, 3 seeds). On MiniGrid, Cond. 1 yields entry std 0.252, quantitatively matching SMACv2 (0.246). The Strike→X preference direction differs (Vanguard-

dominant on MiniGrid vs. Support-dominant on SMACv2), reflecting different role semantics; the replication target is concentration, not the specific role pair.

Cond. 2 routing. Cond. 2 gives entry std 0.286, comparable to Cond. 1 and far above Cond. 3: cross-attention alone is sufficient for routing *concentration*. However, Cond. 2 routes Strike→Vanguard in $4/5$ seeds (0.751 vs. 0.249), opposite to the role-semantic prior and to Cond. 1, consistent with intra-set blocks influencing which role-pair routing is selected.

6. Discussion

Summary. Across two domains, five conditions, and three team sizes, label-conditioned attention yields a stable, role-specific coordination signature: $4.5\times$ more concentrated routing than MLP, positive zero-shot 3v3→9v9 transfer, padding-invariance, and a 14.0 pp gap when labels are removed under shared attention. The signature is robust where win-rate is not – 6v6/9v9 gaps shrink into noise while routing structure persists – and 5-seed re-evaluation aligns 3/4 prior-derived predictions (one tied), reversing two small-n artifacts. Cond. 2 suggests cross-attention supplies *concentration* while intra-set blocks bias *which* role-pair is selected.

Broader implication (speculative). Beyond the specific role-pair setting studied here, the proposed diagnostic offers a general template for measuring whether any learned convention is legible from policy architecture – a property that becomes increasingly relevant as agentic AI systems coordinate in open-ended environments without designer-specified equilibria. Scalable coordination diagnostics of this kind may complement reward-based evaluation in settings where strategic structure, not performance alone, determines system safety and interpretability. We offer this as a future direction, not a demonstrated result.

Limitations. Both domains share an entity-centric schema and three roles; SMACv2 evidence is restricted to Terran and 3v3–9v9. Attention weights are not causal explanations – we report them as architecture-exposed patterns and corroborate via gradient/occlusion attribution; the entry-std and occlusion (r=0.976) metrics are useful but heuristic. Role labels are manually injected; deployments without a designer schema must pair the routing pathway with role discovery (e.g., RODE [9]). The C1 vs. C4 contrast bundles label removal with policy sharing, upper-bounding rather than isolating the label effect, and the Cond. 2 ablation also perturbs capacity. Domain 1 Track 2 and 6v6 C1/C3 are 3-seed exploratory runs; all others use 5. We do not claim the learned convention is an equilibrium or optimal under any solution concept, nor that alignment with the prior holds beyond these designer-specified roles – the contribution is a measurement procedure, not an optimality result.

7. Conclusion

We introduced an empirical diagnostic for the coordination conventions selected by cooperative MARL – a role-routing matrix, formation sensitivity, and a unified occlusion metric – and used it to compare the learned convention against a theory-informed role-semantic prior. Label-conditioned attention produces a substantially more concentrated, role-specific signature than MLP baselines, stable under scaling and zero-shot 3v3→9v9 transfer. A 5-seed re-evaluation aligns 3/4 prior-derived predictions (one tied) and reveals two earlier divergences as small-n artifacts, underscoring the need for seed-aware structural evaluation. The contribution is a measurement procedure, not an equilibrium or causal claim; future work should extend it to learned role schemas and non-Terran domains.

References

[1] Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

[2] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "Other-Play" for zero-shot coordination. In *Proceedings of the International Conference on Machine Learning*, 2020.

[3] Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief learning. In *Proceedings of the International Conference on Machine Learning*, 2021.

[4] Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. UPDeT: Universal multi-agent reinforcement learning via policy decoupling with transformers. In *International Conference on Learning Representations*, 2021.

[5] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2019.

[6] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 2017.

[7] 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 International Conference on Machine Learning*, 2018.

[8] Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. Roma: Multi-agent reinforcement learning with emergent roles. In *Proceedings of the International Conference on Machine Learning*, 2020.

[9] Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. RODE: Learning roles to decompose multi-agent tasks. In *International Conference on Learning Representations*, 2021.

[10] 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*, 2022.

[11] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew Botvinick, Oriol Vinyals, and Peter Battaglia. Deep reinforcement learning with relational inductive biases. In *International Conference on Learning Representations*, 2019.