ICML 2026 WORKSHOP · COOPERATIVE MULTI-AGENT RL

Learned Coordination Conventions in Cooperative MARL

Measuring the Translation Gap Between Theory-Informed Roles and Learned Routing

INSTITUTION LOGO

QR CODE (paper + code)

Yoosung Hong

Independent Researcher · yoosunghong.main@gmail.com · github: yoosung.dev

1. 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.

We ask whether the convention selected by training is **legible from the policy architecture.** Our diagnostic combines a **role-routing matrix**, a **formation-sensitivity** score ($\Delta_{\max}$), and **gradient / occlusion attribution**, evaluated across MiniGrid and SMACv2 (Terran), five MAPPO conditions, and a 3v3–9v9 scaling study.

Label-conditioned attention produces **4.5× more concentrated** role routing than flat MLPs (entry std 0.246 vs. 0.055). The signature survives scaling, transfers **zero-shot** 3v3→9v9 above the from-scratch baseline (0.224 vs. 0.110), and is invariant to ally-slot padding. Under 5-seed re-evaluation **3 of 4** theory-informed predictions hold and one is tied — the diagnostic is itself seed-sensitive.

2. INTRODUCTION & MOTIVATION

The central question is no longer **what equilibrium should exist,** but **what convention a learning system actually settles on** — an equilibrium selection problem that theory alone cannot resolve.

The translation gap. Classical role reasoning gives strategic priors under explicit rationality assumptions. MARL agents select a convention via decentralized gradient updates in non-stationary environments. There is no a priori reason the two should align.

Where existing cooperative MARL falls short:

- **QMIX / MAPPO expose no structural path for role-pair routing.** Credit is assigned to *agents*, never to *role-to-role interactions* — the coordination structure dissolves into weight products.
- **Flat MLPs make "which entity do I condition on, given my role?" implicit.** Concatenating a role one-hot does not create a readable routing channel: C3 vs. C5 differ by only $+1.6$ pp, within noise.
- **Interpretability is retrofitted, not architectural.** Post-hoc saliency over a monolithic policy cannot separate spatial from role-semantic contributions.
- **Structure outlives performance.** At 6v6/9v9 the win-rate gaps collapse into seed noise, yet the routing signature persists — performance alone is a lossy probe of coordination.

Why role-conditioned routing is the right unit of analysis. Placing the role one-hot in **both** `self_tok` (→ Query) and every `ally_tok` slot (→ Key/Value) makes the attention score a **direct function of the role pair** (r_i, r_j) — turning the learned convention into something you can **read off a matrix** rather than infer from behaviour.

SETUP — DOMAINS & FIVE CONDITIONS

Domain 1 — MiniGrid. Custom 15×15 zone-capture task; three role-specialized agents under role-specific reward shaping.
Domain 2 — SMACv2 (Terran). Marine = **Strike**, Marauder = **Vanguard**, Medivac = **Support**. SMACv2 exposes no role labels, so we inject a 3-dim unit-type one-hot into the agent's own token **and** every ally token — preserving the structural condition the routing mechanism requires.

Cond.	Architecture	Labels	Policy
C1 Full	Intra-Set + Cross-Attn	✓	per-role
C2	Cross-Attn only	✓	per-role
C3 Role+MLP	Flat MLP	✓	per-role
C4 No-Role	Intra-Set + Cross-Attn	✗	**shared**
C5 Flat MLP	Flat MLP	✗	per-role

Read the ablations carefully. C1 vs. C4 bundles label removal with a per-role → shared-policy change — it **upper-bounds** the label effect rather than isolating it. C3 vs. C5 is the clean label-in-input test for MLPs: $+1.6$ pp, within noise. C4 vs. C5 additionally confounds a $3\times$ data-per-parameter advantage from sharing.

THEORY-INFORMED PREDICTIONS

Derived from designer role semantics, then tested over **5 seeds.**

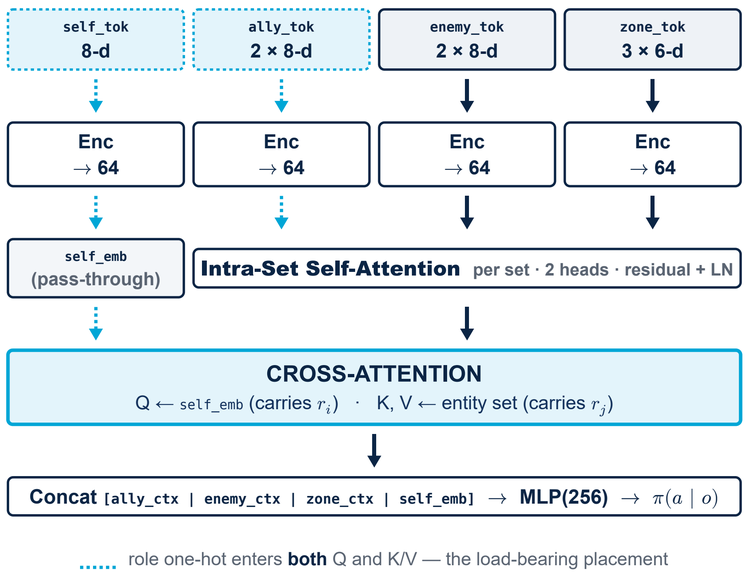
PASS **P1 — Concentration.** Routing entry std $\geq 2\times$ the MLP gradient-attribution std.
Observed **4.5×** — 0.246 vs. 0.055.

PASS **P2 — Label necessity.** Zeroing role labels under shared attention costs > 5 pp.
Observed **−14.0 pp.** Bundles label removal with policy sharing → an **upper bound**, not an isolated effect.

TIED **P3 — Formation sensitivity.** $\Delta_{\max}$ strictly highest for Vanguard.
Support 0.074, Vanguard 0.066 — gap smaller than either seed std. The qualitative intent (**non-DPS roles are formation-sensitive**; Strike 0.024 is not) is recovered.

PASS **P4 — Strike → Support.** Strike routes preferentially to Support.
Holds in **4/5 seeds** at both 3v3 and 9v9. The earlier 3-seed Strike→Vanguard result was a **small-n artifact.**

3. ARCHITECTURE (COND. 1 — FULL)



FORMULATION

Role-conditioned routing. Because the role one-hot r_i is a sub-block of x_i **and** r_j of x_j, the attention score is a direct function of the **role pair**:

$$\alpha_{ij} = \mathrm{softmax}_j\left(\frac{(W_Q x_i)^\top (W_K x_j)}{\sqrt{d_k}}\right)$$

Role-routing matrix. Entry (ρ, σ) averages the self-query's cross-attention over ally slots occupied by role σ; the diagonal is zero by construction.

$$M_{\rho\sigma} = \frac{1}{|S_\sigma|}\sum_{j \in S_\sigma} \alpha_{ij}, \quad i : r_i = \rho$$

$$\text{concentration} = \mathrm{std}\left(\{M_{\rho\sigma}\}_{\rho \neq \sigma}\right)$$

Formation sensitivity. Maximum shift in ally self-attention between the two fixed probe scenarios — **Clustered** vs. **Spread**:

$$\Delta_{\max}^{(\rho)} = \max_{s \in \mathcal{S}} \left| a_s^{\text{clust}} - a_s^{\text{spread}} \right|$$

Training. MAPPO (PPO + centralized critic), 2M steps, 5 seeds, and **identical hyperparameters across all five conditions** — the comparison is structural, not tuned. GAE $\lambda = 0.95$, $\gamma = 0.99$, clip 0.2, entropy 0.01, Adam 3×10^{-4}, batch 512. Reward mixing $r = 0.8\, r_{\text{ind}} + 0.2\, \bar{r}_{\text{team}}$.

4. KEY COMPONENTS

Intra-Set Self-Attention *enriches the Key*

"Who else is on my side?" Each ally token is contextualized against the **other** ally tokens before it is consulted as a Key. Residual + LayerNorm; `key_padding_mask` with slot-0 force-unmask, preventing NaN when a set is fully padded.

Ablation (C2): removing it **preserves** routing concentration (entry std 0.286) but **flips which pair is selected** — C2 routes Strike→Vanguard in 4/5 seeds, opposite to both C1 and the role-semantic prior.

⇒ cross-attention supplies **concentration**; intra-set blocks bias **selection**.

Cross-Attention Routing *exposes the convention*

"Given my role, whose state do I condition on?" The self-embedding queries each entity set. Because the role one-hot sits in **both** Q and K/V, the resulting weight is indexable by (r_i, r_j) — the 3×3 matrix falls out directly.

Slot-invariance: the policy is invariant to ally count by construction, so a 3v3 checkpoint runs unmodified at 9v9. Slot-ordering bias is ruled out by reversal.

Occlusion — the unified metric *apples-to-apples*

Attention weights and MLP gradients are **not** directly comparable. We therefore zero each ally token (mask bit held at 1) and measure policy KL from baseline — well-defined for **both** architectures.

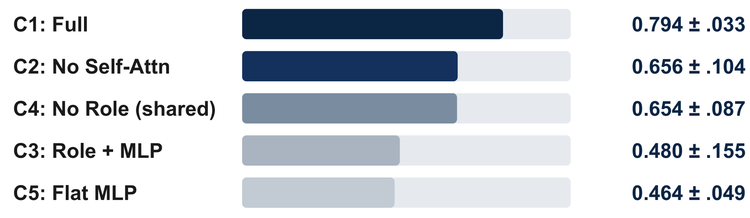
C1: occlusion ↔ attention at row-normalized Pearson $r = 0.976$.
C3 @ 9v9: erratic, one-to-two orders larger (max-row KL ≈ 5.6 vs. ≈ 0.08); the dominant slot varies by seed.

⇒ the MLP is slot-sensitive, but not **role**-structured.

5. EXPERIMENTS & RESULTS

(a) Track 1 — Win rate, SMACv2 3v3

mean ± std, 5 seeds



Attention helps: C1 − C3 = **+31.4 pp**. **Intra-set context helps:** C2 trails C1 by **13.8 pp**. **Labels alone do not:** C3 vs. C5 = 0.480 vs. 0.464, statistically indistinguishable — a role one-hot in an MLP does not by itself produce useful role-conditioned routing.

(b) Scaling, transfer & routing structure

SCALING 3V3 → 9V9

Scale	C1 AUC	C3 AUC
3v3	**.678 ± .019**	.328 ± .059
6v6	**.379 ± .064**	.311 ± .098
9v9	**.119 ± .067**	.089 ± .052

6v6 / 9v9 gaps fall inside seed std — **no significance claimed.** The **structural** signature persists where win-rate does not.

ZERO-SHOT TRANSFER (3V3-TRAINED)

Target	C1	C3	Gap
3v3	**.796**	.468	32.8 pp
6v6	**.212**	.084	12.8 pp
9v9	**.224**	.084	14.0 pp

C1 @ 9v9 zero-shot (**.224**) **beats from-scratch** 9v9 C1 (.110). C3 flatlines at ≈ .084 — MLP weights memorize a fixed slot pattern.

COND. 1 ROLE-ROUTING MATRIX — 5 SEEDS

agent \ ally	Strike	Vanguard	Support
Strike	—	0.473	0.527
Vanguard	0.565	—	0.435
Support	0.633	0.367	—

Strike → Support is Strike's largest entry (P4 ✓, 4/5 seeds). Concentrated, role-specific, stably off-diagonal.

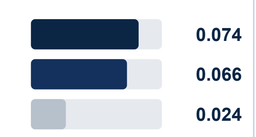
Vanguard **and** Support both route toward **Strike** — the two non-DPS roles attend to the engagement-driver.

CONCENTRATION & FORMATION SENSITIVITY

Routing entry std	value
C1 attention (SMACv2)	**0.246**
C2 no self-attn	0.286
C1 attention (MiniGrid)	0.252
C3 MLP gradient	**0.055**
C1 / C3 ratio	**4.5×**

Row-entropy and KL-from-uniform give the same ranking. MLP spatial/role gradient ratio: 5.8× / 2.3× / 3.7×.

$\Delta_{\max}$ **by role (Cond. 1)**



Support ≈ Vanguard (gap 0.008, below either std) → **P3 TIED**. Strike is clearly formation-**insensitive**, exactly as designed.

6. CONCLUSION & FUTURE WORK

- **An architecture-exposed routing channel makes the learned convention measurable.** Label-conditioned attention yields a 4.5× more concentrated, role-specific signature than MLP baselines.
- **Structure is more robust than performance.** Win-rate gaps dissolve into seed noise at 9v9; the routing signature, the zero-shot transfer advantage, and padding-invariance do not.
- **Seed-aware structural evaluation is mandatory.** Two of four earlier "strategic divergences" were small-n artifacts. A 3-seed routing claim is not a claim.

What this is not. Attention weights are **not** causal explanations; we report them as architecture-exposed patterns, corroborated by gradient and occlusion attribution. We do **not** claim the learned convention is an equilibrium, nor optimal under any solution concept. The contribution is a **measurement procedure.**

Future work. Pair the routing pathway with **learned** role discovery (e.g. RODE) for deployments without a designer schema; move beyond the three-role / Terran restriction; scale the diagnostic to agentic systems coordinating in open-ended environments, where strategic structure — rather than performance alone — may govern safety and interpretability.

7. References

[1] Rashid et al. **QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent RL.** ICML 2018.
[2] Yu et al. **The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games.** NeurIPS 2022.
[3] Ellis et al. **SMACv2: An Improved Benchmark for Cooperative Multi-Agent RL.** NeurIPS D&B 2023.
[4] Wang et al. **RODE: Learning Roles to Decompose Multi-Agent Tasks.** ICLR 2021.
[5] Jain & Wallace. **Attention is not Explanation.** NAACL 2019.

Acknowledgements

The author thanks **Prof. Jinseok Seo** (Dong-Eui University) and **Gyuyeol Jeong** (Netmarble) for insightful discussions and valuable comments that greatly improved this work.

yoosunghong.main@gmail.com · github: yoosung.dev