FROM MARL POLICIES TO SWARM LAWS: THE COMPRESSIBILITY SPECTRUM OF LEARNED COORDINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Swarm intelligence and multi-agent reinforcement learning (MARL) are two common paradigms for multi-agent coordination, yet they construct collective behavior from opposite directions. Swarm rules such as Boids and mean-shift consensus are local, interpretable, and independent of population size in their formulation. However, they must be designed by humans. MARL policies instead learn coordination from environmental rewards and can adapt to different tasks through training. However, the learned policies remain black boxes and may degrade when deployed beyond their training population. These complementary strengths motivate a more specific question. When can a trained collective policy be distilled into a compact local controller that preserves useful organization and transfers across population sizes? Our results show that the answer depends jointly on the policy, its observables, and the rule language. A causal task makes the information boundary explicit: when a hidden bit determines the correct target among locally indistinguishable options, blind mining falls to coin-flip performance, while exposing that bit restores the teacher's performance curve. In SMACv2, three unrelated compressors reveal a sharp Protoss–Terran contrast. On the compressible side, mined local laws preserve learned organization and transfer beyond the training population. In continuous pursuit–encirclement, a sixteen-number law exceeds its own teacher and maintains capture performance up to $32\times$ the training population at constant density. We further remove the hand-designed vocabulary and recover compact executable programs from a frozen grammar of 208 generic local terms. The recovered structure changes systematically across pursuit, coverage, and public MPE Predator–Prey tasks. Finally, direct return optimization and mining from different teachers separate rule-class capacity from teacher-specific content. Together, these results characterize when learned collective behavior can be compressed into a local swarm law, connecting MARL's automatic discovery of coordination with the interpretability and population scalability of swarm intelligence.

1 INTRODUCTION

Swarm intelligence and multi-agent reinforcement learning (MARL) are two common paradigms for multi-agent coordination, but they construct collective behavior from different starting points. Classical swarm systems begin with local interaction rules. Examples include separation–alignment–cohesion (Reynolds, 1987), Vicsek alignment (Vicsek et al., 1995), and mean-shift assembly (Sun et al., 2023). Repeated local interactions then generate group-level organization. These rules are compact and interpretable. Their local form does not grow with population size, which makes them naturally deployable across different population sizes. Their content, however, is supplied by a designer. MARL begins with reward (Sutton and Barto, 2018) and learns coordination from interaction (Rashid et al., 2018; Yu et al., 2022). This allows coordination strategies to be learned directly for different tasks. The resulting neural policies, however, remain black boxes, and their performance can degrade when population size moves beyond the training configuration.

The strengths of these two paradigms are therefore complementary. Swarm intelligence provides compact local structure that can be interpreted and deployed across population sizes, while MARL provides automatic discovery of coordination from reward. This motivates our central question: *when can coordination learned by MARL be converted into a compact local swarm law?* Such a conversion has three linked requirements. First, the local representation must contain the information

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

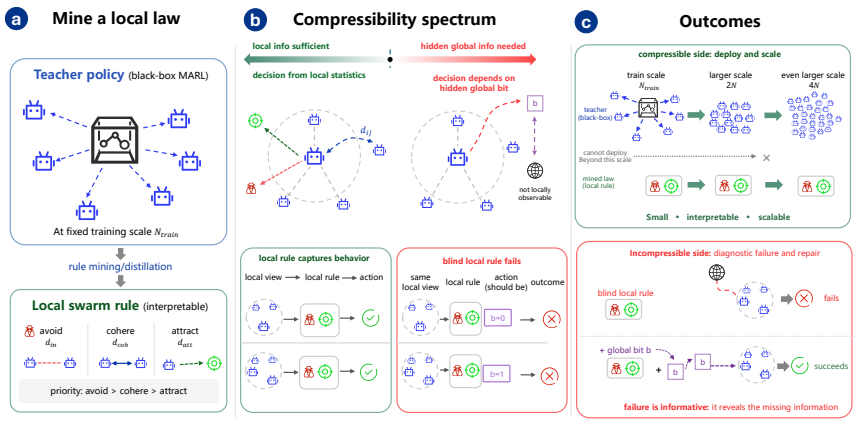

Figure 1: **From a learned policy to a local swarm law. (a)** A trained MARL policy is projected into a compact local rule at the training population. **(b)** Local information determines whether that projection can preserve the learned organization. **(c)** Compressible laws can be deployed at new population sizes, while restoring a missing observable tests an information-limited boundary.

that controls the teacher's decisions. Second, the rule language must be expressive enough to encode that information as a compact executable law. Third, the resulting law must preserve collective behavior in closed loop, including at population sizes beyond the training configuration. These conditions separate three sources of compression failure: missing local information, insufficient rule capacity, and loss of collective behavior under population transfer.

We formalize this idea as a *compressibility spectrum*. A trained policy is projected into a local rule language, and candidate rules are evaluated by collective return across population sizes. Compressibility is therefore relative to the policy, the locally available observables, the rule language, and the probing distribution used to read the policy. The spectrum includes both successful and resistant cases. A successful projection exposes a compact local realization of the learned coordination that can be deployed directly at new population sizes. A resistant projection localizes a bottleneck in the available information or the chosen representation. The spectrum is thus useful in both directions: successful compression produces a deployable controller, while resistance identifies what prevents the learned coordination from being expressed as a local law. Figure 1 summarizes this view.

The experiments follow this logic from boundary to transfer. SMACv2 first shows that the same compression procedures produce different outcomes on two trained policies. This establishes an empirical contrast in compressibility. A minimal hidden-bit task then isolates the information boundary causally. Two positive settings test the compressible side of the spectrum. A discrete decoy task exposes a local sufficient statistic that survives a frozen task family, and a continuous pursuit task yields a compact law that scales to much larger populations. We then examine the role of the rule language by replacing hand-specified behavior names with generic rule dictionaries and learned gates. Coverage and public MPE Predator–Prey test whether the same grammar reads out different organizations from different teachers. Finally, direct return optimization and teacher-content analyses separate the capacity of the rule class from the structure inherited from the trained policy, while organization-level measurements test what the extracted law preserves beyond one-step action agreement.

Contributions. We make three connected contributions. First, we formalize local policy compressibility through a complexity–performance spectrum governed by local information, rule expressivity, and cross-population collective behavior, and we give a causal local-information obstruction. Second, we measure both resistant and compressible regimes across SMACv2, a controlled hidden-bit task, a frozen decoy family, continuous pursuit, and public MPE Predator–Prey. The compressible laws remain executable across population sizes, and the pursuit law remains behaviorally stable up to $32\times$ the training population at constant density. Third, we show that the extracted content is recoverable from broad generic grammars and tracks the teacher's policy-specific decisions. Collective return and

group geometry provide stronger criteria for selecting useful rules when action fidelity alone is nearly unchanged.

2 Related Work

Swarm control laws. Simple local interactions can generate robust collective behavior (Bonabeau et al., 1999; Camazine et al., 2001). Representative examples include Boids (Reynolds, 1987), Vicsek alignment (Vicsek et al., 1995), collective-motion models (Couzin et al., 2002), consensus flocking (Olfati-Saber, 2006), mean-shift assembly (Sun et al., 2023), and kilorobot-scale self-assembly (Rubenstein et al., 2014). Swarm-engineering surveys identify rule design as a recurring bottleneck (Brambilla et al., 2013; Dorigo et al., 2021). Deep RL can learn swarm controllers directly (Hüttenrauch et al., 2019), while system identification can fit neural controllers from observed swarm trajectories (Jiahao et al., 2022). Our study connects these traditions by extracting local executable laws from trained multi-agent policies.

Policy distillation and interpretable control. Policy distillation compresses trained networks into smaller students (Hinton et al., 2015; Rusu et al., 2016), and interpretable RL adds readability as a design objective (Puiutta and Veith, 2020). VIPER (Bastani et al., 2018), differentiable trees (Silva et al., 2020), programmatic policies (Verma et al., 2018), symbolic regression (Landajuela et al., 2021), and free-form law discovery (Schmidt and Lipson, 2009; Brunton et al., 2016; Cranmer et al., 2020) provide compact policy classes. Controller synthesis methods can also optimize such classes directly for return (Shahnazari et al., 2025). Coordinated tree extraction extends this idea to multi-agent policies through MAVIPER (Milani et al., 2022) and HYDRAVIPER (Chen et al., 2025). These methods give us established projection mechanisms and expressive comparison classes. Our spectrum adds two deployment-level questions to that toolkit: whether the information needed by the collective policy is available to a local rule, and whether the compressed controller preserves collective return after population size changes. This distinction matters in multi-agent systems because high one-step agreement can coexist with different group geometry and return. We therefore use action fidelity to characterize the projection, but judge successful compression by closed-loop collective behavior.

Population-scalable MARL. Permutation-invariant, set-based, graph, and attention architectures can accept variable numbers of entities (Zaheer et al., 2017; Agarwal et al., 2020; Veličković et al., 2018; Lee et al., 2019; Sukhbaatar et al., 2016; Iqbal and Sha, 2019; Park et al., 2025). These architectures remove fixed input arity and therefore provide direct neural baselines for population transfer. Variable arity establishes that a controller can be evaluated at a new population. Behavioral stability at that population is a separate property. Our experiments compare this behavioral scale stability with compact local laws extracted from fixed-size teachers. The comparison separates architectural scalability from the distinct question studied here: whether the organization learned at one scale has a population-independent local realization.

Analysis of emergent coordination. Emergent MARL behavior is often studied through qualitative dynamics or cooperative benchmarks (Baker et al., 2020; Samvelyan et al., 2019; Ellis et al., 2023; Yu et al., 2022; Rashid et al., 2018; Lowe et al., 2017; Vinyals et al., 2019; Berner et al., 2019; Jaderberg et al., 2019). Role decomposition and population curricula explicitly build structure into training (Wang et al., 2021; Long et al., 2020). The Agent Response Map analyzes geometric mechanisms in trained swarm policies (Mi et al., 2026), while symbolic regression over GNN surrogates models observed dynamics (Powers et al., 2022). We operationalize learned organization through a local executable projection and evaluate that projection by collective behavior across population sizes.

3 Method

3.1 Local Policy Projection

Let π_θ be a weight-shared MARL policy trained with N_0 agents. Agent i receives observation o_i and takes action $a_i \sim \pi_\theta(\cdot \mid o_i)$. A local law belongs to a human-readable class

$$\mathcal{F}(\mathcal{O}, \mathcal{L}) = \{f_\phi : z_i = \mathcal{O}(o_i) \mapsto a_i\},$$

where $\mathcal{O}$ specifies locally available statistics and $\mathcal{L}$ specifies the rule language. We use decision trees, priorities and thresholds, and sparse combinations of vector-field primitives in different experiments. Given teacher decisions $\mathcal{D}_\pi = \{(o_i, a_i)\}$ on states drawn from a probing distribution μ, mining is the complexity-regularized projection

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{(o_i, a_i) \in \mathcal{D}_\pi} \ell(f(\mathcal{O}(o_i)), a_i) + \lambda\, \Omega(f).$$

The same form covers our least-squares laws and VIPER trees. It also makes the roles of observables, rule language, and program complexity explicit. The probing distribution μ determines which parts of the teacher are visible to the projection, while $\mathcal{O}$ determines which of those distinctions remain available to the local controller. The language $\mathcal{L}$ then determines which local relationships can be expressed compactly. We keep these objects explicit because a poor projection can arise from different sources, and the experiments manipulate them separately.

3.2 Compressibility Spectrum

A compact projection is useful when it preserves collective function. Let $J_N(f)$ be the expected task return when every agent executes f in a population of size N. For a scale set $\mathcal{N}$, we call the teacher (ε, K)-*locally compressible* relative to $(\mathcal{O}, \mathcal{L})$ if there exists $f \in \mathcal{F}$ with $\Omega(f) \le K$ such that

$$J_{N_0}(f) \ge J_{N_0}(\pi_\theta) - \varepsilon \quad \text{and} \quad J_N(f) \text{ remains above matched non-learned controls for all } N \in \mathcal{N}.$$

The first condition measures preservation at the training scale. The second measures behavioral transfer. A local syntax makes execution possible for new N, while the return criterion determines whether the learned organization survives. These are deliberately separate requirements. A rule may be compact and executable at every N yet fail behaviorally after local density or neighbor statistics shift; conversely, a high-fidelity surrogate at N_0 may encode an input interface that cannot even be evaluated at a new population. We summarize the trade-off with the complexity–performance frontier $K \mapsto \max_{\Omega(f) \le K} J_N(f)$. Compressibility is therefore a property of $(\pi_\theta, \mathcal{O}, \mathcal{L}; \mu)$ under a stated deployment family.

3.3 Local Information Boundary

Proposition 1 (Local-information obstruction)**.** *Suppose two equally likely episode states have identical local observations for every agent but require disjoint optimal actions. Any deterministic law $f(o_i^{\text{local}})$ has success probability at most $1/2$.*

Proof. The law receives the same input in both states and selects the same action. That action can be optimal in at most one of the two states. □

The proposition supplies a necessary boundary for local compression. The causal experiment in Section 4.2 realizes the obstruction and then adds the missing observable. The remaining experiments measure how much rule complexity is sufficient once the relevant information is locally available.

4 Results

4.1 SMACv2 Compressibility Contrast

We begin with frozen HARL/MAPPO checkpoints for SMACv2 5v5, using Protoss and Terran policies. Each checkpoint is compressed through three unrelated routes. The local-law route fits regime-dependent movement primitives and an attack gate from teacher decisions. The second route uses VIPER decision-tree distillation at matched budget. The third runs a MAVIPER-style coordinated extraction (Milani et al., 2022): one depth-limited tree per agent slot, trained with joint DAgger rollouts and teacher relabeling at the same data budget (Appendix A details the protocol).

The three routes produce the same qualitative contrast. On Protoss, the teacher scores 0.523, the local law 0.234, VIPER-DAgger 0.117, and VIPER-BC 0.070. On Terran, the teacher scores 0.578, the local law 0.070, and VIPER 0.062. The coordinated per-agent trees reach 0.086 on Protoss and

0.070 on Terran, so the closest multi-agent extraction recipe also remains far below both teachers while preserving the race ordering. Channel substitution gives the same ordering. Replacing all three decision channels changes episodic return by -0.022 on Protoss (paired 95% CI $[-0.907, 0.859]$) and by -3.444 on Terran (CI $[-4.769, -2.111]$). Within the Protoss law, the fitted attack gate and the 1-bit kiting memory carry measurable return: replacing the gate with greedy attack gives 0.148, and removing the memory gives 0.195. On Terran, the corresponding structure scores 0.102 versus 0.023 with the greedy gate. Agreement among the local law, the shared tree, and the coordinated per-agent trees is useful because these compressors make different representational commitments. Their shared race ordering points to a property of the learned policies under the tested observations, while the component ablations show that the recoverable Protoss behavior is carried by specific local decisions rather than by a uniform imitation gain.

Teacher-free return optimization over the same law parameters provides a capacity measurement. With $1{,}800$ search episodes, the best CEM laws reach 0.141 on Terran and 0.219 on Protoss, compared with teachers at 0.578 and 0.523. Under this search budget, the local class remains far below the teachers in both races. SMACv2 therefore establishes an empirical spectrum whose source can lie in missing local information, insufficient language capacity, or difficult projection. The controlled task in the next section varies observability while holding the local-rule question simple, allowing the information axis to be isolated causally.

4.2 Causal Information Boundary

We construct a minimal task in which two enemy subtypes have distinct local HP signatures. A hidden global bit g determines which subtype is mission-critical, and the horizon permits destroying one subtype. Local features describe every unit, while the correct choice depends on g.

A scripted conditional teacher wins 0.992. Mining with local observations alone gives 0.490, matching the coin-flip ceiling from Proposition 1. Adding g to the rule input restores $0.992/0.852/0.626$ at $N = 10/20/30$, alongside the teacher's $0.992/0.849/0.633$. The same pattern appears with a REINFORCE-trained teacher. Its training-scale score is 0.625; blind mining gives 0.505, while the one-bit rule reaches 0.831 ± 0.23 and retains the upper-bound curve at $N = 30$ as the neural teacher falls from 0.625 to 0.078.

This experiment identifies the information boundary directly. The teacher and task structure are unchanged, while the projection gains exactly the variable that distinguishes the two otherwise ambiguous episode states. Recovery after that intervention ties the failure to observability. It also clarifies the role of the remaining experiments: once the required information is locally available, compressibility is determined by whether a compact language can express the useful relation and whether that relation survives closed-loop deployment. The next two tasks examine this compressible side in discrete and continuous settings.

4.3 Discrete Local-Law Transfer

The decoy benchmark provides a discrete positive case. High-HP, zero-threat decoys stand between the agents and their mission targets. Nearest-target greedy wins 0.000 at every size and spends 82–86% of its damage on decoys. A lowest-HP heuristic reaches 0.691 at $N = 10$ and 0.000 at $N = 30$.

Label-free sufficient statistic. Policies are trained from team reward with local candidate features: distance, relative position, HP, sensed damage, and in-range status. No role label is supplied. All five seeds learn threat-first targeting, with decoy-targeting rate 0.022. Mining recovers the same priority order in every seed: threat $\rightarrow$ fragile $\rightarrow$ non-decoy $\rightarrow$ nearest. The observable-attribute rule reaches $0.996/0.992/0.953$ at $N = 10/20/30$, identical to the label-reading version. Ranking by sensed damage alone reproduces these outcomes, showing that the teacher has learned a compact local statistic.

Frozen task-family transfer. The mined rule and thresholds are frozen before evaluation on 300 procedurally varied task instances with randomized role statistics, layouts, sizes, horizons, speeds, and ranges. On the 100 held-out instances, the rule scores 0.833 mean win. The label-reading oracle reaches 0.843, and greedy reaches 0.345. Eighty percent of instances are solved at at least 75%

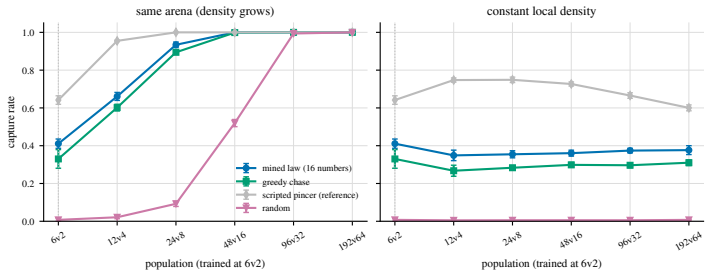

Figure 2: **Behavioral scaling in continuous pursuit.** The same six mined laws are deployed beyond training. Constant-density scaling isolates population growth from density growth; the extracted law remains approximately flat and above greedy through 192v64, or 32× the training population.

success. This test moves beyond reproducing the training instance. The rule must reuse the same local statistic when superficial role properties and geometry change, so performance on the held-out family measures whether the extracted relation persists across configurations.

Population-invariant neural baselines. The frozen MARL per-candidate scorer and a DeepSets policy provide variable-N neural comparisons under identical protocols. At $N = 30$, the mined rule scores 0.953, the teacher scorer 0.723, and DeepSets 0.328. On the held-out family the scores are $0.833/0.772/0.606$; on the worst decile they are $0.075/0.025/0.000$. The local threshold rule therefore preserves more of the teacher's task-family structure under scale and distribution shift. This comparison is important because all three controllers can be evaluated beyond the original population. It therefore measures behavioral stability directly as the entity set changes: the compact statistic remains useful, while the learned neural summaries shift more strongly. Additional noise, attrition, and tree-depth results appear in Appendix A.

4.4 Continuous Local-Law Scaling

We next test a continuous task with different dynamics. On a unit torus, P pursuers with speed 0.030 must capture E faster evaders with speed 0.040. Capture requires two pursuers simultaneously within radius 0.09, so successful behavior requires two-sided, wrap-around encirclement.

The teacher is a fixed-input MLP over sorted neighbor features, trained at 6v2 by greedy behavior cloning followed by REINFORCE. The local law is a regime-gated linear field over four unit primitives: chase, lead, flank, and separate,

$$f(z_i) = \sum_k w_{r(z_i),k}\, \hat{\mathbf{p}}_k(z_i), \qquad r(z_i) \in \{\text{near}, \text{far}\} \times \{\text{engaged}, \text{solo}\},$$

for sixteen fitted numbers. Baselines include greedy chase, random motion, a scripted two-sided pincer, and DeepSets and attention-GNN students trained on the same teacher behavior.

At the training size, the law captures 0.411 ± 0.025 versus the teacher's 0.339 ± 0.053 and greedy's 0.330 ± 0.049; both paired comparisons have $p = 0.031$. At 12v4 the law reaches 0.661 ± 0.021 versus greedy 0.602 ± 0.018, and at 24v8 it reaches 0.934 ± 0.015 versus 0.894 ± 0.009 ($p = 0.031$ at both scales). The law leads in all 18 seed×size comparisons. The attention GNN reaches $0.323/0.556/0.871$, and DeepSets reaches $0.164/0.367/0.520$. The task-informed scripted pincer remains the strongest reference at $0.642/0.955/1.000$, showing that additional coordination headroom remains. The law's improvement over its teacher is consistent with projection smoothing teacher-specific variability while preserving a coordination pattern that is useful in closed loop.

Constant-density scaling. Population growth also changes local density when the arena is fixed. That creates a confound: more agents produce more nearby encounters even if the controller itself has not become more scalable. The appendix shows that this effect eventually makes even random motion saturate in the fixed arena. We therefore scale capture radius, speeds, and the regime distance by $\sqrt{8/(P+E)}$ and deploy the same six mined laws without refitting. This keeps the characteristic

local interaction density comparable while population size changes. Capture remains nearly flat, from 0.349 ± 0.028 at 12v4 to 0.377 ± 0.023 at 192v64, and stays above greedy (0.27–0.31) at every size. This establishes behavioral transfer to $32\times$ the training population under constant density. Figure 2 shows the fixed-arena and constant-density curves. The distinction is central to the spectrum: locality guarantees that the rule can be executed at larger N, while constant-density evaluation tests whether the same local coordination law remains useful there.

Variable-N neural training. A matched attention architecture is also trained end-to-end on return over 6v2, 9v3, and 12v4 using the teacher's REINFORCE recipe and a $6{,}400$-episode budget. The selected model reaches 0.294 ± 0.030 at 6v2 and 0.060 ± 0.022 at 192v64, where the mined law reaches 0.377 ± 0.023. Appendix A reports the configuration sweep and full scaling details.

4.5 Generic Rule-Language Discovery

The continuous task allows us to test the second axis of the spectrum: rule-language expressivity. The sixteen-number controller begins with behaviorally meaningful primitives and hand-set regimes, which makes the resulting law easy to interpret but also supplies substantial prior structure. We therefore remove that structure in stages. First, generic geometric atoms replace named pursuit primitives. Next, learned gates replace the hand-set switching regimes. Finally, a much broader frozen grammar removes both choices at once. This sequence asks whether the same organization can still be recovered as less task knowledge is built into the language.

Primitive and gate discovery. Shrinking the named primitive set reduces training-scale capture from 0.411 ± 0.025 to the chase-only level 0.330. With a generic geometric dictionary, SINDy-style sequentially thresholded least squares (Brunton et al., 2016) reaches 0.405 ± 0.024 at moderate sparsity, using approximately 21 of 28 possible terms. The selected atoms recover chase, lead, and separation structure. Replacing the four hand-set regimes with an M5-style model tree also recovers the switching structure. In all 6 seeds, the root uses ally context, and selected target-distance thresholds fall at 0.24–0.27 around the hand value 0.25. The discovered program reaches 0.372 versus 0.380 for the hand-regime law, exceeds the ungated atoms at 0.349 in all seeds ($p = 0.031$), and reaches 0.381 at 192v64.

Frozen generic grammar. We next freeze a grammar of 208 generic local vector terms over two entity classes. Orthogonal matching pursuit builds a sparse program inside each learned gate. Four terms score 0.330, eight terms 0.378, twelve terms 0.410, and thirty-two terms 0.443. Programs selected from held-out action agreement score 0.399 ± 0.040 at the training size and 0.351 at 192v64 under constant density. Across seeds, the family-level content is stable: pursuit enters every leaf in $6/6$ seeds, velocity prediction in $5/6$, ally-context modulation in $5/6$, and a lateral component in $4/6$. Exact selected terms vary more than these families, which is expected when several parameterizations implement similar local interactions. The compression result is convergence on a small set of recurring mechanical ingredients that preserve collective behavior across equivalent parameterizations.

Task transfer and public benchmark. The same frozen grammar is applied to a coverage task. Every seed selects landmark attraction first; ally-avoidance enters $4/6$ programs; target-velocity terms are absent; and the root gate uses landmark distance with thresholds 0.06–0.09. Eight-term programs recover 96% of teacher coverage and remain stable to 48 agents at constant density. Because the candidate grammar is unchanged, the shift from pursuit-oriented terms to landmark attraction provides a direct content test: the vocabulary supplies a common expressive space, while the teacher and task determine which local organization is selected within it.

We then apply the grammar to public MPE Predator–Prey (`simple_tag`). PPO teachers achieve 4.15 ± 0.41 prey touches per episode per prey at 3v1, compared with 1.44 for greedy. The extracted law retains 73% of teacher performance at 3.02 ± 0.25 and executes unchanged at 6v2 and 12v4. Its margin over the density-dependent greedy anchor is $+1.58/ + 1.59/ + 1.67$ across the three sizes. A MAVIPER-style tree reaches 2.68 ± 0.20 at training size and has fixed input arity. DeepSets reaches 3.93 ± 0.25 at training size, then 1.33 at 6v2 and 1.69 at 12v4. The full program-selection outcomes, including the pre-specified program-size prediction, are reported in Appendix A. Figure 3 summarizes the two consequences that matter for the remainder of the analysis: the rule language

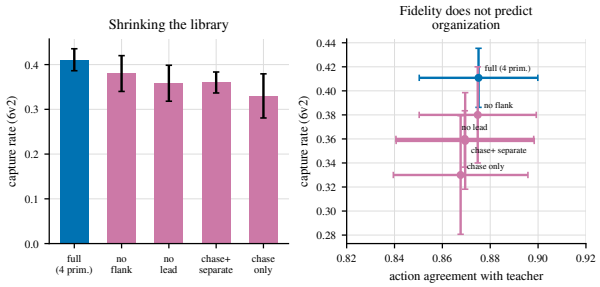

Figure 3: **Rule-language expressivity and collective organization.** Left: reducing the pursuit primitive library lowers closed-loop capture. Right: law variants retain similar 87–88% action agreement while collective capture differs by 24% relative.

constrains the collective behavior that can be expressed, while similar one-step fidelity can conceal substantial differences in collective performance.

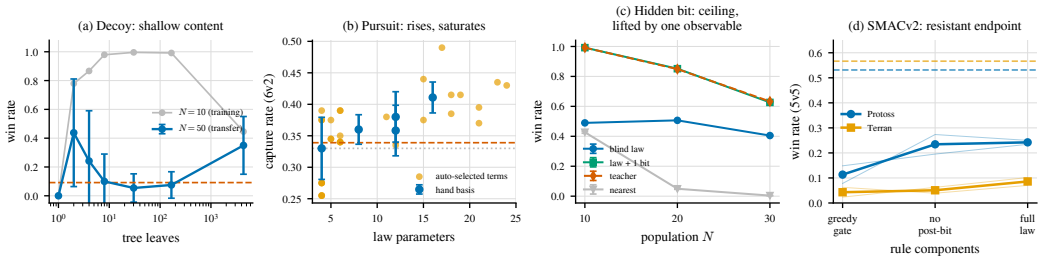

Figure 4: **Compressibility frontiers across tasks.** Decoy and pursuit reach compact high-performing regions, the hidden-bit ceiling is lifted by the missing observable, and SMACv2 provides a resistant endpoint. Dashed horizontal lines mark the corresponding teacher references (decoy teacher at $N{=}50$ in (a); pursuit teacher, with greedy dotted, in (b); Protoss and Terran teachers in (d)). Complexity is task-specific, so panels are compared by frontier shape.

4.6 Teacher Provenance and Collective Organization

The remaining analyses separate two questions that are easy to conflate after a strong local law has been found. The first is provenance: whether the extracted controller reflects the trained teacher or merely the capacity of the rule class. The second is evaluation: whether one-step action agreement is sufficient to identify the rule that best preserves collective organization.

Controller-class capacity and teacher content. Direct return optimization over the same sixteen pursuit parameters reaches capture 1.000 at every size in $6/6$ runs. The rule class therefore has substantial control capacity. Mining the same class from random, greedy, RL, and scripted-pincer teachers yields laws with capture 0.007, 0.376, 0.411, and 0.826, respectively. On a matched teacher-independent probe set, the ordering remains $0.007 \to 0.355 \to 0.382 \to 0.835$. A 2×2 state-distribution$\times$label-source analysis gives a label effect of $+0.052$ ($6/6$ seeds, $p = 0.031$). Direct optimization measures what the class can express when return is the objective, while mining identifies the controller supported by a particular teacher's decisions. The ordering and structural signatures track the teacher, establishing the provenance needed to interpret the mined law as a readout of learned coordination.

Collective organization metrics. Variants of the pursuit law all agree with the teacher on 87–88% of actions (fidelity 0.868–0.875), while collective performance ranges from 0.330 to 0.411, a 24% relative gap. Two geometric measurements separate the variants more clearly. Two-sided engagement rate correlates with return at $r = 0.63$ pooled, and pre-capture bearing spread at $r = 0.86$ pooled;

within-seed means are $0.84/0.83$, positive in all six seeds. Across teachers, action-fidelity correlation with return is $r = -0.51$. This gap is structurally plausible in a collective task: many locally common actions contribute little to coordination, while a smaller set of timing and positioning decisions can determine whether a useful group geometry forms. Collective return and group geometry therefore provide the informative criteria for selecting extracted rules when one-step agreement changes little.

5 DISCUSSION

Together, the results identify three conditions for local policy compressibility: sufficient local information, adequate rule expressivity, and stable collective behavior under population transfer. Observables determine what a local controller can know, the rule language determines what it can express, and closed-loop return determines whether the resulting organization remains useful. The hidden-bit intervention isolates the information boundary, while the generic-grammar and transfer experiments locate the other two dimensions.

Population transfer is stronger than variable-size execution. A local rule can run at new population sizes because its input does not grow with the number of agents, but the learned relation must also remain useful as population statistics change. Constant-density pursuit separates these properties and remains stable from 12v4 to 192v64. The language and provenance results clarify what is transferred: broad grammars provide an expressive space, while different teachers select different structures within it. Collective organization must therefore be judged in closed loop. Pursuit-law variants differ by less than one percentage point in action fidelity while differing by 24% in collective return, and group geometry tracks return more closely.

This view connects the complementary strengths of MARL and swarm intelligence. MARL discovers coordination from reward, while compressibility tests whether that coordination admits a compact local realization. Successful projection yields an interpretable controller that can transfer beyond the training population. Resistant cases localize limitations in information, expressivity, or transfer. The resulting workflow is *learn–diagnose–distill*: learn coordination, diagnose its local requirements, and distill the organization that remains useful across scale.

6 LIMITATIONS

Compressibility is measured relative to the chosen observables, rule language, probing distribution, and deployment family. Terran resistance therefore applies to the tested representations and search budgets. The variable-N comparison covers one attention-based architecture and training recipe, and the largest scale test reaches $32\times$ the training population at constant density. Broader rule languages, richer observables, additional population-scalable teachers, and new coordination domains are needed to map the spectrum more completely.

7 CONCLUSION

This work asks when coordination learned by MARL can be converted into a compact local swarm law. The results reveal a measurable compressibility spectrum governed by local information, rule expressivity, and cross-scale collective behavior. When these conditions align, compact laws preserve collective return and remain effective far beyond the training population. When decisive information is absent, compression reaches a causal ceiling that is lifted by restoring the missing observable. Broad generic grammars further recover policy-specific local structure, while collective return and group geometry distinguish useful laws more clearly than action fidelity alone.

The broader significance is a practical bridge between automatic coordination discovery and swarm-law execution. MARL can discover coordination from reward, and local laws can provide readable, compact, population-scalable controllers when that organization admits a local representation. The same spectrum also makes failure informative by identifying whether the limiting factor is information, expressivity, or transfer.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, provide feedback on experimental design, support result interpretation, assist with code implementation and debugging, and support manuscript drafting, translation, revision, and grammar editing. The research hypotheses and core conceptual framing were developed and refined by the authors. All AI-assisted outputs incorporated into the work were reviewed and verified by the authors, who take responsibility for the final content.

REFERENCES

A. Agarwal, S. Kumar, K. Sycara, and M. Lewis. Learning transferable cooperative behavior in multi-agent teams. In *AAMAS*, 2020.

K. Shahnazari, S. M. Ayyoubzadeh, and M. Keshtparvar. BASIL: Best-action symbolic interpretable learning for evolving compact RL policies. *arXiv:2506.00328*, 2025.

B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch. Emergent tool use from multi-agent autocurricula. In *ICLR*, 2020.

O. Bastani, Y. Pu, and A. Solar-Lezama. Verifiable reinforcement learning via policy extraction. In *NeurIPS*, 2018.

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, et al. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261*, 2018.

C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, et al. Dota 2 with large scale deep reinforcement learning. *arXiv:1912.06680*, 2019.

E. Bonabeau, M. Dorigo, and G. Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, 1999.

M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo. Swarm robotics: a review from the swarm engineering perspective. *Swarm Intelligence*, 7(1):1–41, 2013.

S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS*, 113(15):3932–3937, 2016.

S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. *Self-Organization in Biological Systems*. Princeton University Press, 2001.

R. Chen, S. Milani, Z. Zhang, N. Sadeh, and F. Fang. Making teams and influencing agents: Efficiently coordinating decision trees for interpretable multi-agent reinforcement learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 567–578, 2025.

I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks. Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology*, 218(1):1–11, 2002.

M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho. Discovering symbolic models from deep learning with inductive biases. In *NeurIPS*, 2020.

M. Dorigo, G. Theraulaz, and V. Trianni. Swarm robotics: Past, present, and future [Point of View]. *Proceedings of the IEEE*, 109(7):1152–1165, 2021.

B. Ellis, J. Cook, S. Moalla, M. Samvelyan, M. Sun, A. Mahajan, J. Foerster, and S. Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *NeurIPS Datasets and Benchmarks*, 2023.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

M. Hüttenrauch, A. Šošić, and G. Neumann. Deep reinforcement learning for swarm systems. *Journal of Machine Learning Research*, 20(54):1–31, 2019.

S. Iqbal and F. Sha. Actor-attention-critic for multi-agent reinforcement learning. In *ICML*, 2019.

M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

T. Z. Jiahao, L. Pan, and M. A. Hsieh. Learning to swarm with knowledge-based neural ordinary differential equations. In *ICRA*, 2022.

M. Landajuela, B. K. Petersen, S. Kim, C. P. Santiago, R. Glatt, N. Mundhenk, J. F. Pettit, and D. Faissol. Discovering symbolic policies with deep reinforcement learning. In *ICML*, 2021.

J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019.

Q. Long, Z. Zhou, A. Gupta, F. Fang, Y. Wu, and X. Wang. Evolutionary population curriculum for scaling multi-agent reinforcement learning. In *ICLR*, 2020.

R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*, 2017.

S. Milani, Z. Zhang, N. Topin, Z. R. Shi, C. A. Kamhoua, E. E. Papalexakis, and F. Fang. MAVIPER: Learning decision tree policies for interpretable multi-agent reinforcement learning. In *ECML PKDD*, pages 251–266, 2022.

Y. Mi, J. Li, L. Li, and S. Zhao. Unveiling complex collective behaviors from simple rewards. *arXiv:2607.12861*, 2026.

R. Olfati-Saber. Flocking for multi-agent dynamic systems: Algorithms and theory. *IEEE Transactions on Automatic Control*, 51(3):401–420, 2006.

H. Park, B. Seong, and S.-K. Ko. SPECTra: Scalable multi-agent reinforcement learning with permutation-free networks. *arXiv:2503.11726*, 2025.

S. Powers, J. Smith, and C. Pinciroli. Extracting symbolic models of collective behaviors with graph neural networks and macro-micro evolution. In *Int. Conf. on Swarm Intelligence (ANTS)*, 2022.

E. Puiutta and E. M. S. P. Veith. Explainable reinforcement learning: A survey. In *Machine Learning and Knowledge Extraction (CD-MAKE)*, LNCS 12279, pages 77–95, 2020.

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

C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH*, 1987.

S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

M. Rubenstein, A. Cornejo, and R. Nagpal. Programmable self-assembly in a thousand-robot swarm. *Science*, 345(6198):795–799, 2014.

A. A. Rusu, S. G. Colmenarejo, C. Gulcehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu, and R. Hadsell. Policy distillation. In *ICLR*, 2016.

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

M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

A. Silva, T. Killian, I. Jimenez, S.-H. Son, and M. Gombolay. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *AISTATS*, 2020.

S. Sukhbaatar, A. Szlam, and R. Fergus. Learning multiagent communication with backpropagation. In *NeurIPS*, 2016.

G. Sun, R. Zhou, Z. Ma, Y. Li, R. Groß, Z. Chen, and S. Zhao. Mean-shift exploration in shape assembly of robot swarms. *Nature Communications*, 14:3476, 2023.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, second edition, 2018.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *ICLR*, 2018.

A. Verma, V. Murali, R. Singh, P. Kohli, and S. Chaudhuri. Programmatically interpretable reinforcement learning. In *ICML*, 2018.

T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226–1229, 1995.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019.

T. Wang, T. Gupta, A. Mahajan, B. Peng, S. Whiteson, and C. Zhang. RODE: Learning roles to decompose multi-agent tasks. In *ICLR*, 2021.

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

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. Deep sets. In *NeurIPS*, 2017.

A Supporting Experiments and Diagnostics

The supplementary experiments follow the same logic as the main text. Section A.1 expands the resistant SMACv2 endpoint, Section A.2 extends the discrete transfer and rule-complexity analysis, and Section A.3 details the scaling protocol for continuous pursuit. Section A.4 examines rule-language complexity and public-benchmark transfer. Section A.5 then extends the provenance and organization analyses used to determine what the compressed controller preserves.

A.1 SMACv2 Capacity and Deployment Diagnostics

The main text uses SMACv2 to establish a resistant endpoint of the measured spectrum. The following analyses separate the capacity available to the tested rule class from the fact that the extracted law remains syntactically executable beyond the teacher's fixed input size.

Capacity search. We optimize the SMACv2 law parameters directly against environment return with CEM. Both from-scratch and warm-started runs use $1{,}800$ search episodes, and selected controllers are re-evaluated on 128 episodes. The best Terran law reaches 0.141 (teacher 0.578, mined law 0.070), and the best Protoss law reaches 0.219 (teacher 0.523, mined law 0.234). The finite search budget provides an empirical lower bound on the capacity available to the tested law family.

Raw scale transfer. At 10v10 and 20v20, the mined SMACv2 law remains above random and below greedy in win rate; at 10v10 it reaches 0.062 versus greedy 0.094. Its focus-fire signature is 0.712 versus greedy 0.680. Fixed-input teachers and trees retain their 5v5 input arity, whereas the local law remains executable at the larger populations.

MAVIPER-style protocol. The coordinated route fits one depth-16 CART classifier per agent slot (five trees) rather than a single shared tree over pooled samples. After a behavior-cloning round of 100 episodes, each of two DAgger rounds rolls out all five current trees jointly for 25 episodes, so every tree's training distribution reflects the other trees' closed-loop behavior, and visited states are relabeled by the teacher. The data budget (150 collection episodes), depth, leaf constraint, and 128 evaluation episodes match the shared-tree VIPER baseline. The original MAVIPER weights DAgger resampling by a criticality criterion computed from centralized per-action critics; the MAPPO teacher

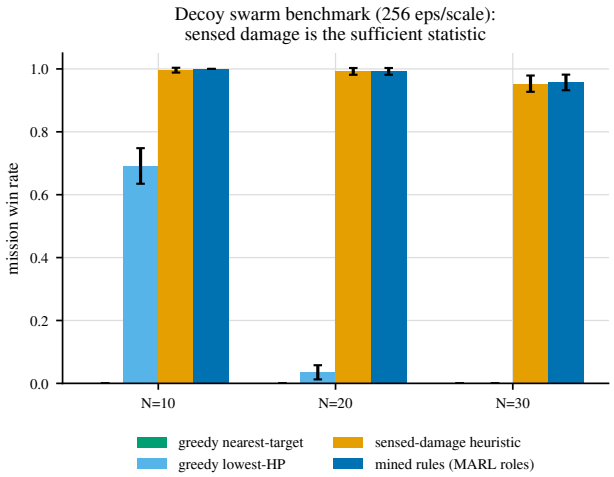

Figure 5: **Decoy-task population transfer.** Win rate across population sizes for nearest-target greedy and the rule mined from label-free MARL policies (256 episodes per scale). The mined local rule remains near the task ceiling across the evaluated sizes.

exposes a state-value head only, so resampling is uniform, a close VIPER variant. Final trees have 183–296 leaves per slot. Win rates progress from 0.039 (BC) to 0.086 on Protoss and from 0.016 to 0.070 on Terran; coordination-aware trees improve on BC but remain far below both teachers.

A.2 Decoy Transfer and Rule Complexity

The decoy task provides the discrete compressible case. Here we extend the population-transfer result and test how much rule complexity is needed to preserve it.

A complexity analysis shows that the transferable content is shallow. A depth-1 tree transfers best to $N = 50$ at 0.438, while the full teacher and a depth-5 tree reach 0.092 and 0.054 at that scale.

A.3 Continuous Scaling Protocols

The continuous pursuit result depends on distinguishing population growth from density growth. This section reports the variable-N neural baseline and the fixed-arena versus constant-density protocols used for that separation.

Variable-N neural training. The attention baseline is trained end-to-end on return over 6v2, 9v3, and 12v4 using greedy-BC initialization, the teacher's REINFORCE recipe, and a 6,400-episode budget. A 3×2 learning-rate$\times$width sweep selects the configuration on training-distribution validation; the published recipe is within 8% of the best configuration, while large learning rates diverge. The selected model reaches 0.294 ± 0.030 at 6v2 and decreases to 0.060 ± 0.022 at 192v64, where the mined law reaches 0.377 ± 0.023.

Scaling protocol. In a fixed arena, random motion already reaches 0.52 at 48v16, and all controllers saturate at 1.0 by 96v32. The constant-density test scales capture radius, speeds, and the regime distance by $\sqrt{8/(P+E)}$. Under this protocol the local law changes from 0.349 ± 0.028 at 12v4 to 0.377 ± 0.023 at 192v64, while greedy remains in the 0.27–0.31 range.

A.4 Generic Grammar and Cross-Task Transfer

The main text progressively removes hand-specified primitives and regimes. The analyses below quantify the resulting program sizes and show how the same frozen grammar changes its selected structure across pursuit, coverage, and public MPE Predator–Prey.

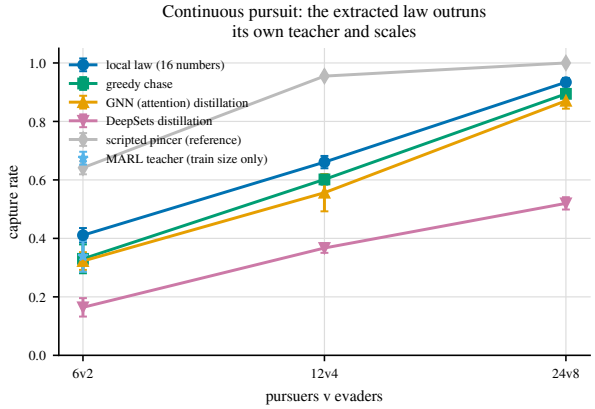

Figure 6: **Training-scale and near-scale pursuit comparison.** Capture rate for the sixteen-number law, its MARL teacher, greedy pursuit, size-invariant neural students, and the scripted pincer reference at 6v2, 12v4, and 24v8 (6 seeds, 100 episodes per point). This comparison complements the constant-density scaling curve in the main text by showing the teacher and neural baselines at the smaller evaluated populations.

Generic program size. In the 208-term grammar, held-out agreement with a pre-specified one-percent tolerance selects programs of 4–12 terms. Sweeping the tolerance over 0–2% moves the median program from 32 to 8 terms while closed-loop capture remains within 0.38–0.44. A BIC stop selects 34–64 terms. The exact selected modulators vary across seeds (mean pairwise Jaccard 0.211), while family-level ingredients remain stable.

Coverage-task details. Pursuit and coverage produce different program families from the same 208-term grammar. In coverage, landmark attraction is the first selected term in every leaf for $6/6$ seeds, ally-avoidance enters $4/6$ programs, and target-velocity terms enter $0/6$. All root gates use nearest-landmark distance, with thresholds 0.06–0.09, near the 0.09 coverage radius. Program-family overlap with pursuit is 0.09–0.23 Jaccard. Eight-term programs recover 96% of teacher coverage and remain stable to 48 agents at constant density.

MPE program structure. On public MPE Predator–Prey, the first term in every leaf is distance-weighted prey attraction in $6/6$ seeds. Prey-velocity prediction appears in $6/6$ seeds and is the second selected term in every seed; teammate-context terms enter $5/6$. All gate trees split on prey speed, with thresholds near 0.38 and 0.60. Refitting the selected terms without gates increases held-out disagreement in $6/6$ seeds (agreement 0.794 versus 0.762), yet does not reduce closed-loop performance: the ungated refit scores 3.24 versus 3.02 at 3v1 (higher in $6/6$ seeds) and 5.24 versus 5.09 at 12v4 ($5/6$). The discovered prey-speed gate is therefore fidelity-relevant but not return-critical in this task, in contrast to pursuit, where gating improves return in all seeds. The selected programs contain 8–24 terms (median 20); the observed maximum of 24 exceeds the pre-specified 4–20 program-size band.

Against the same teacher rollouts, a MAVIPER-style coordinated tree reaches 2.68 ± 0.20 at the training size. A DeepSets clone reaches 3.93 ± 0.25 at 3v1, then 1.33 at 6v2 and 1.69 at 12v4. The greedy anchor is $1.44/2.11/3.42$ across those sizes. The local law maintains margins $+1.58/+1.59/+1.67$, while the DeepSets margin changes from $+2.49$ to -0.78 and -1.73. Sweeping the selection tolerance over 0–2% changes the median MPE program from 32 to 10 terms while performance remains within 2.92–3.15.

A.5 Teacher Provenance and Organization Metrics

The final supplementary analyses address two related questions: how strongly the mined controller follows teacher-specific decisions, and which measurements best track the collective organization preserved by the projection.

Matched-state decomposition. The teacher-content ordering persists under two alternative probe mixtures in all six seeds. RL-visited probe states increase the RL–greedy gap to $+0.076$ ($5/6$, $p = 0.062$). In the 2×2 state-distribution$\times$label-source factorization, the label effect is $+0.052$ ($6/6$, $p = 0.031$), the state main effect is -0.006, and the interaction is $+0.059$ ($5/6$, $p = 0.094$). On the common support of the two state distributions, covering approximately 90% of records, the label effect is $+0.039$ ($6/6$, $p = 0.031$), and the interaction is $+0.047$ ($4/6$, $p = 0.125$).

Organization anchors. The two-sided engagement rate is the per-step frequency with which at least two pursuers hold an alive evader within 0.2 at bearings separated by at least $90°$. In this organization-metric analysis, its values are 0.033 for greedy, 0.038–0.042 for the five law variants, and 0.175 for the scripted pincer, with corresponding returns 0.255, 0.330–0.411, and 0.650. The pooled correlation with return is $r = 0.63$, and the within-seed mean is $r = 0.84$. Pre-capture bearing spread is $78°$ for greedy, 82–$85°$ for the variants, and $93°$ for the pincer. Its pooled correlation with return is $r = 0.86$; the within-seed mean is $r = 0.83$, with seed-bootstrap 95% CI $[0.67, 0.96]$. Within one teacher, fidelity still ranks variants with within-seed mean $r = 0.71$; across teachers, the pooled correlation changes to $r = -0.51$.

Exploratory localization analysis. A six-seed intervention replaced teacher-disagreeing actions at selected locations and increased return from 0.330 to 0.386; a count-matched random replacement gave $p \approx 0.27$. A pre-specified replication on twelve fresh seeds produced $+0.014$, 95% CI $[-0.009, +0.039]$, $p = 0.31$, leaving the location-specific effect unresolved. At the regime level, swapping the closure regime produces the largest observed regime cost, -0.022. We therefore report the localization result as exploratory.

B Reproducibility

Every experiment writes JSON outputs; every number in this paper is copied verbatim from those files and re-validated by `paper/verify_numbers.py`. Within each comparison, all controllers share identical environment classes, parameters, and evaluation seeds. An end-to-end regeneration check re-ran the full extraction–selection–evaluation pipeline for one representative seed per task (pursuit and MPE) from the stored teacher checkpoints. All fitting-stage quantities, selected term sets, gate thresholds, and closed-loop returns reproduced exactly ($24/24$ compared quantities; script and report included in the release).