LOCAL-TO-COLLECTIVE PREDICTABILITY FROM HETEROGENEOUS MULTI-AGENT INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Collective behavior in multi-agent systems emerges from repeated local interactions, yet existing approaches lack an explicit representation of how heterogeneous local interaction changes propagate to collective-level degradation. To address this gap, this work introduces *Interaction Predictability* (IP), an agent-level representation of how strongly an agent's spatiotemporal interaction dynamics support prediction of its future motion. Building on this representation, **Local-to-Collective Predictability (L2CP)** learns IP from ego-centric interaction histories and selectively aggregates agent-level predictability through learned relevance into a *Collective Predictability Score* (CPS). Both representations are learned without using degradation labels or collective-state metrics as training targets. Experiments show that IP distinguishes heterogeneous local interaction conditions and achieves stronger degraded-agent localization than simple local-statistics baselines, while relevance-weighted CPS captures collective degradation structure consistently across multiple interaction dynamics. These results establish local interaction predictability as a learned intermediate representation that connects distributed agent-level dynamics to collective degradation.

1 INTRODUCTION

Modern multi-agent systems, including autonomous swarms and distributed robotic teams, achieve coordinated behavior through repeated local interactions among agents (Zheng et al., 2024; Li et al., 2014). As the number of agents and interactions grows, coordination depends on increasingly diverse local interaction contexts distributed across the collective. Global properties such as connectivity, coordination, and group organization emerge from the evolving structure of these interactions (Ren & Beard, 2005). When local interactions are disrupted by sensing noise, motion perturbations, limited observability, or other disturbances, their effects can propagate through the interaction network and alter the collective state, leading to fragmentation, connectivity loss, or coordination loss (Santilli et al., 2022; Zheng et al., 2025; Rezaee et al., 2020). Figure 1(a) illustrates this process, where persistent local interaction degradation leads to collective consequences that continue to evolve over time. The challenge is therefore not only to detect local interaction changes, but to determine which changes matter at the collective level. Because interaction conditions vary across agents and over time, similar disturbances can have different collective effects depending on local context (Zheng et al., 2021; Wen et al., 2024). This motivates the central question: *How can heterogeneous local interaction dynamics be represented and connected to collective-level degradation?*

Existing approaches provide complementary but incomplete views of this problem. Collective-state monitoring characterizes global properties such as connectivity, cohesion, or order, but typically relies on system-level state information and describes the resulting collective state rather than the distributed interaction dynamics from which it emerges (Rahimi et al., 2025; Cui et al., 2024; He et al., 2025; Cavagna et al., 2022). Fault- and anomaly-detection methods identify deviations from nominal agent behavior, but typically focus on whether individual agents are abnormal rather than on how heterogeneous interaction changes relate to collective consequences (Arrichiello et al., 2014; Tarapore et al., 2019; Carminati et al., 2024; O'Keeffe, 2025). Interaction-aware trajectory-prediction models explicitly represent inter-agent dependencies, but use these interactions primarily to improve or characterize future-motion prediction (Wen et al., 2024; Xiang et al., 2024; Thakkar et al., 2025; Tang et al., 2021). What remains missing is an intermediate representation that connects distributed local interaction dynamics to the evolving collective state.

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.

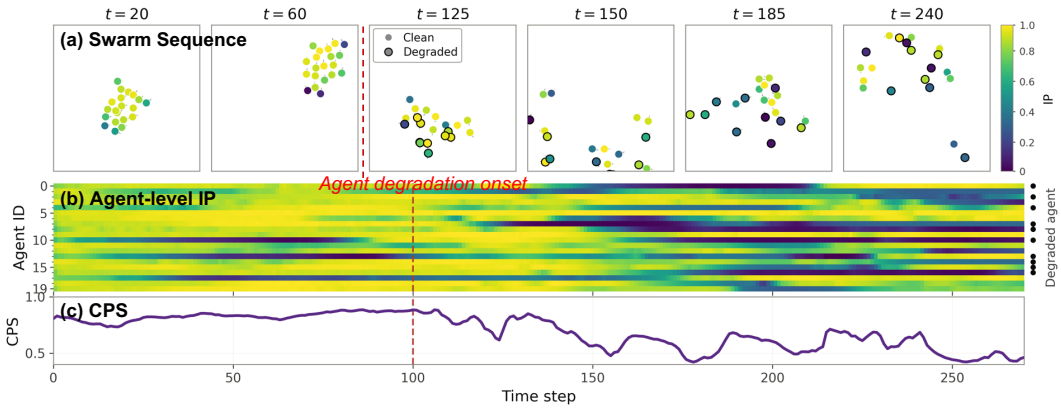


Figure 1: **Local-to-collective predictability under partial interaction degradation.** (a) A representative illustrative swarm sequence showing the evolution of collective structure under partial interaction degradation. (b) The agent-level Interaction Predictability (IP) map captures heterogeneous changes in local interaction predictability across individual agents following degradation onset. (c) The Collective Predictability Score (CPS) aggregates these distributed agent-level signals into a collective readout associated with the evolving collective state.

To address this gap, this paper introduces *Interaction Predictability*, an agent-level representation of how strongly an agent's observed spatiotemporal interactions support prediction of its future motion. Rather than characterizing local interaction condition with a single hand-crafted neighborhood statistic, Interaction Predictability is inferred from the temporal evolution of ego-centric relative configurations and motion. Reliable and temporally consistent interactions provide stronger predictive structure, whereas disrupted or inconsistent interactions provide weaker support for subsequent motion. Interaction Predictability is related to, but not equivalent to, trajectory uncertainty. Trajectory uncertainty characterizes ambiguity in the predicted future trajectory, whereas Interaction Predictability is inferred from the observed interaction history and reflects how strongly that history supports future-motion prediction. Because collective behavior and its degradation emerge from distributed local interactions, this predictive structure provides a link between heterogeneous agent-level dynamics and the evolving collective state.

Building on this representation, the **Local-to-Collective Predictability (L2CP)** framework learns agent-level interaction predictability (IP) from local spatiotemporal interaction histories and selectively aggregates it into a *Collective Predictability Score* (CPS). L2CP operates on ego-centric local observations without reconstructing the full global system state. Because agents occupy heterogeneous local contexts, their predictability signals need not contribute equally at the collective level; L2CP therefore learns relative agent relevance to weight each agent's contribution to CPS. Both IP and CPS are learned through future-motion-derived supervision, without using degradation labels or collective-state metrics as training targets. As illustrated in Figure 1(b–c), the resulting representation connects heterogeneous local interaction dynamics to collective degradation. Experiments validate this local-to-collective representation at both agent and system levels. Agent-level IP distinguishes degraded and nominal interaction conditions even when simple instantaneous neighborhood statistics are similar, demonstrating that it captures local interaction condition beyond individual instantaneous descriptors. Learned relevance identifies agents whose predictability signals contribute more strongly to the collective readout, while relevance-weighted CPS captures collective degradation structure consistently across multiple interaction dynamics.

Our main contributions are:

- We introduce **Interaction Predictability** (IP), an agent-level representation of how heterogeneous local spatiotemporal interaction dynamics support future-motion prediction, providing an interaction-history-based representation complementary to conventional trajectory uncertainty.
- We introduce **L2CP**, a local-to-collective representation framework that preserves agent-level interaction heterogeneity and learns how individual predictability signals contribute to a collective readout through relative relevance.
- We show that IP distinguishes heterogeneous local interaction conditions and achieves the highest mean degraded-agent localization AUROC among the evaluated local-statistics baselines, while learned relevance reveals unequal agent contributions and the resulting CPS consistently reflects collective degradation across multiple interaction dynamics.

2 Related Work

Prior work related to L2CP spans three complementary directions: collective-state characterization, fault and anomaly detection, and interaction-aware trajectory prediction. These directions respectively emphasize system-level state, agent-level abnormalities, and interaction-conditioned future motion. What remains missing is an explicit local-to-collective representation that preserves heterogeneous interaction dynamics at the agent level while relating them to collective-level change. L2CP addresses this gap through agent-level predictability modeling, learned relevance, and collective aggregation. Table 1 summarizes this distinction.

Table 1: Comparison of related research directions. △ denotes interaction modeling used internally without an explicit representation of local interaction condition.

Approach	Interaction Condition	Agent Variation	Collective Readout	Agent Relevance
Collective-state characterization	△	✗	✓	✗
Fault & anomaly detection	✓	✓	✗	✗
Trajectory prediction	△	✓	✗	✗
L2CP	✓	✓	✓	✓

2.1 Collective Behavior and State Characterization

Collective behavior is commonly modeled as an emergent consequence of local interactions. Classical models such as Boids (Reynolds, 1987) and the Vicsek model (Vicsek et al., 1995) demonstrate how simple neighborhood-level rules produce flocking, alignment, and group structure. Recent learning-based methods further infer pairwise, higher-order, or causal interaction structure from multi-agent observations (Uwamichi et al., 2024; Lee & Mukhopadhyay, 2026; Rahimi et al., 2025). At the system level, collective behavior is characterized through aggregate state distributions, population-level dynamics, or global properties such as connectivity, cohesion, polarization, and fragmentation (Cui et al., 2024; He et al., 2025; Vicsek et al., 1995; Cavagna et al., 2022). These approaches describe interaction structure or collective state, but do not explicitly represent which heterogeneous local interaction conditions are most strongly associated with collective degradation.

2.2 Fault and Anomaly Detection in Multi-Agent Systems

Fault and anomaly detection methods focus on identifying abnormal agents from deviations in local state or behavior. Prior work includes distributed diagnosis using local and inter-agent information (Arrichiello et al., 2014), behavioral outlier detection in physical swarms (Tarapore et al., 2019), data-driven detection of faulty swarm members (Carminati et al., 2024), and predictive approaches to gradual failures (O'Keeffe, 2025). These methods capture agent-level variation in abnormality, but primarily determine whether an agent is faulty or likely to fail rather than linking heterogeneous local changes to collective degradation.

2.3 Interaction-Aware Trajectory Prediction and Uncertainty

Multi-agent trajectory prediction models increasingly capture inter-agent dependencies and multi-modal future uncertainty. Trajectron++ (Salzmann et al., 2020) models dynamic relational structure, AgentFormer (Yuan et al., 2021) captures temporal and social dependencies, and recent methods incorporate adaptive graphs, interaction-conditioned latent variables, and higher-order relations (Wen et al., 2024; Xiang et al., 2024; Thakkar et al., 2025). Interaction-induced uncertainty has also been explicitly modeled in probabilistic forecasting (Tang et al., 2021). These approaches use interactions primarily to improve or characterize future-motion prediction. L2CP instead treats local interaction condition as an explicit learned representation and aggregates agent-level predictability using learned relevance to collective dynamics.

3 Problem Formulation

Consider a partially observed multi-agent system with N interacting agents. At time t, the global state is $\mathbf{X}^t = (x_1^t, \ldots, x_N^t)$, where $x_n^t \in \mathbb{R}^d$ denotes the state of agent n, such as its position and velocity. Agent n observes only ego-centric relative states within its local neighborhood $\mathcal{N}_n^t$,

$$\tilde{x}_n^t = \left\{ \phi(x_n^t, x_m^t) \;\middle|\; m \in \mathcal{N}_n^t \right\},$$

where $\phi(\cdot,\cdot)$ denotes a relative observation function. The temporal sequence of these observations defines the agent's spatiotemporal interaction history, denoted by $\tilde{\mathbf{X}}_n^{\leq t}$. The model receives only the collection of local interaction histories, $\tilde{\mathbf{X}}^{\leq t} = \{\tilde{\mathbf{X}}_n^{\leq t}\}_{n=1}^N$. Global absolute coordinates and the full interaction graph are not assumed to be available during inference.

The objective is to learn an agent-level representation $r_n^t = F_\theta(\tilde{\mathbf{X}}_n^{\leq t})$ that preserves heterogeneous differences in local interaction dynamics. The resulting set of agent representations, $\mathcal{R}^t = \{r_n^t\}_{n=1}^N$, is then aggregated into a collective representation $c^t = G_\theta(\mathcal{R}^t)$ which summarizes the interaction structure distributed across agents at the system level.

Both representations are learned through future-motion prediction. Given interaction histories up to time t, the model predicts future trajectories $\mathbf{Y}^{t+1:t+T}$ over a horizon T,

$$p_\theta\left(\mathbf{Y}^{t+1:t+T} \mid \tilde{\mathbf{X}}^{\leq t}\right).$$

Collective degradation is introduced only for evaluation and is computed from future global states as $D^t = \Psi(\mathbf{X}^{t+1:t+T})$. The central question is whether predictive supervision can induce agent-level representations $\mathcal{R}^t$ that reflect heterogeneous interaction conditions and a collective representation c^t that relates these distributed local dynamics to subsequent collective degradation, without using D^t as a training target.

4 Local-to-Collective Predictability Modeling

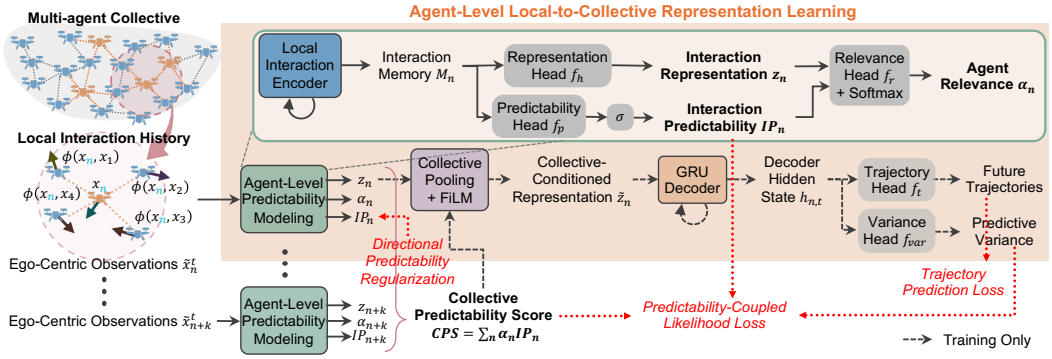


Figure 2: **L2CP architecture.** Agent-level IP and relevance are inferred from ego-centric interaction histories and aggregated into CPS, with future-motion prediction used only for training.

As illustrated in Figure 2, L2CP learns agent-level interaction predictability from ego-centric spatiotemporal interaction histories. Each agent produces an interaction representation z_n, an Interaction Predictability score IP_n, and a relevance weight α_n. Relevance-weighted aggregation of agent-level IP yields the Collective Predictability Score (CPS). During training, these representations condition future-motion prediction, which serves as the representation-learning objective. At inference, IP and relevance are inferred from local interaction histories and aggregated into CPS without requiring the trajectory decoder or future states.

4.1 Interaction Predictability and Collective Aggregation

Local Interaction Encoding. At each time step, the ego-centric relative states of neighboring agents define a local interaction field. Relative positions and velocities are embedded using Fourier features (Li et al., 2021) and processed by a linear-attention set encoder (Shen et al., 2021) to obtain a compact representation of the instantaneous interaction structure. At inference time, these ego-centric observations are sufficient for constructing the learned predictability representations. L2CP does not reconstruct the global system state or require the full interaction graph.

An associative memory inspired by streaming event representations (Kamal et al., 2023) accumulates the interaction history through attention-based access and gated erase/add updates, preserving persistent structure while adapting to neighborhood changes. The local interaction history of agent n is encoded in an interaction memory $\mathbf{M}_n \in \mathbb{R}^{D_m \times D_m}$.

Interaction Predictability. The interaction memory is mapped to an interaction representation $z_n = f_h(\mathbf{M}_n)$ and an agent-level predictability score $\mathrm{IP}_n = \sigma(f_p(\mathbf{M}_n))$. IP is a learned scalar representation of the predictability of an agent's local interaction dynamics. A higher IP indicates that the observed interaction history provides more consistent predictive support for the agent's subsequent motion, whereas a lower value indicates weaker or less consistent predictive structure. Rather than encoding a particular hand-crafted neighborhood property, IP is inferred from the full spatiotemporal interaction memory and can jointly reflect multiple aspects of local interaction dynamics. Agents with similar instantaneous interaction statistics may therefore exhibit different IP when their interaction histories differ.

IP is related to, but not identical to, trajectory uncertainty. Trajectory uncertainty characterizes ambiguity in the predicted future motion, whereas IP is computed from the observed local interaction history and is trained to reflect how strongly that history supports subsequent motion prediction. Accordingly, IP is not itself a trajectory variance, although its learning objective intentionally couples it to predictive difficulty.

Collective Predictability. Agent-level interaction predictability need not contribute equally to the collective dynamics. Some local interaction changes may remain confined to a small neighborhood, whereas others may involve agents more strongly coupled to the evolving group behavior. L2CP therefore learns a relative relevance weight for each agent, $\alpha_n = \mathrm{softmax}_n(f_r([z_n, \mathrm{IP}_n]))$, where z_n provides interaction context for interpreting the corresponding predictability score.

The agent-level predictability signals are aggregated as

$$\mathrm{CPS} = \sum_{n=1}^{N} \alpha_n \, \mathrm{IP}_n.$$

CPS is therefore a collective readout of distributed interaction predictability rather than an independently learned degradation score. The learned weights selectively aggregate heterogeneous agent-level IP signals while preserving IP as an individual-agent representation. Importantly, this aggregation operates on the learned agent-level signals and does not require global absolute coordinates or reconstruction of the full interaction graph.

Collective-Conditioned Trajectory Prediction. During training, future-motion prediction incorporates both the broader collective interaction context and its predictability. A permutation-invariant collective context z_c is obtained by pooling the agent representations, and each local representation is conditioned on both z_c and CPS through feature-wise linear modulation (FiLM) (Perez et al., 2018), $\tilde{z}_n = \mathrm{FiLM}(z_n; [z_c, \mathrm{CPS}])$. Here, z_c captures the broader configuration of interactions across agents, whereas CPS captures their collective predictability.

The conditioned representations $\tilde{z}_n$ are decoded autoregressively using a GRU-based trajectory decoder. Let $h_{n,t}$ denote the decoder hidden state of agent n at prediction step t. Starting from the last observed state, the decoder integrates the current agent state and aggregated inter-agent information and predicts a position increment $\Delta x_n^t = f_t(h_{n,t})$. A separate variance head produces the base predictive log-variance $f_{\mathrm{var}}(h_{n,t})_d$.

4.2 Learning Objective

Neither IP nor CPS uses degradation labels or collective-state metrics as gradient-based training targets. Both are learned from future-motion prediction through trajectory reconstruction, predictability-coupled likelihood, directional regularization, and representation diversity.

Trajectory Prediction. The trajectory prediction loss $\mathcal{L}_{\mathrm{mse}}$ minimizes the mean squared error between predicted and ground-truth future trajectories, propagating predictive supervision through both agent-level representations and collective aggregation.

Predictability-Coupled Likelihood. In addition to the predicted trajectory mean $\hat{x}_{n,t,d}$, the decoder models each future coordinate with a Gaussian variance $\sigma^2_{n,t,d}$. Its log-variance is parameterized as

$$\log \sigma^2_{n,t,d} = f_{\mathrm{var}}(h_{n,t})_d - \beta_i \mathrm{IP}_n - \beta_c \mathrm{CPS},$$

where $f_{\text{var}}(h_{n,t})_d$ represents the trajectory-dependent base uncertainty and $\beta_i = 0.5$, $\beta_c = 0.25$. The resulting mean and variance are optimized using the Gaussian negative log-likelihood

$$\mathcal{L}_{\text{nll}} = \frac{1}{2}\,\text{mean}_{n,t,d}\left[\log \sigma^2_{n,t,d} + \frac{(x_{n,t,d} - \hat{x}_{n,t,d})^2}{\sigma^2_{n,t,d}}\right].$$

This coupling associates higher local and collective predictability with lower predictive variance.

Directional Predictability Regularization. To stabilize the semantic direction of IP across agents, a pairwise ranking loss is applied using stop-gradient trajectory errors. Let e_i denote the mean trajectory error of agent i and $z_i = \text{logit}(\text{IP}_i)$:

$$\mathcal{L}_{\text{IPDir}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \text{softplus}\left(\text{sign}(e_i - e_j)(z_i - z_j)\right),$$

where $\mathcal{P} = \{(i,j) : i \neq j,\ |e_i - e_j| > \tau\}$ and $\tau = 0.05$. This encourages lower-error agents to receive higher IP without using degradation supervision.

Overall Objective. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda_{\text{nll}}\mathcal{L}_{\text{nll}} + \lambda_{\text{IPDir}}\mathcal{L}_{\text{IPDir}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}},$$

where $\mathcal{L}_{\text{div}}$ discourages degenerate agent representations, with $\lambda_{\text{nll}} = 0.1$, $\lambda_{\text{IPDir}} = 0.01$, and $\lambda_{\text{div}} = 10^{-4}$. Additional implementation details are provided in Appendix C.2.

5 Simulation Results

5.1 Simulation Setup

Dataset. Existing multi-agent trajectory datasets are not designed to capture degradation in local interaction dynamics and its subsequent collective consequences. A synthetic dataset is therefore constructed using a Boids-based simulator (Reynolds, 1987), where collective behavior emerges from decentralized neighborhood interactions. This setting enables controlled interaction degradation while preserving the underlying interaction mechanism, allowing direct examination of its relationship with subsequent collective behavior. Full simulation and dataset-generation details are provided in Appendix A.1.

Interaction Degradation. Interaction degradation is induced through sensing delay, motion perturbations, and limited field of view. These mechanisms alter the spatiotemporal interactions available to each agent while leaving the underlying interaction policy unchanged. Degradation parameters are used only to generate the simulated dynamics and are never provided to L2CP as inputs or supervision. Within each sequence, degradation is activated after an initial nominal period and then held fixed while its collective consequences continue to evolve. Exact ranges and implementation details are provided in Appendix A.2.

Evaluation Metrics. Collective degradation is evaluated from future ground-truth simulator states using three complementary metrics: *fragmentation*, *connectivity loss*, and *order loss*. Let $\widetilde{C}(t)$ denote the normalized number of connected components, $\text{LCC}(t) = |\mathcal{C}_{\max}(t)|/N$ the largest-connected-component ratio, and $O(t)$ the velocity polarization. For each time t, degradation over the subsequent $T = 15$ steps is defined as

$$D^t_{\text{frag}} = \max_{\tau=t+1}^{t+T} \widetilde{C}(\tau), \qquad D^t_{\text{conn}} = 1 - \min_{\tau=t+1}^{t+T} \text{LCC}(\tau), \qquad D^t_{\text{order}} = 1 - \min_{\tau=t+1}^{t+T} O(\tau),$$

corresponding respectively to future fragmentation, connectivity loss, and order loss. Each target captures the most severe degradation within the future horizon, with larger values indicating stronger degradation. Predictive alignment is measured using Spearman rank correlation (Spearman, 1987). Full metric definitions are provided in Appendix B.1.

5.2 Agent-Level Predictability under Heterogeneous Degradation

Agent-level IP is evaluated under heterogeneous interaction conditions within the same collective. Half of the agents operate under nominal dynamics, while the remaining agents experience a sensing delay of 5, motion perturbation of 0.1, and a 280° field of view. Because IP represents local interaction predictability rather than degradation severity itself, degradation need not uniformly reduce IP across affected agents. We therefore examine both the overall shift in IP and whether it persists after accounting for simple local interaction statistics.

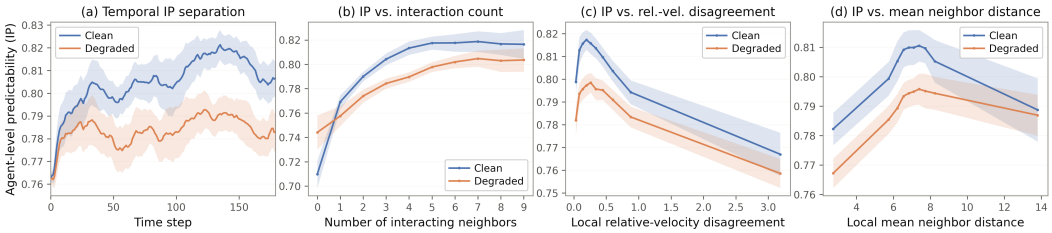


Figure 3: **Agent-level IP under heterogeneous degradation.** (a) Degradation induces a systematic shift in agent-level IP. The clean–degraded separation generally persists at similar (b) interaction counts, (c) local relative-velocity disagreement, and (d) mean neighbor distances.

Figure 3 shows that degraded agents exhibit lower IP on average, with the clean–degraded separation generally persisting at similar interaction counts, local velocity disagreement, and mean neighbor distances. Consistent with this observation, degraded-agent localization using $1 - \text{IP}$ achieves the highest mean AUROC among the evaluated signals, including the combined logistic-regression baseline (Table 2). These results demonstrate that IP captures agent-level information beyond simple instantaneous neighborhood descriptors.

Table 2: **Degraded-agent localization.** Mean AUROC under heterogeneous agent-level degradation.

Signal	AUROC
Interaction counts	0.512
Velocity disagreement	0.571
Neighbor distance	0.529
3 local statistics	0.557
L2CP $(1 - \text{IP})$	**0.596**

5.3 Heterogeneous Contribution of Agent-Level Predictability

Learned relevance is evaluated by progressively retaining agents according to the weights used in CPS aggregation and comparing relevance-selected subsets with random subsets at matched retention rates. Because agents occupy different local interaction contexts, their predictability signals need not contribute equally to the collective readout.

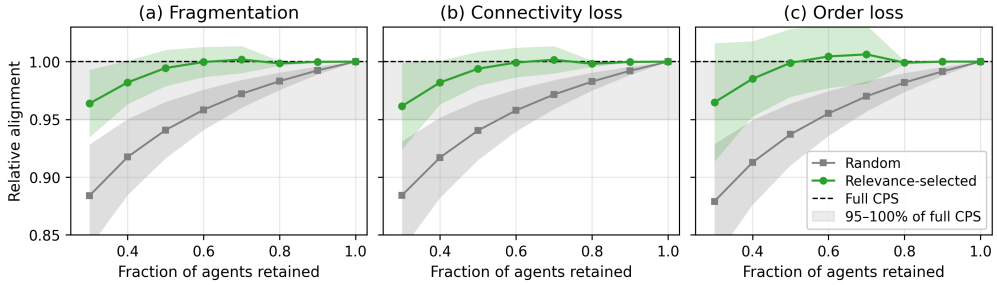


Figure 4: **Partial aggregation of agent-level predictability.** Relevance-selected subsets recover full-agent degradation alignment more rapidly than random subsets.

As shown in Figure 4, relevance-selected subsets recover full-agent degradation alignment more rapidly than random subsets across all three metrics. A small fraction of selected agents preserves much of the full-collective alignment, indicating that agent-level predictability contributes unevenly to the collective readout. These results demonstrate that learned relevance identifies agent-level predictability signals that contribute disproportionately to the collective readout.

5.4 Collective Predictability and Future Degradation

The collective readout is evaluated by comparing CPS with predictive uncertainty from four probabilistic trajectory-prediction models: Trajectron++ (Salzmann et al., 2020), AgentFormer (Yuan et al., 2021), NMRF (Fang et al., 2025), and SingularTrajectory (Bae et al., 2024). For consistent

directionality, the L2CP risk signal is defined as $1 - \text{CPS}$, while trajectory-model uncertainty is measured by the variance across sampled future trajectories (Appendix B.2 and Appendix D.1).

Table 3: **Future collective degradation alignment.** Spearman correlation with future collective degradation. L2CP is reported as mean ± standard deviation over nine seeds.

Model	Modelling Approach	Fragmentation	Connectivity	Order
Trajectron++	Graph-structured CVAE	-0.144	-0.187	-0.048
AgentFormer	Agent-aware Transformer CVAE	0.552	0.448	**0.514**
NMRF	Relational probabilistic modeling	0.419	0.325	0.461
SingularTrajectory	Diffusion-based trajectory modeling	0.067	0.108	0.098
L2CP $(1 - \text{CPS})$	Local-to-collective predictability	$\mathbf{0.688 \pm 0.089}$	$\mathbf{0.600 \pm 0.082}$	0.512 ± 0.063

Table 3 shows that $1 - \text{CPS}$ is more strongly aligned with future fragmentation and connectivity loss than the trajectory-uncertainty baselines, with correlations of 0.688 and 0.600, respectively. For order loss, L2CP and AgentFormer show comparable alignment, indicating that trajectory uncertainty can also reflect aspects of collective coordination. Overall, these results establish CPS as a collective representation that captures degradation-related structure not consistently reflected by conventional trajectory-uncertainty signals.

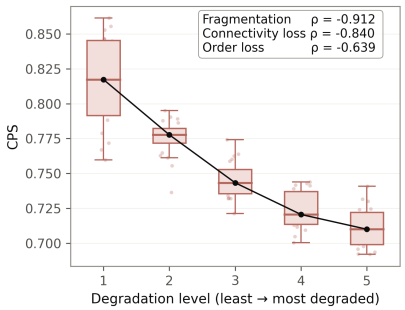


Figure 5: **Collective predictability under interaction degradation.**

Controlled degradation levels are further generated by jointly varying sensing delay, motion perturbation, and field of view. Figure 5 shows a monotonic decrease in CPS with increasing degradation severity despite the absence of direct supervision from degradation levels or collective-state metrics. Together, these results demonstrate that CPS responds systematically to distributed interaction degradation and preserves a consistent relationship with subsequent collective behavior. Additional analyses of its temporal and sequence-level associations are provided in Appendix D.3.

5.5 Evaluation across Collective Dynamics

To evaluate whether the local-to-collective relationship is observed beyond the primary Boids setting, the framework is additionally evaluated under Vicsek (Vicsek et al., 1995) and Cucker–Smale (Cucker & Smale, 2007) dynamics, which differ in their interaction mechanisms. Across all three dynamics, lower collective predictability consistently corresponds to stronger future fragmentation, connectivity loss, and order loss, as shown in Table 4. The correlations are weaker and more variable under Vicsek and Cucker–Smale, indicating that the strength of the relationship depends on the underlying interaction law while remaining qualitatively consistent across distinct collective dynamics.

Table 4: **Evaluation across Collective Dynamics.** Spearman correlation between $1-\text{CPS}$ and future collective degradation over nine random seeds.

Dynamics	Fragmentation	Connectivity Loss	Order Loss
Boids	0.688 ± 0.089	0.600 ± 0.082	0.512 ± 0.063
Vicsek	0.569 ± 0.189	0.498 ± 0.144	0.532 ± 0.146
Cucker	0.435 ± 0.174	0.355 ± 0.149	0.462 ± 0.206

5.6 Ablation Analysis

Table 5 evaluates the roles of relevance weighting, CPS conditioning, and predictability coupling. Each ablation reduces alignment with future fragmentation, connectivity loss, and order loss, demonstrating that all three components contribute to the local-to-collective representation. Full L2CP consistently achieves the strongest degradation alignment across all three metrics.

Table 5: **Ablation analysis.**

Model	Frag.	Conn.	Order
Uniform relevance	0.570	0.492	0.435
No CPS conditioning	0.560	0.480	0.431
No pred. coupling	0.600	0.522	0.457
Full L2CP	**0.688**	**0.600**	**0.512**

5.7 DISCUSSION

The central role of L2CP is to provide an intermediate representation that connects heterogeneous local interaction dynamics to collective behavior.

Interaction predictability represents heterogeneous local interaction condition. IP summarizes how local spatiotemporal interaction structure supports future motion, allowing heterogeneous interaction conditions to remain distinguishable even when instantaneous descriptors are similar. Rather than reproducing a predefined local state metric, IP provides a compact predictive representation of how an agent is embedded in its surrounding interaction dynamics.

Collective predictability connects distributed local signals to the collective level. Because local interaction conditions vary across agents, their predictability signals need not contribute equally to collective behavior. Learned relevance allows CPS to emphasize agent-level signals that contribute more strongly to the collective readout rather than treating the population as homogeneous. The resulting CPS provides a collective summary of distributed interaction predictability that remains associated with collective degradation under persistent interaction changes. Additional analyses show that this relationship is not fully accounted for by realized trajectory prediction error, particularly for fragmentation and connectivity (Appendix D.2). Its temporal and sequence-level structure is examined further in Appendix D.3.

Future-motion-derived supervision organizes the local-to-collective representation. Future prediction serves as a representation-learning objective rather than the end task. Directional predictability regularization and predictability coupling help establish a consistent semantics for the learned representation (Appendix D.8), while degradation labels and collective-state metrics remain outside the training objective. The persistence of the CPS–degradation relationship across Boids, Vicsek, and Cucker–Smale further indicates that the local-to-collective representation is observed across distinct interaction dynamics.

5.8 LIMITATIONS

L2CP is designed to represent how heterogeneous local interaction dynamics relate to collective degradation rather than to optimize a scalar degradation predictor. IP captures agent-level interaction condition, while CPS provides a relevance-weighted collective readout from ego-centric interaction histories. This representation is complementary to handcrafted local statistics and global state or graph metrics, which characterize different aspects of collective behavior.

The relationships among IP, CPS, and collective degradation are associative and do not establish a causal pathway from local predictability changes to system-level degradation. Evaluation is currently limited to simulated collective dynamics with controlled interaction laws and degradation mechanisms. Extending the framework to real multi-agent systems with richer sensing, communication, and control interactions remains an important direction for future work.

6 CONCLUSIONS

This work addresses how heterogeneous local interaction dynamics can be represented and connected to collective degradation in multi-agent systems. By introducing *Interaction Predictability* and relevance-weighted collective aggregation, L2CP learns an intermediate representation linking agent-level interaction histories to collective system behavior through future-motion-derived supervision. This provides a complementary perspective to approaches based on global state variables or explicit fault labels by characterizing how distributed local interaction dynamics relate to subsequent collective degradation.

The resulting representation may support identifying local interaction regions more strongly associated with emerging collective degradation, motivating targeted sensing, communication, and control. More broadly, local interaction predictability could support adaptive resource allocation and closed-loop coordination by providing a structured view of evolving distributed interactions. Extending these ideas to real multi-agent systems remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with experimental and analysis code. We have not used generative AI tools to generate datasets, formulate mathematical claims, or produce mathematical proofs, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for literature search and summarization, code editing, figure refinement, and drafting and editing portions of the manuscript for clarity and readability. We have reviewed all AI-assisted work. All methodological decisions, experimental conclusions, and technical claims were independently evaluated by the authors, and AI-assisted code was verified through implementation and experimental testing. We take responsibility for the final content of this work, including text, claims, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive personal data. The proposed method is evaluated using simulated multi-agent systems. We have not identified additional ethical concerns specific to the methodology or experiments presented in this work.

REPRODUCIBILITY STATEMENT

We provide details supporting the reproducibility of this work throughout the main paper and appendix. The model architecture and learning objectives are described in the main paper, while additional implementation details, experimental settings, data generation procedures, evaluation protocols, and hyperparameters are provided in the appendix. The source code and datasets used in this work will be made publicly available upon acceptance.

REFERENCES

Filippo Arrichiello, Alessandro Marino, and Francesco Pierri. Distributed fault detection and recovery for networked robots. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3734–3739. IEEE, 2014.

Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17890–17901. IEEE, 2024.

Alessandro Carminati, Davide Azzalini, Simone Vantini, and Francesco Amigoni. A distributed approach for fault detection in swarms of robots. In *Proceedings of the 23rd international conference on autonomous agents and multiagent systems*, pp. 253–261, 2024.

Andrea Cavagna, Antonio Culla, Xiao Feng, Irene Giardina, Tomas S Grigera, Willow Kion-Crosby, Stefania Melillo, Giulia Pisegna, Lorena Postiglione, and Pablo Villegas. Marginal speed confinement resolves the conflict between correlation and control in collective behaviour. *Nature Communications*, 13(1):2315, 2022.

Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Transactions on automatic control*, 52(5):852–862, 2007.

Kai Cui, Sascha Hauck, Christian Fabian, and Heinz Koeppl. Learning decentralized partially observable mean field control for artificial collective behavior. In *International Conference on Learning Representations*, volume 2024, pp. 40755–40794, 2024.

Zilin Fang, David Hsu, and Gim H Lee. Neuralized markov random field for interaction-aware stochastic human trajectory prediction. In *International Conference on Learning Representations*, volume 2025, pp. 73878–73893, 2025.

Joël Foramitti. Agentpy: A package for agent-based modeling in python. *Journal of Open Source Software*, 6(62):3065, 2021.

Feixiang He, Jiangbei Yue, Jialin Zhu, Armin Seyfried, Dan Casas, Julien Pettré, and He Wang. Learning extremely high density crowds as active matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 540–550. IEEE, 2025.

Uday Kamal, Saurabh Dash, and Saibal Mukhopadhyay. Associative memory augmented asynchronous spatiotemporal representation learning for event-based perception. In *The Eleventh International Conference on Learning Representations*, 2023.

Minah Lee and Saibal Mukhopadhyay. Sigmas: Second-order interaction-based grouping for overlapping multi-agent swarms. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '26, pp. 125–133, Richland, SC, 2026. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400723179. doi: 10.65109/MZXX4786. URL https://doi.org/10.65109/MZXX4786.

Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems*, 34:15816–15829, 2021.

Zhongkui Li, Guanghui Wen, Zhisheng Duan, and Wei Ren. Designing fully distributed consensus protocols for linear multi-agent systems with directed graphs. *IEEE Transactions on Automatic Control*, 60(4):1152–1157, 2014.

James O'Keeffe. Anticipating degradation: a predictive approach to fault tolerance in robot swarms. *IEEE Robotics and Automation Letters*, 10(9):8954–8961, 2025.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Ahmad Rahimi, Po-Chien Luan, Yuejiang Liu, Frano Rajič, and Alexandre Alahi. Sim-to-real causal transfer: A metric learning approach to causally-aware interaction representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17271–17281, 2025.

Wei Ren and Randal W Beard. Consensus seeking in multiagent systems under dynamically changing interaction topologies. *IEEE Transactions on automatic control*, 50(5):655–661, 2005.

Craig W Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pp. 25–34, 1987.

Hamed Rezaee, Thomas Parisini, and Marios M Polycarpou. Almost sure resilient consensus under stochastic interaction: Links failure and noisy channels. *IEEE Transactions on Automatic Control*, 66(12):5727–5741, 2020.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European conference on computer vision*, pp. 683–700. Springer, 2020.

Matteo Santilli, Pratik Mukherjee, Ryan K Williams, and Andrea Gasparri. Multirobot field of view control with adaptive decentralization. *IEEE Transactions on Robotics*, 38(4):2131–2150, 2022.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3531–3539, 2021.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987.

Bohan Tang, Yiqi Zhong, Ulrich Neumann, Gang Wang, Siheng Chen, and Ya Zhang. Collaborative uncertainty in multi-agent trajectory forecasting. *Advances in Neural Information Processing Systems*, 34:6328–6340, 2021.

Danesh Tarapore, Jon Timmis, and Anders Lyhne Christensen. Fault detection in a swarm of physical robots based on behavioral outlier detection. *IEEE Transactions on Robotics*, 35(6):1516–1522, 2019.

Neerja Thakkar, Tara Sadjadpour, Jathushan Rajasegeran, Shiry Ginosar, and Jitendra Malik. Poly-autoregressive prediction for modeling interactions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12402–12412. IEEE, 2025.

Masahito Uwamichi, Simon K Schnyder, Tetsuya J Kobayashi, and Satoshi Sawai. Integrating gnn and neural odes for estimating non-reciprocal two-body interactions in mixed-species collective motion. *Advances in Neural Information Processing Systems*, 37:72625–72647, 2024.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.

Di Wen, Haoran Xu, Zhaocheng He, Zhe Wu, Guang Tan, and Peixi Peng. Density-adaptive model based on motif matrix for multi-agent trajectory prediction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14822–14832. IEEE, 2024.

Wei Xiang, YIN Haoteng, He Wang, and Xiaogang Jin. Socialcvae: Predicting pedestrian trajectory via interaction conditioned latents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6216–6224, 2024.

Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9813–9823, 2021.

Fang Zheng, Le Wang, Sanping Zhou, Wei Tang, Zhenxing Niu, Nanning Zheng, and Gang Hua. Unlimited neighborhood interaction for heterogeneous trajectory prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13148–13157. IEEE, 2021.

Zhicheng Zheng, Yongjian Zhou, Yalun Xiang, Xiaokang Lei, and Xingguang Peng. Emergence of collective behaviors for the swarm robotics through visual attention-based selective interaction. *IEEE Robotics and Automation Letters*, 9(11):9399–9406, 2024.

Zhicheng Zheng, Tao Wang, Yalun Xiang, Xiaokang Lei, and Xingguang Peng. Shaping collective behaviors in swarm robotics through probabilistic motion decision-making. *IEEE Robotics and Automation Letters*, 2025.

A Simulation and Dataset Details

A.1 Simulation Environment

Table 6: Simulation parameters used for dataset generation.

Parameter	Value
Number of agents	20
Spatial domain	100×100
Simulation length	300 steps
Discarded initial steps	100
Retained sequence length	200 steps
Inner interaction radius	5
Outer interaction radius	10
Cohesion strength	0.01
Separation strength	0.1
Alignment strength	0.2
Preferred speed range	$[0.8, 1.2]$
Center dead-zone radius	5
Center-pull gain	0.005

The dataset is generated using AgentPy (Foramitti, 2021), a two-dimensional Boids-style flocking simulator. Each simulation contains 20 agents initialized at random positions and headings. Agent motion is governed by cohesion, separation, and alignment interactions, together with a weak center-directed steering term that prevents the flock from drifting away from the interaction region. Agent orientation is defined by the instantaneous velocity direction, and interactions are restricted by both spatial proximity and the agent's field of view (FOV).

Each simulation is executed for 300 time steps. Since the initial portion mainly captures transient flock formation from random initialization, the first 100 steps are discarded. The remaining 200 steps are retained for model training and evaluation. Table 6 summarizes the simulation parameters.

A.2 Dataset Generation and Interaction Degradation

Table 7: Interaction degradation ranges used for dataset generation.

Parameter	Range
Sensing delay	0–5 steps
Motion perturbation (heading std.)	0–0.2 rad
Field of view	200°–360°
Latent interaction quality	$q \sim \mathcal{U}(0, 1)$
Component-wise quality perturbation	$\mathcal{N}(0, 0.05^2)$

A total of 1,200 simulation sequences are generated and subsequently partitioned into 960 training sequences, 120 validation sequences, and 120 test sequences. L2CP processes the retained interaction sequence sequentially as a stream rather than using a fixed observation window.

Interaction conditions are varied continuously across sequences. For each simulation, a latent interaction-quality variable $q \sim \mathcal{U}(0, 1)$ is first sampled. Independent Gaussian perturbations with standard deviation 0.05 are added to q for the delay, motion perturbation, and FOV components, followed by clipping to $[0, 1]$. The resulting component-wise quality values are mapped to the degradation ranges summarized in Table 7. This construction produces correlated degradation across the three interaction mechanisms while retaining moderate variation among individual factors.

The initial 100-step transient period uses the nominal interaction condition with zero sensing delay, no motion perturbation, and a 360° FOV. After this period, the sampled degradation parameters are activated and remain fixed for the rest of the simulation. Sensing delay is implemented using historical agent states, such that a delay of d uses interaction information from d time steps in the past. Motion perturbation is introduced by rotating the updated velocity direction by an angle sampled from a zero-mean Gaussian distribution, after which the velocity magnitude is restored to the agent-specific preferred speed. FOV degradation restricts perceived neighbors to an angular sector centered on the current heading.

B Evaluation Details

B.1 Collective Degradation Metrics and Evaluation Protocol

Collective degradation is evaluated from the ground-truth simulator state rather than from the degraded local observations available to the model. At each time step, an undirected interaction graph is constructed from the current agent positions, with an edge between agents i and j when

$$\|\mathbf{x}_i^t - \mathbf{x}_j^t\|_2 < r,$$

where $r = 10$ is the outer interaction radius.

Fragmentation. Let $C(t)$ denote the number of connected components in the interaction graph. The component count is normalized within each simulated sequence as

$$\widetilde{C}(t) = \frac{C(t) - 1}{\max_\tau C(\tau) - 1 + \epsilon},$$

where $\epsilon = 10^{-8}$. Larger values indicate stronger fragmentation.

Connectivity loss. Let $\mathcal{C}_{\max}(t)$ denote the largest connected component. The largest-connected-component ratio is defined as

$$\mathrm{LCC}(t) = \frac{|\mathcal{C}_{\max}(t)|}{N}.$$

A lower $\mathrm{LCC}(t)$ indicates greater loss of collective connectivity.

Order loss. Collective order is quantified using the polarization of agent velocity directions,

$$O(t) = \left\| \frac{1}{N} \sum_{n=1}^{N} \frac{\mathbf{v}_n^t}{\|\mathbf{v}_n^t\|_2 + \epsilon} \right\|_2.$$

A lower $O(t)$ indicates weaker directional agreement across the collective.

Predictive signals at time t are evaluated against the worst collective degradation occurring over the subsequent $T = 15$ time steps. The future evaluation interval is $\tau \in \{t+1, \ldots, t+T\}$. The corresponding future-degradation targets are defined as

$$D_{\mathrm{frag}}^t = \max_{\tau=t+1}^{t+T} \widetilde{C}(\tau),$$

$$D_{\mathrm{conn}}^t = 1 - \min_{\tau=t+1}^{t+T} \mathrm{LCC}(\tau),$$

and

$$D_{\mathrm{order}}^t = 1 - \min_{\tau=t+1}^{t+T} O(\tau).$$

Thus, each target captures the most severe degradation observed within the future horizon. Timesteps without a complete T-step future interval are excluded from evaluation.

Each valid $(\mathrm{sequence}, t)$ pair is treated as one system-level sample, and samples from all test sequences are pooled. Predictive alignment is measured using signed Spearman rank correlation. For L2CP, the collective predictability score is converted to a risk signal as $1 - \mathrm{CPS}$, such that larger values indicate greater predicted degradation risk.

B.2 Trajectory-Prediction Uncertainty

To compare L2CP with probabilistic trajectory-prediction baselines, each baseline prediction is reduced to a single system-level uncertainty score at each time step. For sample-based predictors, predictive uncertainty is computed as the empirical variance across predicted future trajectories and then averaged over agents, prediction steps, and spatial dimensions. Specifically,

$$U^t = \log\left(\frac{1}{NTD} \sum_{n=1}^{N} \sum_{\tau=1}^{T} \sum_{d=1}^{D} \mathrm{Var}_k \left[\hat{x}_{n,k,d}^{t+\tau} \right] + \epsilon \right),$$

where k indexes trajectory samples, $D = 2$ is the spatial dimension, and $\epsilon = 10^{-8}$. This formulation is used for AgentFormer (Yuan et al., 2021), NMRF (Fang et al., 2025), and SingularTrajectory (Bae et al., 2024) with 20 predicted trajectories and a 15-step prediction horizon.

Trajectron++ (Salzmann et al., 2020) directly parameterizes a Gaussian predictive distribution. Its uncertainty score is therefore obtained from the model-predicted coordinate-wise log variances rather than from empirical trajectory samples:

$$U^t_{\mathrm{T++}} = \frac{1}{NTD} \sum_{n=1}^{N} \sum_{\tau=1}^{T} \sum_{d=1}^{D} \log \sigma^2_{n,\tau,d},$$

with $T = 15$. In all cases, uncertainty is aggregated into one scalar per $(\mathrm{sequence}, t)$ pair, with larger values indicating greater predictive uncertainty. These scalar uncertainty scores are then evaluated against the same future collective-degradation targets described above.

C Implementation Details

C.1 L2CP Architecture

Table 8: L2CP architecture configuration.

Parameter	**Value**
Neighbor input features	4
Set-encoder layers	3
Set-encoder attention heads	4
Interaction memory size	4×4
Agent representation dimension	128
Collective representation dimension	128
Collective pooling	Mean, max, std
Decoder hidden dimension	128
Prediction horizon	15

L2CP receives ego-centric relative interaction features consisting of forward/lateral relative position and velocity. Position and velocity components are normalized by $(10, 10)$ and $(2.4, 2.4)$, respectively, and independently encoded using learnable Fourier features. The resulting neighbor tokens are processed by a three-layer linear-attention set encoder with four attention heads.

Each agent maintains a 4×4 associative interaction memory that is updated sequentially through attention-based read/write operations and gated recurrent fusion. The flattened memory state is mapped to a 128-dimensional agent representation z_n. Agent-level interaction predictability IP_n is produced by a two-layer MLP followed by a sigmoid activation.

Collective context is obtained by concatenating the mean, maximum, and standard deviation of the agent representations and mapping the resulting feature to a 128-dimensional collective representation. Agent relevance is computed from $[z_n, \mathrm{IP}_n]$ and normalized across agents using softmax. The collective predictability score is then

$$\mathrm{CPS} = \sum_{n=1}^{N} \alpha_n \mathrm{IP}_n,$$

where α_n denotes the learned relevance weight.

The collective representation and CPS jointly condition each agent representation through FiLM. The conditioned representation is decoded using a 128-dimensional GRUCell with inter-agent message passing to autoregressively predict 15 future displacements. A separate variance head outputs diagonal log-variance for the two spatial dimensions. The predictive variance is coupled to agent- and collective-level predictability as

$$\log \sigma^2_{n,t,d} = f_{\mathrm{var}}(h_{n,t})_d - \beta_i \mathrm{IP}_n - \beta_c \mathrm{CPS},$$

where $\beta_i = 0.5$ and $\beta_c = 0.25$ are fixed coefficients.

C.2 Training and Optimization

L2CP is trained in a streaming manner without a fixed observation window. The interaction memory is initialized once at the beginning of each sequence and retained as the sequence is processed. Training sequences are divided into 15-step chunks. After each chunk, the model predicts the subsequent 15-step trajectory, performs one optimization update, and detaches the memory state from the computational graph while retaining its content. This limits backpropagation through time while preserving interaction information accumulated across the sequence.

Trajectory Reconstruction Loss. The predicted future trajectory is supervised using mean squared error,

$$\mathcal{L}_{\text{mse}} = \frac{1}{NT} \sum_{n=1}^{N} \sum_{\tau=1}^{T} \left\| x_n^{t+\tau} - \hat{x}_n^{t+\tau} \right\|_2^2 .$$

This term trains the trajectory decoder and propagates predictive supervision through the interaction representation and collective conditioning used to generate the future trajectory.

Predictability-Coupled Likelihood Loss. For each agent, prediction step, and spatial dimension, the decoder outputs a trajectory mean together with a base log-variance $f_{\text{var}}(h_{n,\tau})_d$. The final predictive log-variance is modulated by agent- and collective-level predictability as

$$\log \sigma_{n,\tau,d}^2 = f_{\text{var}}(h_{n,\tau})_d - \beta_i \text{IP}_n - \beta_c \text{CPS},$$

where $\beta_i = 0.5$ and $\beta_c = 0.25$. The corresponding Gaussian negative log-likelihood is

$$\mathcal{L}_{\text{nll}} = \frac{1}{2} \, \text{mean}_{n,\tau,d} \left[\log \sigma_{n,\tau,d}^2 + \frac{\left(x_{n,\tau,d} - \hat{x}_{n,\tau,d} \right)^2}{\sigma_{n,\tau,d}^2} \right].$$

This parameterization imposes a monotonic relationship in which higher interaction predictability corresponds to lower modeled predictive variance, while the base variance head retains trajectory-dependent uncertainty.

Directional Predictability Regularization. To stabilize the semantic direction of IP, we additionally use a pairwise ranking loss based on agent-level trajectory errors. For agent i, the trajectory error is computed as

$$e_i = \frac{1}{T} \sum_{\tau=1}^{T} \left\| \hat{x}_i^{t+\tau} - x_i^{t+\tau} \right\|_2 ,$$

with the trajectory error detached from the computational graph before constructing the ranking targets.

Let $z_i = \text{logit}(\text{IP}_i)$. For ordered agent pairs whose trajectory errors differ by more than a threshold τ_e, we define

$$\mathcal{P} = \{ (i,j) \mid i \neq j, \ |e_i - e_j| > \tau_e \} .$$

The directional regularization loss is

$$\mathcal{L}_{\text{IPDir}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \text{softplus} \left[\text{sign}(e_i - e_j)(z_i - z_j) \right] .$$

This encourages agents with lower trajectory error to receive higher IP while avoiding direct supervision from degradation labels or collective-state metrics. We use $\tau_e = 0.05$.

Representation Diversity. A small diversity regularization term $\mathcal{L}_{\text{div}}$ is applied to the agent representations to discourage representation collapse and preserve variation across agents.

Overall Objective. The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda_{\text{nll}} \mathcal{L}_{\text{nll}} + \lambda_{\text{IPDir}} \mathcal{L}_{\text{IPDir}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}},$$

with $\lambda_{\text{nll}} = 0.1$, $\lambda_{\text{IPDir}} = 0.01$, and $\lambda_{\text{div}} = 10^{-4}$.

Optimization uses AdamW with an initial learning rate of 10^{-3} and weight decay of 0.03. The learning rate is reduced by a factor of 0.5 every 25 epochs. Training uses a batch size of 256 and gradient clipping with a maximum norm of 100. The checkpoint achieving the lowest validation ADE is used for evaluation.

Trajectory-prediction baselines. Trajectron++, AgentFormer, NMRF, and SingularTrajectory are trained using their official implementations with the original model architectures and training procedures, replacing only the dataset with the same swarm train/validation/test splits used for L2CP. No degradation labels are provided to the baselines. Predictive uncertainty is extracted from each trained model.

D Additional Experimental Results

D.1 Trajectory Prediction as Predictive Supervision

Table 9: Trajectory prediction accuracy on the swarm test set. ADE/FDE denotes the error of a single deterministic prediction for L2CP and the mean error across sampled predictions for stochastic baselines.

Model	**ADE**	**FDE**
L2CP	4.11	7.17
Trajectron++	3.58	7.83
AgentFormer	3.55	7.32
NMRF	3.44	7.17
SingularTrajectory	2.80	5.98

Trajectory prediction is used as a training objective for representation learning rather than as the end task of L2CP. We nevertheless report trajectory accuracy to contextualize the predictive supervision used during training. All results are reported in raw simulator coordinate units. Since L2CP produces a single deterministic rollout, the comparison uses the mean trajectory error across samples for stochastic baselines. All models are evaluated on the same 120 test sequences and 180 retained timesteps per sequence.

D.2 Relation to Realized Trajectory Prediction Error

Because IP is learned using future-motion-derived objectives, we examine whether the association between CPS and collective degradation can be explained by trajectory prediction error alone. For each time t, we define the mean realized trajectory error

$$E^t_{\mathrm{mean}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T}\sum_{\tau=1}^{T}\left\|\hat{x}_n^{t+\tau} - x_n^{t+\tau}\right\|_2 .$$

Table 10: Correlation with future collective degradation and partial correlation after controlling for realized trajectory error. Values are mean $\pm$ standard deviation over nine seeds.

Signal	Fragmentation	Connectivity loss	Order loss
$\rho(1-\mathrm{CPS}, D)$	0.688 ± 0.089	0.600 ± 0.082	0.512 ± 0.063
$\rho(E_{\mathrm{mean}}, D)$	0.746 ± 0.011	0.646 ± 0.024	0.660 ± 0.025
$\rho(1-\mathrm{CPS}, D \mid E_{\mathrm{mean}})$	0.390 ± 0.134	0.303 ± 0.108	0.129 ± 0.048

Although realized trajectory error exhibits stronger raw correlation with degradation, it requires access to future ground-truth trajectories and is therefore unavailable at inference time. In contrast, CPS is computed solely from interaction histories observed up to the current time. Partial Spearman correlation is computed by rank-transforming the variables and correlating the residuals after regressing both variables on E_{mean}. After controlling for realized trajectory error, $1-\mathrm{CPS}$ retains positive partial correlation with future fragmentation and connectivity loss, indicating that its association with structural degradation is not fully explained by trajectory prediction error alone. The residual association is weaker for order loss.

D.3 Between- and Within-Sequence Association

Because each degradation condition remains fixed within a sequence, the pooled correlation between $1-\mathrm{CPS}$ and future collective degradation can reflect both differences across sequences and temporal variation within a sequence. We therefore decompose the association using the same test samples and nine final checkpoints as in the main evaluation.

For the within-sequence analysis, sequence-specific means are removed from both $1-\mathrm{CPS}$ and each degradation target before computing the pooled Spearman correlation. For the between-sequence

analysis, each quantity is first averaged over time within a sequence and the correlation is then computed across sequences.

Table 11: Between- and within-sequence association between $1 - \mathrm{CPS}$ and future collective degradation. Values are mean $\pm$ standard deviation over nine seeds.

Metric	Pooled	Within-seq.	Between-seq.
Fragmentation	0.688 ± 0.089	0.252 ± 0.137	0.900 ± 0.031
Connectivity	0.600 ± 0.082	0.228 ± 0.110	0.803 ± 0.046
Order	0.512 ± 0.063	0.148 ± 0.091	0.694 ± 0.049

The association is stronger across sequences than within individual sequences, while positive within-sequence correlations remain across all three metrics. This decomposition characterizes the time scales reflected by CPS under the persistent degradation setting considered here.

D.4 Agent-Level Localization Analysis

We quantify degraded-agent localization in the heterogeneous setting of Table 2. The evaluation contains 20 sequences with 20 agents each, half nominal and half degraded, yielding 72,000 agent–time samples. To ensure identical evaluation conditions across all methods, we retain the 69,184 samples for which all local statistics are defined, excluding zero-neighbor windows.

We compare $1-\mathrm{IP}$ with per-agent neighbor count, local relative-velocity disagreement, mean neighbor distance, and a logistic-regression baseline combining all three statistics. The logistic-regression baseline is trained using nominal/degraded labels, whereas $1 - \mathrm{IP}$ is evaluated directly from the pretrained L2CP model without localization-specific fitting or degradation-label supervision. All methods use the same 20 repeated sequence-level splits, with 25% of sequences held out for testing and feature standardization estimated from training data only.

AUROC is reported on the shared held-out splits. For IP, results are aggregated over nine model seeds and 20 splits. For the local-statistics baselines, results are reported over the same 20 splits.

D.5 Comparison with Simple Local Statistics

To contextualize the collective readout, we compare $1 - \mathrm{CPS}$ with simple aggregate statistics computed from the same ego-centric local observations. Mean neighbor count, local velocity disagreement, and mean neighbor distance are averaged across agents and evaluated using the same future-degradation targets.

Table 12: Spearman correlation with future collective degradation. Simple statistics are computed from the same local observations available to L2CP.

Signal	Fragmentation	Connectivity	Order
$1 - \mathrm{CPS}$	0.688 ± 0.089	0.600 ± 0.082	0.512 ± 0.063
Mean neighbor count	0.834	0.779	0.581
Velocity disagreement	0.586	0.475	0.521
Mean neighbor distance	0.232	0.211	0.220

Simple aggregate statistics exhibit different alignment patterns across degradation measures, with interaction density showing particularly strong association with fragmentation and connectivity. This comparison further positions CPS as a learned collective representation of distributed interaction dynamics rather than a task-specific scalar degradation score.

D.6 Degradation-Level Experimental Details

Table 13: Degradation settings used in the severity-sweep experiment. Each condition jointly varies sensing delay, motion perturbation, and field of view.

Degradation Level	**Delay**	**Motion Perturbation**	**FOV (deg.)**
1	0–1	0.020–0.057	318–345
2	1–2	0.054–0.094	276–306
3	2–3	0.104–0.141	251–274
4	3–4	0.146–0.177	217–259
5	4–5	0.181–0.200	200–212

The degradation-level analysis jointly varies sensing delay, motion perturbation, and field of view as summarized in Table 13. Each level contains 20 independently generated sequences with 180 retained timesteps per sequence. The levels are ordered from the least to the most degraded condition.

D.7 Additional Partial-Aggregation Analysis

Agents are retained by thresholding the learned relevance score used in CPS aggregation. For each retained fraction, CPS is recomputed from the selected agents, and alignment is measured using the pooled Spearman correlation between $1 - \mathrm{CPS}$ and each future collective-degradation target. Relative alignment is defined as

$$R_f = \frac{\rho_f}{\rho_{\mathrm{full}}},$$

where ρ_f and ρ_{full} denote the partial- and full-agent correlations, respectively. A rate-matched random-selection baseline is evaluated over 20 independent random seeds.

D.8 Disentangling IPDir and Predictability Coupling

Table 14: Ablation of directional predictability regularization and predictability coupling. Results are Spearman correlation between $1 - \mathrm{CPS}$ and future collective degradation, reported as mean $\pm$ standard deviation over nine seeds.

Variant	Fragmentation	Connectivity	Order
Full L2CP	0.688 ± 0.089	0.600 ± 0.082	0.512 ± 0.063
w/o IPDir	0.591 ± 0.201	0.525 ± 0.178	0.454 ± 0.130
w/o pred. coupling	0.600 ± 0.226	0.522 ± 0.200	0.457 ± 0.176
w/o both	0.410 ± 0.400	0.356 ± 0.356	0.343 ± 0.302

To examine the roles of directional predictability regularization and predictability coupling, we compare Full L2CP with variants that remove IPDir, predictability coupling, or both. Table 14 reports the alignment between $1 - \mathrm{CPS}$ and future collective degradation over nine random seeds.

Removing either IPDir or predictability coupling reduces the average collective degradation alignment and increases variation across seeds. Removing both produces the weakest and least stable alignment. These results show that the two objectives improve average degradation alignment and reduce seed-level variability when used together.

D.9 Hyperparameter Sensitivity

Table 15: Sensitivity to predictability coupling and IPDir. Values are Spearman correlations between $1 - \mathrm{CPS}$ and future collective degradation, reported as mean $\pm$ standard deviation over nine seeds.

Setting	Fragmentation	Connectivity	Order
Predictability coupling ($\lambda_{\mathrm{IPDir}} = 0.01$)			
$(\beta_i, \beta_c) = (0.25, 0.125)$	0.469 ± 0.227	0.409 ± 0.201	0.361 ± 0.187
$(\beta_i, \beta_c) = (0.50, 0.25)$	$\mathbf{0.688 \pm 0.089}$	$\mathbf{0.600 \pm 0.082}$	$\mathbf{0.512 \pm 0.063}$
$(\beta_i, \beta_c) = (1.00, 0.50)$	0.555 ± 0.220	0.474 ± 0.194	0.423 ± 0.134
Directional regularization ($(\beta_i, \beta_c) = (0.50, 0.25)$)			
$\lambda_{\mathrm{IPDir}} = 0.005$	0.507 ± 0.207	0.443 ± 0.182	0.405 ± 0.124
$\lambda_{\mathrm{IPDir}} = 0.01$	$\mathbf{0.688 \pm 0.089}$	$\mathbf{0.600 \pm 0.082}$	$\mathbf{0.512 \pm 0.063}$
$\lambda_{\mathrm{IPDir}} = 0.02$	0.431 ± 0.430	0.363 ± 0.382	0.315 ± 0.345

We evaluate sensitivity around the selected setting $(\beta_i, \beta_c) = (0.5, 0.25)$ and $\lambda_{\mathrm{IPDir}} = 0.01$. The selected values are chosen on the validation split using collective-degradation alignment. Table 15 reports performance under neighboring coefficient settings.