

Topology Enhanced MARL for Multi-Agent Cooperative Decision-Making of CAVs

Ye Han, Lijun Zhang*, Dejian Meng, Zhuang Zhang

Abstract—Decentralized multi-agent cooperative decision-making in continuous environments is fundamentally bottlenecked by the curse of dimensionality, where undirected exploration typically converges to conservative local optima. We propose Topology-Enhanced Multi-Agent Reinforcement Learning (TPE-MARL) to reformulate multi-agent exploration as a structured topological traversal. We introduce the Game Topology Tensor, utilizing locality-sensitive hashing to project the continuous physical manifold into a discrete quotient space. This abstraction operates as an adversarial Information Bottleneck, decoupling strategic coordination intents from environmental noise. Within this space, a dual intrinsic reward mechanism drives exploration: a novelty reward maximizes the marginal entropy of visited topologies, while a collaboration reward, optimized via a variational Evidence Lower Bound (ELBO), minimizes conditional entropy to exploit cooperative joint configurations. Evaluations demonstrate that TPE-MARL achieves near-optimal decision distributions, closely approximating the theoretical bounds established by a Monte Carlo Tree Search (MCTS) oracle. Furthermore, physical testbed experiments validate the framework’s zero-shot out-of-distribution (OOD) generalization. Supported by a spatial relaxation mechanism, the learned representations reliably execute dynamic negotiations, such as cooperative zipper-merging, exhibiting inherent robustness against real-world covariate shifts and actuation latencies.

I. INTRODUCTION

The pursuit of artificial cooperative intelligence in open, dynamic environments remains one of the most formidable challenges in machine learning [1]. Formulated as Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs), multi-agent systems must synthesize coherent joint policies from coupled local observations [2], [3]. While deep Multi-Agent Reinforcement Learning (MARL) paradigms, such as value factorization [4], [5] and multi-agent policy gradients [6], [7], have achieved remarkable success in controlled domains, their deployment in safety-critical cyber-physical systems exposes a fundamental vulnerability: the exponential explosion of the continuous joint state-action space, notoriously known as the *curse of dimensionality*.

Navigating this intractable space via passive ϵ -greedy exploration invariably leads to sample inefficiency and convergence to conservative local optima, often described as the sparse reward problem [8], [9]. To mitigate this, recent advancements have introduced intrinsic motivation, encouraging agents to explore novel states [10], [11], [12] or maximize behavioral diversity [13], [14], [15], [16], [17]. However, in tightly coupled, non-stationary multi-agent games, undirected diversity

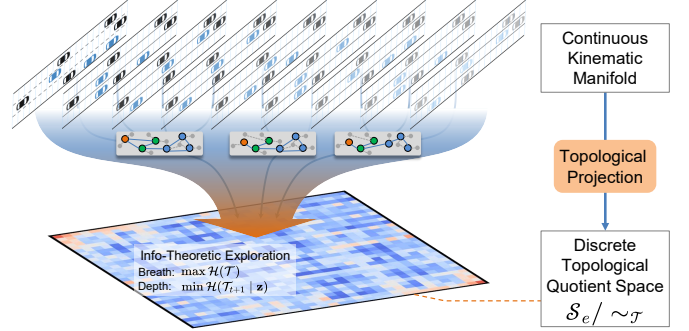


Fig. 1: Conceptual illustration of TPE-MARL. The framework projects the noisy, continuous kinematic manifold (Top) through an LSH-based Information Bottleneck (Middle) into a discrete topological quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$ (Bottom). Within this space, a dual information-theoretic engine drives exploration by balancing marginal entropy maximization (breadth) and conditional entropy minimization (depth).

often proves catastrophic. Blindly maximizing the mutual information between latent skills and environmental trajectories indiscriminately scatters computational resources across strategically trivial configurations, fundamentally failing to answer the critical question of *how to proactively explore* the most profound game-theoretic interactions.

Our core insight is that the continuous kinematic manifold of the physical world is highly redundant for strategic decision-making. Microscopic positional shifts rarely alter the underlying Nash equilibrium of the local game. Therefore, optimal coordination does not require exploring the full continuous space, but rather discovering the correct equivalence classes of spatial interactions.

In this paper, we propose a paradigm shift (illustrated in Fig. 1): transferring the exploration imperative from the continuous physical manifold to a discrete, structured *quotient space*. We establish this mapping by extracting the *minimal sufficient statistic* of the multi-agent interaction, instantiated as the Game Topology Tensor. By leveraging locality-sensitive hashing [18], we construct a topology-preserving, collision-resistant discrete representation that inherently filters out high-frequency physical noise while strictly preserving low-frequency game-theoretic intents.

Operating within this topological quotient space, we formulate the exploration process as a mathematically rigorous joint information capacity optimization problem, termed **Topology-Enhanced MARL (TPE-MARL)**. Driven by a dual intrinsic reward mechanism, the framework seamlessly resolves the exploration-exploitation dilemma. First, a topology novelty reward systematically maximizes the marginal entropy of the

Ye Han, Lijun Zhang, Dejian Meng, Zhuang Zhang are with the School of Automotive Studies, Tongji University, Shanghai 201804, China. {hanye_leohancnjs, tjedu_zhanglijun, mengdejian, zhuang_zhang}@tongji.edu.cn

*Corresponding author: Lijun Zhang

visited topologies, guaranteeing exploration breadth. Second, a collaboration reward, optimized via a tractable variational Evidence Lower Bound (ELBO), minimizes conditional entropy, guiding agents to exploit deeply cooperative, maximum-mutual-information joint configurations.

The main contributions of this paper are fundamentally four-fold:

- **Quotient Space Abstraction for MARL:** We theoretically formalize the Game Topology Tensor, proposing a mathematically rigorous mechanism to map continuous, non-stationary Dec-POMDPs into discrete topological quotient spaces. This structural abstraction inherently bounds the exploration space while preserving sufficient statistics required for optimal policy invariance.
- **Information-Theoretic Exploration Engine:** We design an adversarial Variational Information Bottleneck (VIB) to explicitly decouple strategic coordination intents from environmental stochasticity. This dual intrinsic reward mechanism advances spatial MARL architectures by equipping them with a proactive, structure-aware exploration engine.
- **Rigorous Theoretical Optimality Bounds:** Beyond generic heuristic benchmarks, we evaluate the learned decentralized policies against an exact, full-state oracle powered by a MCTS solver. We empirically prove that TPE-MARL systematically avoids conservative local optima, aligning its Q-value distributions with the theoretical upper bounds across joint action spaces.
- **Physical Deployment and OOD Robustness:** Validated in a cyber-physical testbed, the framework achieves zero-shot simulation-to-real transfer. It successfully executes high-dimensional dynamic negotiation, demonstrating inherent resilience against real-world covariate shifts, including localization latencies and tracking errors.

The remainder of this paper is organized as follows. Section II reviews related work. Section III mathematically formulates the Dec-POMDP and the information space. Section IV details the Game Topology Tensor, the variational mutual information estimation, and the complete TPE-MARL framework. Section V presents extensive algorithmic evaluations and MCTS optimality bounds. Section VI provides physical cyber-physical validations. Finally, Section VII discusses theoretical boundaries, and Section VIII concludes the paper.

II. RELATED WORKS

A. Structural Representation in Multi-Agent Systems

The fundamental challenge in Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) lies in effectively fusing highly coupled, varying-length observations into coherent joint representations. Historically, to overcome the limitations of rudimentary Multi-Layer Perceptrons (MLPs), the MARL community has extensively adopted Graph Neural Networks (GNNs) and attention mechanisms to explicitly model spatial and topological dependencies.

A substantial body of literature has demonstrated the efficacy of GNNs in extracting local interaction features

and enabling representational synergy. For instance, GNN-parameterized policies and centralized critics have been deployed to mitigate non-stationarity in large-scale multi-robot coordination [19], while various attention-enhanced architectures successfully perform dynamic neighborhood aggregation and edge-importance weighting in complex cooperative tasks [20], [21], [22], [23], [24]. This structural paradigm has been widely extended to heterogeneous graphs [25], distributed traffic engineering [26], and edge resource allocation [27]. Furthermore, advanced representation frameworks have continuously pushed the boundaries of perceptual accuracy. Recent breakthroughs incorporate causal decoupling to filter confounding features [28], utilize contrastive learning for dynamic role disentanglement [16], and employ Inverse Attention/Theory of Mind (ToM) to infer teammate mental states in ad-hoc teaming [7]. Similarly, foundational spatiotemporal denoising mechanisms [29], [30], [31] and Set Transformers [32] have theoretically mitigated the “curse of many agents” by guaranteeing permutation invariance. Recent works have even optimized subgraph searching [33] or random probabilistic graph embeddings [34], while others [35] formally analyzed the 1-WL expressiveness limits of GNN communications, emphasizing the need for symmetry-breaking mechanisms. These prior structural foundations have indisputably solved the *passive observation* problem.

The Theoretical Gap: However, *representation does not equate to exploration*. A critical theoretical vulnerability unites all the aforementioned architectures: they exclusively function as passive state denoisers or structural filters. When deployed in exponentially expanding continuous kinematic manifolds, these sophisticated networks still inherently rely on primitive, undirected exploration mechanisms. Whether utilizing heuristic ϵ -greedy sampling [16], [25], [28], [33], [36], maximum entropy regularization (SAC) [19], [20], [21], stochastic PPO action sampling [7], [23], [26], [27], or parameter space noise [34], the fundamental exploration paradigm remains relegated to unstructured random trial-and-error. An attention mask merely dictates *what to observe in the current topology*, but fails to instruct the agent on *where to proactively explore in the unknown topology*. This critical deficiency necessitates a paradigm shift from passive structural representation to active, structure-aware topological exploration, serving as the foundational motivation for the TPE-MARL framework.

B. Information-Theoretic Exploration in MARL

To overcome the inefficiency of random perturbations, intrinsic motivation has become the de facto standard for exploration in sparse-reward environments. Early paradigms largely relied on state visitation pseudo-counts or curiosity-driven prediction errors to encourage the traversal of novel states [10], [11], [37]. However, as theoretical studies on the geometry of neural reinforcement learning indicate, policies in continuous environments strictly evolve on low-dimensional manifolds [38]. In exponentially expanding continuous state-action spaces, attempting to uniformly measure prediction errors or episodic temporal distances [12] without structural dimensionality reduction becomes computationally intractable.

Furthermore, non-stationary intrinsic bonuses inevitably violate the Markov property, leading to biased objectives or learning collapse unless rigorously regularized via state augmentation [39], [40]. While prior efforts have attempted to guide exploration by projecting into restricted discrete subspaces [9], providing subgoals from replay buffers [8], or utilizing graph Betti numbers for static grid traversals [41], they fail to encapsulate the continuous dynamic topology required for highly coupled multi-agent interactions.

The Trap of Undirected Diversity: Recently, the frontier of MARL exploration has been dominated by Information-Theoretic approaches, most notably Mutual Information (MI) maximization and Contrastive Learning. A prevailing methodology aims to maximize the mutual information between an agent’s identity and its trajectory or skills, explicitly forcing agents to exhibit highly heterogeneous behaviors. Representative frameworks such as CDS [13], ACORM [16], LIIR [17], and CTR [42] rely on InfoNCE bounds or contrastive roles to push agent representations apart, while others combine influence and curiosity heuristically [43] or filter visual noise via Multi-View Information Bottlenecks [44].

This “diversity-seeking” paradigm, however, suffers from a fatal theoretical flaw when applied to tightly coupled cyber-physical games (e.g., autonomous driving). Maximizing mutual information to indiscriminately distance agent behaviors constitutes *undirected diversity*. As game-theoretic analyses reveal, completely unaligned or purely autotelic exploration in multi-agent settings typically collapses into exploitable, highly unequal, or unsafe equilibria [45], [46]. In dense traffic, an agent actively pursuing behavioral heterogeneity solely for the sake of “being different” will inevitably execute dangerous, non-cooperative maneuvers (e.g., sudden braking or aggressive cut-ins). Recognizing this, some works have attempted to constrain MI maximization using value-aware weighting (RACE) [47], team-agnostic marginalizations (MIPI) [48], or retrospective successor features [15]. Conversely, other extremes attempt to completely suppress exploration-induced non-stationarity by forcing opponents into singular fixed points (SILI) [49] or demanding absolute expectation alignment (ELIGN) [50]. Yet, scalar value weightings and rigid predictability constraints still fail to capture the true geometric structure of the joint strategy space.

The Theoretical Gap: The fundamental missing link is *structure-aware topological guidance*. Rather than maximizing blind behavioral divergence, exploration must be bounded by the underlying interaction topology. Recent isolated insights in single-agent meta-RL and communication protocols have demonstrated that the Information Bottleneck (IB) principle [51], [52] can efficiently decouple task-critical states from environmental noise [53], [54], [55], and that structural entropy provides far superior guidance compared to flat Shannon entropy [14], [56]. Inspired by these principles, but crucially diverging from heuristic reward distillation [57] or adaptive scalar shaping [58], TPE-MARL introduces the discrete *topological quotient space* as the ultimate multi-agent Information Bottleneck. By optimizing a tractable Variational Evidence Lower Bound (ELBO) within these topological equivalence classes, our approach guarantees that mutual information max-

imization is strictly constrained to discovering deep, globally optimal coordinations, permanently resolving the catastrophe of undirected diversity in continuous MARL.

C. Robustness and Generalization in Physical MARL

Beyond the theoretical challenges of exploration, deploying MARL in physical cyber-physical systems exposes policies to severe covariate shifts and unmodeled non-stationary transition kernels [59]. The classical “Sim-to-Real” paradigm typically addresses these discrepancies through domain randomization or post-hoc safety constraints. For instance, in real-world traffic smoothing deployments, encountering Out-of-Distribution (OOD) behaviors—such as unmodeled vehicle cut-ins—often forces the system to rely on post-training controller modifications or rigid heuristic fallbacks [59]. These safety layers inherently override the RL policy, sacrificing cooperative efficiency for conservative survival [59]. Similarly, bridging the dynamic gap in robotic interactions frequently requires low-level signal approximations (e.g., zero-order holds) or adversarial probing agents to handle physical discrepancies [60]. Furthermore, comprehensive multi-agent exploration benchmarks reveal that even state-of-the-art algorithms remain highly brittle, with their efficacy heavily dictated by environmental scale and obstacle density [3].

To formally address these robustness issues, recent machine learning advancements have attempted to structurally decouple environmental noise from decision-making representations. In communication-constrained MARL, Dual Mutual Information Estimators (Du-MIE) have been proposed to explicitly maximize the impact of lossless messages while minimizing the interference of lossy noise bounds [55]. In the broader context of OOD generalization, context meta-RL frameworks utilize non-parametric density estimation to structure latent variables for OOD task enhancement [61]. Concurrently, causal RL aims to disentangle invariant causal factors from volatile environmental dynamics [62].

The Theoretical Gap: While these methods mathematically isolate continuous latent factors, they still operate entirely within the continuous physical or latent manifold. Consequently, as demonstrated in recent causal RL studies, strict zero-shot OOD generalization typically fails when the test environment introduces entirely unseen causal topologies or topological constraints, forcing models to rely on subsequent few-shot adaptation [62]. The fundamental limitation is that continuous representations remain overly sensitive to high-frequency physical perturbations (e.g., actuator latency, localization jitter).

TPE-MARL fundamentally circumvents this vulnerability by establishing *topological homomorphism*. Rather than attempting to disentangle continuous causal factors or applying post-hoc conservative safety heuristics, our framework maps the physical state into a discrete topological quotient space. This SimHash-based projection inherently functions as a structural Information Bottleneck, naturally absorbing high-frequency physical covariate shifts into invariant topological equivalence classes. This topological immunity is precisely what empowers TPE-MARL to seamlessly execute highly ag-

gressive, zero-shot dynamic negotiations (such as the physical zipper-merge) in real-world deployments.

III. PROBLEM FORMULATION

A. Dec-POMDP in Dynamic and Open Environments

The multi-vehicle cooperative decision-making problem in a continuous traffic flow fundamentally differs from conventional multi-agent reinforcement learning (MARL) tasks due to its open boundaries and mixed-autonomy nature. To formally capture these characteristics, we model the system as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) equipped with a time-varying agent set. The problem is defined by the topological tuple:

$$\mathcal{G} \triangleq \langle \{\mathcal{V}_t\}_{t \geq 0}, \mathcal{S}, \{\mathcal{A}_t\}_{t \geq 0}, \mathcal{P}, \mathcal{R}, \Omega, \mathcal{O}, \gamma \rangle$$

where each component is formulated to address the dynamic nature of the environment.

Open boundaries and time-varying active sets. Unlike standard Dec-POMDPs that assume a fixed and static set of interacting agents, the coordination zone in our problem is an open spatial domain where vehicles stochastically enter and exit. Let $\mathbb{V}$ denote the universal set of all potential entities. The active vehicle set at discrete time step $t \in \mathbb{N}_0$ is a time-varying subset $\mathcal{V}_t \subset \mathbb{V}$. This active set is strictly partitioned into two disjoint subsets: the controllable Connected and Automated Vehicles (CAVs) denoted by $\mathcal{N}_t$, and the Human-Driven Vehicles (HDVs) denoted by $\mathcal{H}_t$, satisfying $\mathcal{V}_t = \mathcal{N}_t \cup \mathcal{H}_t$ and $\mathcal{N}_t \cap \mathcal{H}_t = \emptyset$. The cardinality $N_t \triangleq |\mathcal{N}_t|$ denotes the instantaneous number of controllable CAVs, which fundamentally dictates the dimensionality of the joint action space.

Non-stationarity via mixed autonomy and latent intents. Let $\mathcal{S}$ denote the global state space and $\mathcal{A}_t$ denote the joint action space of the active agent set $\mathcal{N}_t$. The state transition kernel $\mathcal{P} : \mathcal{S} \times \mathcal{A}_t \times \mathcal{S} \rightarrow [0, 1]$ encapsulates a profound source of environmental non-stationarity driven by the mixed autonomy of the traffic flow. Specifically, the environment comprises both CAVs, which are governed by the learned coordination policy, and HDVs, whose behaviors are inherently stochastic and strictly non-controllable.

Instead of a deterministic progression, the unobservable internal intentions, varying reaction times, and bounded rationality of HDVs act as confounding stochastic processes that dictate the environment transition. We formalize this by introducing a latent stochastic variable $\mathbf{z}_t^{(j)} \sim P_{\mathcal{Z}}(\cdot)$ for each HDV $j \in \mathcal{H}_t$. The joint action of HDVs, denoted as $\mathbf{u}_t \in \mathcal{U}_t$, is sampled from an unknown, time-varying behavioral distribution π_{HDV} conditioned on these latent variables. The true state transition kernel is thus derived by marginalizing over the HDV action space:

$$\begin{aligned} \mathcal{P}(s_{t+1} \mid s_t, \mathbf{a}_t) &= \int_{\mathcal{U}_t} \mathcal{T}(s_{t+1} \mid s_t, \mathbf{a}_t, \mathbf{u}_t) \\ &\times \prod_{j \in \mathcal{H}_t} \pi_{\text{HDV}}^{(j)}(u_t^{(j)} \mid s_t, \mathbf{z}_t^{(j)}) d\mathbf{u}_t \end{aligned} \quad (1)$$

where $\mathcal{T}(\cdot)$ is the deterministic kinematic transition function. This integral representation explicitly demonstrates why the

Markov game is highly dynamic, as the transition probabilities constantly shift according to the time-varying spatial distribution and composition of the heterogeneous traffic flow.

Partial observability and objective. The decision-making process is further constrained by partial observability. Let $\mathcal{O} : \mathcal{S} \times \mathcal{N}_t \rightarrow \Omega$ denote the observation emission function, mapping the global state to a joint observation space based on the limited sensor perception range of each individual CAV. The joint reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A}_t \rightarrow \mathbb{R}$ evaluates the collective performance, and $\gamma \in [0, 1)$ is the discount factor.

The primary mathematical challenge posed by this formulation lies not only in the stochastic state transitions caused by the mixed autonomy but also in the severe dimensionality explosion associated with the time-varying joint action space $\{\mathcal{A}_t\}_{t \geq 0}$, which will be formalized in the subsequent subsection.

B. Information Spaces and the Curse of Dimensionality

State manifold and ego-centric observations. The global state $s_t \in \mathcal{S}$ aggregates the physical and kinematic properties of all active entities $\mathcal{M}_t = \mathcal{N}_t \cup \mathcal{H}_t$ within the system. Each entity $j \in \mathcal{M}_t$ is represented by a high-dimensional feature vector $\mathbf{x}_t^{(j)} \in \mathbb{R}^8$, which encapsulates spatial coordinates, velocities, accelerations, and latent driving intentions. However, due to the strict partial observability constraint inherent to the Dec-POMDP, each CAV i cannot access this complete global state manifold.

Instead, the decision-making process relies on a localized, ego-centric projection $\mathbf{o}_t^{(i)} \in \Omega$. To balance perceptual completeness with computational tractability, we truncate the perceptual field to a topological neighborhood $\mathcal{N}_K(i)$, comprising up to $K = 6$ most strategically relevant adjacent vehicles. The local observation is mathematically formalized as a concatenation of the ego-state $\mathbf{x}_{t,\text{ego}}^{(i)} \in \mathbb{R}^{d_e}$ and a set of relational feature vectors $\Delta \mathbf{x}_t^{(ij)} \in \mathbb{R}^4$. These relational vectors encode the spatial and kinematic asymmetries between the ego agent i and its neighbor j :

$$\mathbf{o}_t^{(i)} \triangleq \left(\mathbf{x}_{t,\text{ego}}^{(i)}, \left\{ \Delta \mathbf{x}_t^{(ij)} \right\}_{j \in \mathcal{N}_K(i)} \right) \in \Omega \subseteq \mathbb{R}^{d_o}, \quad (2)$$

where d_o denotes the total dimensionality of the observation space.

The combinatorial explosion of joint actions. The crux of the multi-agent coordination challenge emerges from the temporal evolution of the decision space. We formalize the control space for each CAV as a decoupled discrete action set spanning both lateral ($\mathcal{A}_{\text{lat}}$) and longitudinal ($\mathcal{A}_{\text{long}}$) dimensions. With three lateral maneuvers (left change, lane keeping, right change) and three longitudinal commands (accelerate, maintain, decelerate), the individual action space is the Cartesian product $\mathcal{A}^{(i)} = \mathcal{A}_{\text{lat}} \times \mathcal{A}_{\text{long}}$, yielding a constant cardinality of $|\mathcal{A}^{(i)}| = 9$.

At any given time step t , the joint action space $\mathcal{A}_t$ is constructed as the Cartesian product of the individual action spaces over the active controllable agent set:

$$\mathcal{A}_t \triangleq \prod_{i \in \mathcal{N}_t} \mathcal{A}^{(i)} \implies |\mathcal{A}_t| = |\mathcal{A}^{(i)}|^{N_t} = 9^{N_t}. \quad (3)$$

This exponential scaling induces a severe “curse of dimensionality”. As the density of the multi-agent system increases, the vastness of this joint space systematically confounds undirected exploration mechanisms and renders the exhaustive evaluation of joint Q-values computationally intractable. Solving this fundamental mathematical bottleneck—learning coordinated joint policies without being paralyzed by the exponentially growing search space—necessitates the topological dimensionality reduction methodology proposed in Section IV.

C. Centralized Objective and Extrinsic Reward

Centralized training with decentralized execution. We formulate the decision-making paradigm within the Centralized Training with Decentralized Execution (CTDE) framework. During the training phase, a central decision unit aggregates the joint observation $\mathbf{o}_t = \{\mathbf{o}_t^{(i)}\}_{i \in \mathcal{N}_t}$ and accesses the global state s_t to learn a parameterized joint policy π_θ . During execution, each CAV operates independently based strictly on its local observation $\mathbf{o}_t^{(i)}$, satisfying the communication and latency constraints required for real-world deployment.

Extrinsic reward and optimal policy invariance. Under the high-frequency control regime ($\Delta t = 100$ ms), the system inherently suffers from a severe sparse gradient problem, as the macroscopic state evolution resulting from high-frequency discrete actions is easily masked by temporal lag and observation noise. To mitigate this issue without altering the underlying Nash equilibrium of the Dec-POMDP, we adopt the Hybrid Differential Reward (HDR) framework established in prior work [63], [64] as the extrinsic environment reward $r_t = \mathcal{R}(s_t, \mathbf{a}_t)$. This framework fundamentally reshapes the sparse task reward $\mathcal{R}_{\text{task}}$ by incorporating a state-potential function $\Phi : \mathcal{S} \rightarrow \mathbb{R}$:

$$r_t \triangleq \mathcal{R}_{\text{task}}(s_t, \mathbf{a}_t) + \gamma \Phi(s_{t+1}) - \Phi(s_t). \quad (4)$$

This formulation inherently synthesizes a temporal reward difference signal and an action-reward gradient signal, theoretically ensuring optimal policy invariance [65] while providing a dense causal signal for policy optimization.

Unified optimization formulation. The ultimate objective is to find optimal parameters θ^* that maximize the expected cumulative discounted reward $J(\theta)$ over the decision horizon T . Let $\tau \triangleq (s_0, \mathbf{a}_0, r_0, s_1, \dots)$ denote a trajectory generated by the joint policy. The multi-vehicle coordination problem is thus rigorously cast as the following constrained stochastic optimization problem:

$$\begin{aligned} \max_{\theta} \quad & J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right], \\ \text{s.t.} \quad & \mathcal{N}_t = \mathcal{N}_t^{\text{CAV}} \cup \mathcal{N}_t^{\text{HDV}}, \quad \forall t \in [0, T-1], \\ & \mathcal{N}_{t+1} \sim \mathcal{F}_{\text{flow}}(\mathcal{N}_t), \quad \forall t \in [0, T-1], \\ & s_{t+1} \sim \mathcal{P}(\cdot | s_t, \mathbf{a}_t), \quad \forall t \in [0, T-1], \\ & \mathbf{o}_t^{(i)} = \mathcal{O}(s_t, i), \quad \forall i \in \mathcal{N}_t, \\ & \mathbf{a}_t = \pi_\theta(\mathbf{o}_t), \quad \mathbf{a}_t \in \mathcal{A}_t, \\ & a_t^{(i)} \sim \pi_\theta^{(i)}(\cdot | \mathbf{o}_t^{(i)}), \quad \forall i \in \mathcal{N}_t. \end{aligned} \quad (5)$$

This formulation rigorously encapsulates the compounding mathematical challenges of the task: optimizing a globally coherent policy under partial observability, non-stationary stochastic traffic flow, and an exponentially exploding joint action space. Direct optimization of this objective via conventional MARL techniques is mathematically intractable due to the curse of dimensionality. Consequently, discovering a low-dimensional structural representation of the multi-agent interactions is not merely beneficial, but strictly necessary. This mathematical imperative directly motivates the game topology construction and the dual intrinsic reward mechanism presented in Section IV.

IV. METHODOLOGY

Section III established that the multi-vehicle coordination problem is governed by a Dec-POMDP with a time-varying agent set, where the joint action space grows exponentially with the number of agents ($|\mathcal{A}_t| = 9^{N_t}$), and the high-frequency control regime ($\Delta t = 100$ ms) induces a sparse gradient problem in reward design. To address these challenges, we now introduce the *Topology-enhanced Multi-Agent Reinforcement Learning* (TPE-MARL) framework. The core philosophy of this framework is to exploit the structural redundancy inherent in cooperative driving: while the raw state space is continuous and high-dimensional, the underlying game-theoretic interactions between vehicles admit a compact discrete representation. By constructing this representation and using it to guide exploration, the effective search space is dramatically reduced.

The methodology proceeds in three stages. First, Section IV-A develops a *Game Topology Tensor* that compresses the continuous physical state into a structured, low-dimensional topological representation while preserving policy-critical equivalence relations. Second, Section IV-B introduces a dual intrinsic reward mechanism that leverages this topological representation to guide exploration: a *novelty reward* based on visit counts in the topological space encourages breadth of exploration, while a *collaboration reward* based on conditional mutual information encourages depth of exploration in strategically relevant interactions. Third, Section IV-C presents a variational inference framework for estimating the mutual information terms required by the collaboration reward, and Section IV-D consolidates all components into the unified TPE-MARL training algorithm.

A. Game Topology Tensor Construction

The central observation motivating this methodology is that the vast majority of continuous physical states encountered during multi-agent coordination are *game-theoretically redundant*. While the raw state space $\mathcal{S}_e$ is a continuous, high-dimensional manifold, the underlying game-theoretic interactions—the mutual influence relationships defined by physical constraints and right-of-way—admit a compact discrete representation. By mapping the continuous physical states to a structured, low-dimensional topological representation, the effective search space for policy optimization can be dramatically compressed.

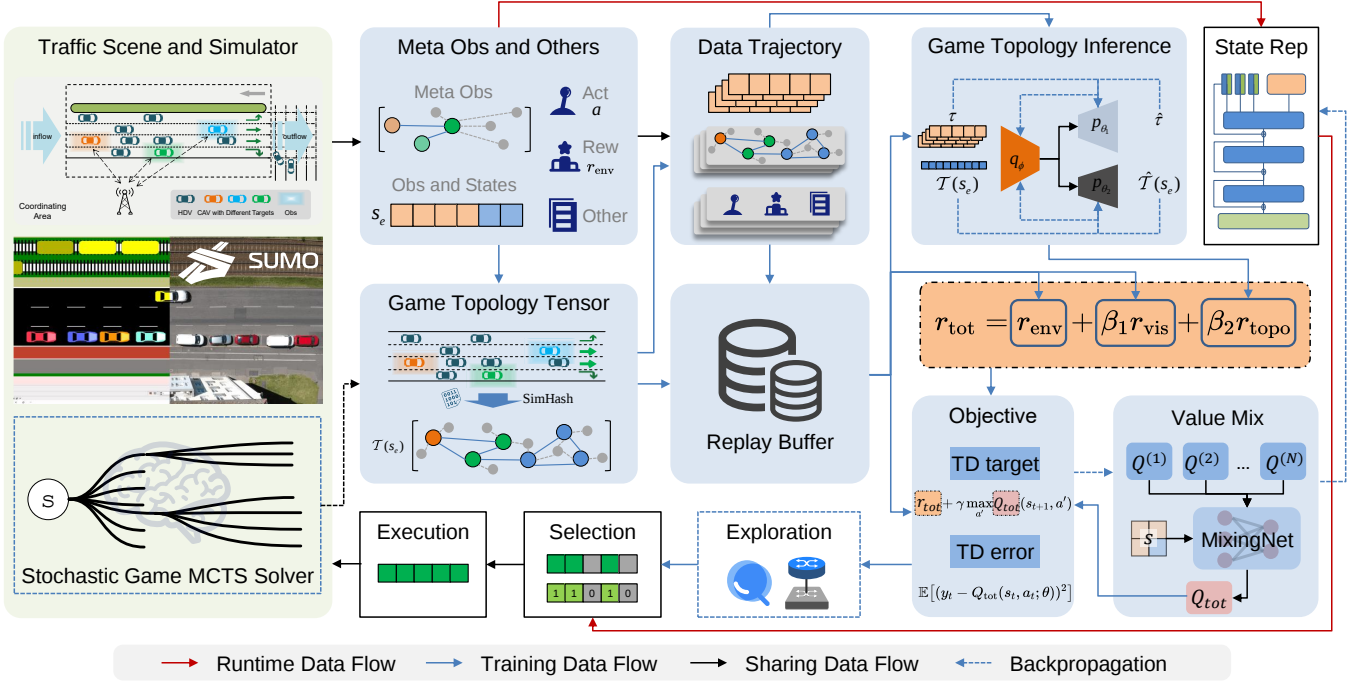


Fig. 2: TPE-MARL framework

This section formalizes this dimension reduction through the construction of a *Game Topology Tensor*. The process proceeds in three stages: (i) abstracting the raw sensory data into a domain-agnostic *meta-observation*; (ii) encoding the pairwise geometric and kinematic asymmetries into game relation vectors via Locality-Sensitive Hashing (LSH); and (iii) aggregating these relations into a global topology tensor that maps the continuous state manifold into a discrete quotient space, effectively acting as an information bottleneck that preserves policy-critical information.

1) *Meta-Observation and Game Relation Vector*: The initial step transforms the raw, high-dimensional local observation $\mathbf{o}_t^{(i)}$ (defined in Section III-B) into a compact relational representation. We define a *meta-observation* vector $\mathbf{o}_{t,\text{meta}}^{(i)} \in \mathbb{R}^{d_m}$ for each CAV i , which encapsulates only the feature subspace relevant to game-theoretic interactions.

To ensure the generalizability of the representation, we abstract the domain-specific traffic variables into a unified multi-agent feature space. The meta-observation is partitioned into ego-centric information $\mathbf{s}_{t,\text{ego}}^{(i)}$ and neighbor information $\mathbf{s}_{t,\text{nbr}}^{(i)}$. The ego information is formulated as:

$$\mathbf{s}_{t,\text{ego}}^{(i)} \triangleq [\mathbf{k}_t^{(i)}, \mathbf{b}_t^{(i)}, \mathbf{e}_{\text{intent}}^{(i)}]^\top, \quad (6)$$

where $\mathbf{k}_t^{(i)}$ denotes the *kinematic manifold* (encapsulating normalized spatial coordinates and velocities), $\mathbf{b}_t^{(i)}$ represents the *local spatial boundary constraints* (e.g., time headways to surrounding topological boundaries), and $\mathbf{e}_{\text{intent}}^{(i)}$ is a *semantic intent embedding* encoding the agent's categorical properties (such as agent type and navigation targets).

The neighbor information $\mathbf{s}_{t,\text{nbr}}^{(i)}$ captures the relative dynamics of the K most strategically relevant entities within

a truncated perception field $\mathbf{c}^{(i)}$:

$$\mathbf{s}_{t,\text{nbr}}^{(i)} \triangleq \text{Concat}_{j \in \mathbf{c}^{(i)}} [\Delta \mathbf{k}_t^{(ij)}, \mathbf{e}_{\text{intent}}^{(j)}]^\top, \quad (7)$$

where $\Delta \mathbf{k}_t^{(ij)} = \mathbf{k}_t^{(j)} - \mathbf{k}_t^{(i)}$ measures the kinematic asymmetry between agent i and neighbor j . The complete meta-observation is the concatenation $\mathbf{o}_{t,\text{meta}}^{(i)} = [\mathbf{s}_{t,\text{ego}}^{(i)}, \mathbf{s}_{t,\text{nbr}}^{(i)}] \in \mathbb{R}^{d_m}$.

To capture the pairwise interaction structure, we define a *game relation vector* $\mathbf{D}_t^{ij}$ between any two CAVs i and j . This vector decomposes the difference in their meta-observations into a continuous magnitude and a discretized angular direction:

$$\mathbf{D}_t^{ij} \triangleq \left(\left\| \mathbf{o}_{t,\text{meta}}^{(i)} - \mathbf{o}_{t,\text{meta}}^{(j)} \right\|_2, \langle \mathbf{o}_{t,\text{meta}}^{(i)} - \mathbf{o}_{t,\text{meta}}^{(j)} \rangle \right) \in \mathbb{R} \times \mathbb{Z}. \quad (8)$$

The Euclidean norm quantifies the discrepancy in their perceptual states, while the angular mapping operator $\langle \cdot \rangle : \mathbb{R}^{d_m} \rightarrow \mathbb{Z}$ discretizes the structural type of this asymmetry into a finite categorical set via SimHash, as detailed below.

2) *SimHash-Based Topology Encoding*: The angular mapping operator $\langle \cdot \rangle$ requires a mechanism that maps directionally similar continuous vectors to identical discrete codes, fundamentally preserving the local neighborhood structure of the original metric space. We employ the SimHash algorithm [66], an LSH technique.

By generating a random projection matrix $\mathbf{V}_{\text{hash}} \in \mathbb{R}^{m \times d_m}$, where each row is sampled uniformly from the unit hypersphere, the continuous meta-observation difference is projected

into an m -bit binary hash code:

$$\mathbf{h}_t^{ij} \triangleq \text{sign} \left(\mathbf{V}_{\text{hash}} \cdot \frac{\mathbf{o}_{t,\text{meta}}^{(i)} - \mathbf{o}_{t,\text{meta}}^{(j)}}{\|\mathbf{o}_{t,\text{meta}}^{(i)} - \mathbf{o}_{t,\text{meta}}^{(j)}\|_2} \right) \in \{0, 1\}^m. \quad (9)$$

The theoretical foundation of this encoding is the *angle preservation property*: the probability of a bit collision directly correlates with the cosine similarity of the vectors. The binary vector $\mathbf{h}_t^{ij}$ is subsequently mapped to a decimal integer $h_{t,\text{int}}^{ij} \in \{0, 1, \dots, 2^m - 1\}$. The game relation vector is thus operationalized as $\mathbf{D}_t^{ij} = (d_{t,\text{mag}}^{ij}, h_{t,\text{int}}^{ij})$.

3) *Local Topology and Quotient Space Mapping*: For each CAV i , the *local topology vector* $\mathcal{T}_t^{(i)}$ aggregates the game relation vectors within its attention set $\mathbf{c}^{(i)}$:

$$\mathcal{T}_t^{(i)} \triangleq (\mathbf{D}_t^{ij_0}, \mathbf{D}_t^{ij_1}, \dots, \mathbf{D}_t^{ij_{k-1}}). \quad (10)$$

The global *Game Topology Tensor* $\mathcal{T}(\mathbf{s}_e)$ is the ordered collection of all local topology vectors across the active CAV set:

$$\mathcal{T}(\mathbf{s}_e) \triangleq (\mathcal{T}_t^{(0)}, \mathcal{T}_t^{(1)}, \dots, \mathcal{T}_t^{(N_t-1)}). \quad (11)$$

Crucially, the construction of $\mathcal{T}(\cdot)$ is not merely a feature engineering heuristic; it constitutes a surjective mapping from the continuous exploration state manifold $\mathcal{S}_e$ to a compact topological space, serving as the mathematical foundation for equivalence preservation.

4) Equivalence Preservation and Information Bottleneck:

To rigorously justify the dimensionality reduction, we formalize the mapping $\mathcal{T}(\cdot)$ through the lens of equivalence relations and representation learning. We define a binary relation $\sim_{\mathcal{T}}$ on the exploration state space $\mathcal{S}_e$:

$$\mathbf{s}_1 \sim_{\mathcal{T}} \mathbf{s}_2 \iff \mathcal{T}(\mathbf{s}_1) = \mathcal{T}(\mathbf{s}_2), \quad \forall \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}_e. \quad (12)$$

Lemma 1 (Game Topology Equivalence and Quotient Space). *The relation $\sim_{\mathcal{T}}$ is a strict equivalence relation. It partitions the continuous state manifold $\mathcal{S}_e$ into a set of disjoint equivalence classes, inducing a discrete quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$.*

Proof. The reflexivity ($\mathcal{T}(\mathbf{s}) = \mathcal{T}(\mathbf{s})$), symmetry ($\mathcal{T}(\mathbf{s}_1) = \mathcal{T}(\mathbf{s}_2) \Rightarrow \mathcal{T}(\mathbf{s}_2) = \mathcal{T}(\mathbf{s}_1)$), and transitivity ($\mathcal{T}(\mathbf{s}_1) = \mathcal{T}(\mathbf{s}_2) \wedge \mathcal{T}(\mathbf{s}_2) = \mathcal{T}(\mathbf{s}_3) \Rightarrow \mathcal{T}(\mathbf{s}_1) = \mathcal{T}(\mathbf{s}_3)$) naturally hold, formally establishing $\sim_{\mathcal{T}}$ as an equivalence relation. $\square$

The theoretical significance of the quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$ relies on its ability to preserve the optimal decision structure while discarding redundant continuous variations. We formulate this fundamental property as follows:

Conjecture 1 (Optimal Policy Invariance via Information Bottleneck). *The topology tensor $\mathcal{T}(\mathbf{s})$ acts as an optimal Information Bottleneck [67]. It minimizes the mutual information with the raw state manifold $\mathcal{I}(\mathcal{T}(\mathbf{s}); \mathbf{s})$ to discard strategically redundant physical noise, while maximizing the mutual information $\mathcal{I}(\mathcal{T}(\mathbf{s}); Q^*)$ with the optimal action-value distribution. Consequently, for any $\mathbf{s}_1 \sim_{\mathcal{T}} \mathbf{s}_2$, the tensor acts as a minimal sufficient statistic for the optimal joint policy, leading to:*

$$Q^*(\mathbf{s}_1, \mathbf{a}) \approx Q^*(\mathbf{s}_2, \mathbf{a}), \quad \forall \mathbf{a} \in \mathcal{A}_t. \quad (13)$$

Empirical Corroboration of Policy Invariance. While Equation (13) involves a lossy compression and cannot be established as a strict equality in continuous limits, we provide rigorous empirical validation to substantiate this information-theoretic conjecture. We sampled 28,743 pairwise vehicle interaction instances from the simulation environment and computed their respective SimHash codes. For each state, we derived the ground-truth optimal action-value distribution $\pi^*(\cdot | \mathbf{s}) = \text{softmax}(Q^*(\mathbf{s}, \cdot) / \tau)$ utilizing an exact Monte Carlo Tree Search (MCTS) oracle [68] with full state access.

To quantify the intra-group policy discrepancy, we evaluated the Jensen-Shannon Divergence (JSD) between the optimal policy distributions of any two states $\mathbf{s}_1, \mathbf{s}_2$ mapped to the same topological equivalence class (i.e., $\mathbf{s}_1 \sim_{\mathcal{T}} \mathbf{s}_2$). As illustrated in Figure 3, the statistical distribution reveals that over 92% of the topologically equivalent states exhibit an intra-group JSD below 0.3. This minimal divergence confirms that states sharing the same Game Topology Tensor reliably share near-identical optimal Q-value landscapes, effectively preserving the sufficient statistic of the game-theoretic structure.

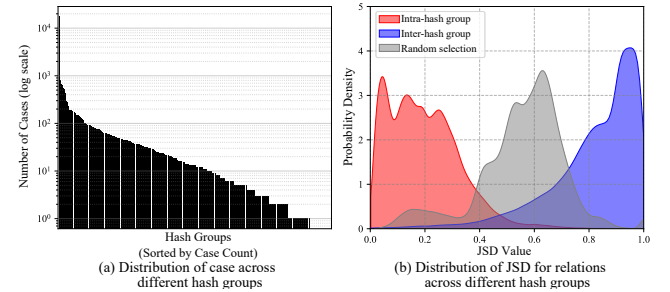


Fig. 3: Distribution of Jensen-Shannon Divergence (JSD) among optimal policy distributions within identical topological equivalence classes. The heavy concentration below JSD=0.3 corroborates the optimal policy invariance conjecture.

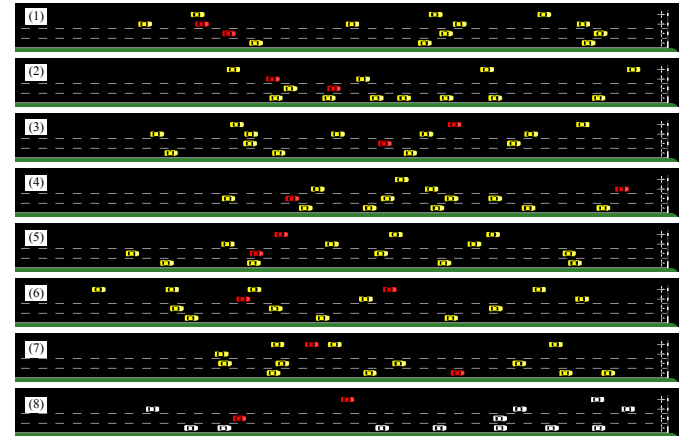


Fig. 4: Eight physically distinct traffic scenarios mapped to the identical Game Topology Tensor. The quotient space representation filters out absolute spatial variations while strictly preserving the underlying multi-agent interaction geometry.

To visually demystify this equivalence preservation, Figure 4 presents eight physically distinct traffic scenarios that

are structurally aggregated into the identical SimHash group. Despite substantial variations in absolute spatial coordinates, lane assignments, and neighbor identities, the intrinsic game-theoretic topography—e.g., the ego vehicle is adjacent to a slower leading entity with a feasible cooperative lane-changing path—remains strictly invariant. Consequently, the optimal cooperative strategy dictating which agent should yield or accelerate is identical across all eight scenarios.

By formalizing the problem within the discrete quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$, we effectively circumvent the curse of dimensionality. This mathematical transformation renders count-based exploration tractable and establishes a foundational structure for the intrinsic reward mechanisms developed in the subsequent sections.

B. Dual Intrinsic Reward Mechanism

The Game Topology Tensor constructed in Section IV-A furnishes a discrete quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$ that encapsulates the multi-agent interaction structure. However, this compact representation requires a principled mechanism to translate the structural information into actionable exploration incentives. Conventional undirected exploration strategies (e.g., ϵ -greedy) are sample-inefficient, while raw-state count-based methods succumb to the curse of dimensionality.

To overcome these limitations, we introduce a *dual intrinsic reward mechanism* grounded in information theory. This mechanism guides exploration via two complementary optimization objectives: a *topology novelty reward* (Section IV-B2) that maximizes the marginal entropy of the visited topological states (exploration breadth), and a *collaboration exploration reward* (Section IV-B1) that maximizes the conditional mutual information between interacting agents (exploration depth). Section IV-B3 unifies these components into a coherent information-seeking objective.

1) *Collaboration Exploration Reward via Mutual Information*: In partially observable Markov games, the inherent uncertainty regarding the unobservable intentions of neighboring agents fundamentally limits coordination. Reducing this uncertainty necessitates actively acquiring information about neighbors' behavioral states. We formulate this "depth-oriented" exploration through the lens of mutual information: agents should be intrinsically rewarded for interactions that yield high predictive information gain about their future topological surroundings.

Information gain via conditional mutual information. For any interacting pair (i, j) , we quantify the information gain agent j provides to agent i by the conditional mutual information between i 's future local topology $\mathcal{T}_{t+1}^{(i)}$ and j 's latent behavioral state $\mathbf{z}_t^{(j)}$:

$$\mathcal{G}_{\text{info},t}^{(i,j)} \triangleq \mathcal{I} \left(\mathcal{T}_{t+1}^{(i)}; \mathbf{z}_t^{(j)} \mid \tau_t^{\mathbf{c}^{i+}}, \mathbf{z}_t^{\mathbf{c}^{i+} \setminus \{j\}} \right), \quad (14)$$

where $\mathbf{c}^{i+} \triangleq \mathbf{c}^{(i)} \cup \{i\}$ denotes the extended attention set, $\tau_t^{\mathbf{c}^{i+}}$ aggregates their observation-action histories, and $\mathbf{z}_t^{(j)}$ is a compressed latent variable encoding j 's history (formally derived in Section IV-C). The conditioning ensures we isolate the *unique* marginal information contributed by j .

Proposition 1 (Entropy Reduction Equivalence). *Let $C_{ij} \triangleq (\tau_t^{\mathbf{c}^{i+}}, \mathbf{z}_t^{\mathbf{c}^{i+} \setminus \{j\}})$. The conditional mutual information precisely quantifies the reduction in conditional entropy of i 's future topology:*

$$\mathcal{I} \left(\mathcal{T}_{t+1}^{(i)}; \mathbf{z}_t^{(j)} \mid C_{ij} \right) = \mathcal{H} \left(\mathcal{T}_{t+1}^{(i)} \mid C_{ij} \right) - \mathcal{H} \left(\mathcal{T}_{t+1}^{(i)} \mid \mathbf{z}_t^{(j)}, C_{ij} \right). \quad (15)$$

Proof. This follows directly from the definition of conditional mutual information and the standard chain rule of Shannon entropy. $\square$

Tractable estimation via Variational Inference. Computing Equation (15) requires integrating over unknown high-dimensional transition dynamics. We derive a tractable surrogate by introducing a variational predictive distribution $p_{\theta_1}(\mathcal{T}_{t+1}^{(i)} \mid \mathbf{z}_t^{\mathbf{c}^{i+}})$, parameterized by a neural network decoder.

Proposition 2 (Variational Evidence Lower Bound). *By substituting the true posterior with the variational decoder p_{θ_1} , the conditional mutual information admits a strict Evidence Lower Bound (ELBO):*

$$\mathcal{I}(\mathcal{T}_{t+1}^{(i)}; \mathbf{z}_t^{(j)} \mid C_{ij}) \geq \mathbb{E}_{\mathbf{z} \sim q_{\phi}} \left[\log p_{\theta_1} \left(\mathcal{T}_{t+1}^{(i)} \mid \mathbf{z}_t^{\mathbf{c}^{i+}} \right) - \log p_{\theta_1} \left(\mathcal{T}_{t+1}^{(i)} \mid \mathbf{z}_t^{\text{partial}} \right) \right], \quad (16)$$

where $\mathbf{z}_t^{\text{partial}}$ is constructed by replacing the informative latent $\mathbf{z}_t^{(j)}$ with a random sample $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ from the prior.

Proof. Applying the non-negativity of the Kullback-Leibler divergence $D_{\text{KL}}(P \parallel Q) \geq 0$, the conditional entropy is bounded by the cross-entropy with the variational distribution p_{θ_1} . The first term represents the expected log-likelihood given full latent contexts, while the second term serves as a baseline log-likelihood with j 's specific information ablated. The difference forms the exact ELBO. $\square$

We deploy this ELBO as the pairwise collaboration estimator $\hat{\mathcal{I}}(\cdot)$. The global collaboration exploration reward $r_{\text{topo},t}$ is then aggregated across all active CAVs and their topological neighbors:

$$r_{\text{topo},t} \triangleq \frac{1}{N_t \cdot |\mathbf{c}^{i+}|} \sum_{i \in \mathcal{N}_t} \sum_{j \in \mathbf{c}^{i+}} \hat{\mathcal{I}} \left(\mathcal{T}_{t+1}^{(i)}; \mathbf{z}_t^{(j)} \mid C_{ij} \right). \quad (17)$$

2) *Topology Novelty Reward via Visit Count*: While the collaboration reward promotes *depth* by deepening the understanding of current interactions, it does not inherently prevent the policy from collapsing into a subset of highly predictable, yet globally sub-optimal, local topologies. Discovering superior coordination paradigms requires *breadth-oriented* exploration.

As established in Section IV-A4, the Game Topology Tensor projects the intractable continuous state manifold $\mathcal{S}_e$ onto a discrete quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$. This structural discretization is what theoretically legitimizes count-based exploration: the empirical visitation frequencies in the quotient space provide a valid density estimator for the marginal distribution of topological interactions.

By further discretizing the continuous magnitude d_{mag}^{ij} into $B = 20$ bins, we map the tensor to a fully discrete representation $\tilde{\mathcal{T}}_t$. We maintain a hash table $N(\tilde{\mathcal{T}}_t)$ to track the visitations of these quotient states. Following the Optimism in the Face of Uncertainty (OFU) principle [69], the topology novelty reward is defined as an Upper Confidence Bound (UCB) bonus:

$$r_{\text{vis},t} \triangleq \frac{1}{\sqrt{N(\tilde{\mathcal{T}}_t)}}. \quad (18)$$

This reward structure systematically drives the multi-agent system to traverse unexplored regions of the quotient space, maximizing the marginal entropy of the visited topologies $\mathcal{H}(\mathcal{T})$.

3) Synergy: A Unified Information-Theoretic Objective:

The two intrinsic reward signals are unified into a compound exploration objective:

$$r_{\text{tot},t} = r_{\text{env},t} + \beta_1 r_{\text{vis},t} + \beta_2 r_{\text{topo},t}, \quad (19)$$

where β_1 and β_2 govern the relative incentive strengths.

Crucially, this combination is not merely a heuristic aggregation, but represents a holistic information-theoretic objective for the exploration policy π_{explore} . The novelty reward acts as a proxy for maximizing the marginal state entropy $\mathcal{H}(\mathcal{T})$, encouraging the policy to diversify its topological experiences. Simultaneously, the collaboration reward maximizes the conditional mutual information $\mathcal{I}(\mathcal{T}_{t+1}; \mathbf{z}_t)$, encouraging the policy to seek informative and predictable multi-agent geometries within those topologies.

Together, the dual mechanism transforms exploration from an undirected random walk into a targeted optimization process that maximizes the joint information capacity of the multi-agent system, resolving the exploration-exploitation dilemma within the exponentially large action space.

C. Mutual Information Estimation via Variational Inference

The collaborative exploration reward defined in Section IV-B1 relies on estimating the conditional mutual information $\mathcal{I}(\mathcal{T}_{t+1}^{(i)}; \mathbf{z}_t^{(j)} \mid C_{ij})$. Proposition 2 established a tractable Evidence Lower Bound (ELBO) for this quantity via a variational decoder p_{θ_1} . However, a crucial prerequisite for this estimation is the extraction of the latent behavioral state $\mathbf{z}_t^{(j)}$.

To ensure the mutual information estimator strictly quantifies the *topology-predictive* information rather than redundant trajectory noise, we introduce an adversarial variational inference framework. This framework casts the latent state extraction as an Information Bottleneck (IB) optimization problem, actively decoupling the game-theoretic intents from raw kinematic variations.

1) *Topology Inference Network Architecture*: The topology inference network is mathematically structured as an adversarial Variational Autoencoder (VAE) [70]. It learns a probabilistic mapping from the temporal observation-action history to a compact latent space. Let $\boldsymbol{\tau}_t^{(i)} \triangleq (\mathbf{o}_0^{(i)}, \mathbf{a}_0^{(i)}, \dots, \mathbf{o}_t^{(i)})$ denote the trajectory sequence. The architecture comprises three distinct differentiable operators:

Temporal Sequence Encoder $q_\phi(\mathbf{z}_t^{(i)} \mid \boldsymbol{\tau}_t^{(i)})$. This operator models the posterior distribution of the latent state given the trajectory. We parameterize it using a recurrent sequence encoder to capture temporal dependencies, mapping $\boldsymbol{\tau}_t^{(i)}$ to the parameters of a multivariate Gaussian: $\mathbf{z}_t^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}_t^{(i)}, \text{diag}(\boldsymbol{\sigma}_t^{(i)})^2)$. The capacity of the latent space is constrained by a prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Topology Prediction Decoder $p_{\theta_1}(\mathcal{T}_{t+1}^{(i)} \mid \mathbf{z}_t^{i+})$. This is the core estimator for the ELBO (Equation (16)). It aggregates the latent states of all agents within the extended attention set $\mathbf{c}^{i+}$ to predict i 's future local topology. To ensure robustness against varying neighbor quantities and permutations, we apply a permutation-invariant set aggregation operator (e.g., mean-pooling over feature representations) before projecting to the topological space:

$$\hat{\mathcal{T}}_{t+1}^{(i)} \triangleq f_{\theta_1} \left(\frac{1}{|\mathbf{c}^{i+}|} \sum_{j \in \mathbf{c}^{i+}} \psi(\mathbf{z}_t^{(j)}) \right). \quad (20)$$

The predictive fidelity is supervised by a Mean Squared Error (MSE) loss, denoted as $\mathcal{L}_{\text{TP}}(\phi, \theta_1)$.

Adversarial Reconstruction Decoder $p_{\theta_2}(\boldsymbol{\tau}_t^{(i)} \mid \mathbf{z}_t^{(i)})$. Acting as an antagonist, this decoder attempts to reconstruct the raw trajectory $\boldsymbol{\tau}_t^{(i)}$ solely from the marginalized latent $\mathbf{z}_t^{(i)}$. The reconstruction loss is defined as $\mathcal{L}_{\text{RG}}(\phi, \theta_2) = \text{MSE}(\boldsymbol{\tau}_t^{(i)}, \hat{\boldsymbol{\tau}}_t^{(i)})$.

2) *Information Decoupling via Adversarial Information Bottleneck*: The fundamental requirement for our mutual information estimator is that $\mathbf{z}_t^{(i)}$ must act as a minimal sufficient statistic for predicting $\mathcal{T}_{t+1}^{(i)}$. If $\mathbf{z}_t^{(i)}$ acts as a lossless compression of $\boldsymbol{\tau}_t^{(i)}$, the pairwise mutual information evaluated in Equation (14) would trivially collapse into measuring raw trajectory correlations.

We enforce this selectivity by optimizing the Information Bottleneck (IB) Lagrangian [67]:

$$\min_{q(\mathbf{z}|\boldsymbol{\tau})} \mathcal{I}(\boldsymbol{\tau}; \mathbf{z}) - \beta \mathcal{I}(\mathbf{z}; \mathcal{T}), \quad (21)$$

where $\mathcal{I}(\mathbf{z}; \mathcal{T})$ is maximized by the predictive loss $\mathcal{L}_{\text{TP}}$. The critical challenge lies in explicitly minimizing $\mathcal{I}(\boldsymbol{\tau}; \mathbf{z})$. By definition, $\mathcal{I}(\boldsymbol{\tau}; \mathbf{z}) = \mathcal{H}(\boldsymbol{\tau}) - \mathcal{H}(\boldsymbol{\tau} \mid \mathbf{z})$. Since the marginal entropy of the trajectory $\mathcal{H}(\boldsymbol{\tau})$ is fixed given the environment, minimizing the mutual information is mathematically equivalent to *maximizing the conditional entropy* $\mathcal{H}(\boldsymbol{\tau} \mid \mathbf{z})$.

Under the assumption that the adversarial decoder p_{θ_2} formulates a Gaussian likelihood $p_{\theta_2}(\boldsymbol{\tau} \mid \mathbf{z}) \propto \exp(-\text{MSE}(\boldsymbol{\tau}, \hat{\boldsymbol{\tau}}))$, maximizing the conditional entropy corresponds exactly to *maximizing the reconstruction error* $\mathcal{L}_{\text{RG}}$. To ensure the optimization remains bounded and theoretically sound, the KL divergence $\mathcal{L}_{\text{KL}}(\phi) = D_{\text{KL}}(q_\phi \parallel p(\mathbf{z}))$ restricts the latent norm. The composite optimization objective for the topology network is thus rigorously defined as:

$$\mathcal{L}_{\text{TopoNet}}(\phi, \theta_1, \theta_2) \triangleq \mathcal{L}_{\text{TP}} + \mathcal{L}_{\text{KL}} - \lambda_{\text{GF}} \mathcal{L}_{\text{RG}}. \quad (22)$$

This adversarial dynamic is practically implemented using a Gradient Reversal Layer (GRL) [71]. During backpropagation, the gradients from $\mathcal{L}_{\text{RG}}$ are negated for the encoder q_ϕ . Con-

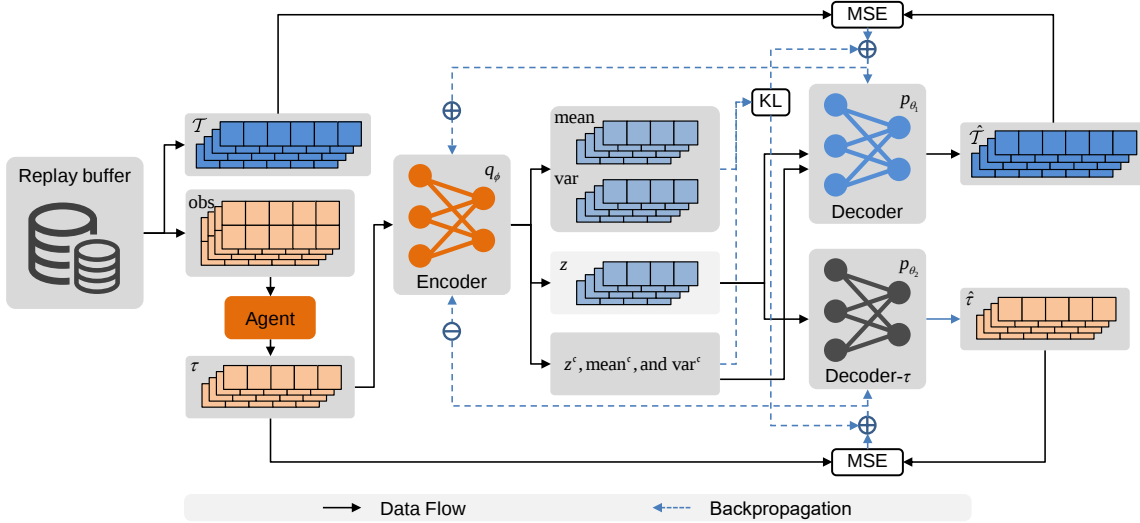


Fig. 5: Topology Inference Network, The Topology Inhanced Network is a VAE-like architecture. It samples the historical observations of all agents in the current traffic scenario from the replay buffer. For each agent i , the encoder q_ϕ extracts a latent variable z_t^i from its trajectory τ_t^i . The decoder p_{θ_2} reconstructs the original observation trajectory from the latent representation z_t^i . Meanwhile, the decoder p_{θ_1} takes the concatenated latent set z_t^{i+} as input to predict the future topological state $\hat{\mathcal{T}}_{t+1}^i$.

sequently, the latent variable $z_t^{(i)}$ is actively forced to discard trajectory-specific micro-variations (increasing reconstruction error) and exclusively encode the semantic intents vital for the macroscopic topology prediction.

3) *Integration into the TPE-MARL Training Loop*: The adversarial inference network operates centrally during training to quantify the collaboration reward $r_{\text{topo},t}$ without impeding decentralized execution. Given a sampled batch of trajectories, the integration proceeds as follows:

- **Latent encoding**: Map historical trajectories $\tau_t^{(i)}$ of all agents to latent representations $z_t^{(i)}$ via the encoder q_ϕ .
- **Joint prediction (Full context)**: For each interaction pair (i, j) , evaluate the log-likelihood of predicting the true future topology $\log p_{\theta_1}(\mathcal{T}_{t+1}^{(i)} | z_t^{i+})$ using the intact extended set.
- **Marginalized prediction (Ablated context)**: Construct an ablated context z_t^{partial} by resubstituting $z_t^{(j)}$ with a decorrelated noise sample $\tilde{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Evaluate the marginalized log-likelihood $\log p_{\theta_1}(\mathcal{T}_{t+1}^{(i)} | z_t^{\text{partial}})$.
- **ELBO computation**: Compute the pairwise estimator $\hat{\mathcal{I}}(\cdot)$ via the log-likelihood difference (Equation (16)) and aggregate it to formalize the global reward $r_{\text{topo},t}$ (Equation (17)).

To stabilize the temporal difference (TD) learning of the main reinforcement learning policy, the collaboration reward $r_{\text{topo},t}$ is evaluated using a slowly updated target inference network (parameterized by ϕ^-, θ_1^-). The concurrent optimization of the main RL policy and the adversarial topology network is systematically detailed in the overall algorithm architecture (Section IV-D).

D. Overall TPE-MARL Architecture and CTDE Optimization

The preceding sections have established the three fundamental pillars of our methodology: the discrete quotient space

representation via the Game Topology Tensor (Section IV-A), the information-theoretic dual intrinsic reward mechanism (Section IV-B), and the adversarial inference framework for mutual information estimation (Section IV-C). This section synthesizes these components into an end-to-end framework, termed *Topology-enhanced Multi-Agent Reinforcement Learning* (TPE-MARL).

Operating under the Centralized Training with Decentralized Execution (CTDE) paradigm, TPE-MARL leverages global topological structures during training to shape the value landscape, while strictly relying on localized ego-centric observations during execution.

1) *Asymptotic Policy Invariance and Hybrid TD Learning*: The core RL objective is driven by a composite reward signal that dynamically balances the extrinsic environmental constraints with the intrinsic exploration imperatives. The joint reward $r_{\text{tot},t}$ is formalized as a linear scalarization:

$$r_{\text{tot},t} \triangleq r_{\text{env},t} + \beta_1 r_{\text{vis},t} + \beta_2 r_{\text{topo},t}, \quad (23)$$

where $r_{\text{env},t} = \mathcal{R}(s_t, \mathbf{a}_t)$ governs the macroscopic task objectives (Section III-C), and $\beta_1, \beta_2 \in \mathbb{R}^+$ dictate the exploration intensities.

Asymptotic Policy Invariance. A primary theoretical concern when introducing intrinsic bonuses is the potential alteration of the original Dec-POMDP Nash equilibrium. However, the proposed intrinsic rewards guarantee *asymptotic policy invariance*. As the state space coverage expands, the topological visitation count $N(\mathcal{T}_t) \rightarrow \infty$, strictly bounding $r_{\text{vis},t} \rightarrow 0$. Concurrently, as the latent representations converge, the conditional mutual information $\hat{\mathcal{I}}(\cdot)$ plateaus, transforming $r_{\text{topo},t}$ into a stationary potential-based shaping term. Consequently, the optimal policy derived from the hybrid reward asymptotically converges to the optimal policy of the extrinsic task.

Monotonic Value Factorization. To optimize the joint

action-value function under the CTDE paradigm, we employ a monotonic value factorization architecture inspired by QMIX [4]. Each active CAV $i \in \mathcal{N}_t$ maintains a decentralized utility network $Q^{(i)}(\mathbf{o}_t^{(i)}, a_t^{(i)}; \theta_i)$, parameterized by a GRU-MLP cascade.

During centralized training, a hypernetwork conditionally generates the weights of a mixing network f_{mix} based on the global state s_t . To satisfy the Individual-Global-Max (IGM) condition, the hypernetwork outputs strictly non-negative weights, ensuring a monotonic relationship between the decentralized utilities and the joint action-value:

$$Q_{\text{tot}}(s_t, \mathbf{a}_t; \theta) \triangleq f_{\text{mix}}\left(\left\{Q^{(i)}(\mathbf{o}_t^{(i)}, a_t^{(i)}; \theta_i)\right\}_{i \in \mathcal{N}_t}, s_t; \theta_{\text{mix}}\right), \quad (24)$$

where $\frac{\partial Q_{\text{tot}}}{\partial Q^{(i)}} \geq 0$, $\forall i \in \mathcal{N}_t$, and $\theta = \theta_{\text{mix}} \cup \{\theta_i\}$ encapsulates all RL parameters.

The RL policy is optimized by minimizing the temporal difference (TD) loss over batches sampled from a replay buffer $\mathcal{D}$:

$$\mathcal{L}_{\text{TD}}(\theta) = \mathbb{E}_{\tau \sim \mathcal{D}} \left[\left(r_{\text{tot},t} + \gamma \max_{\mathbf{a}'} Q_{\text{tot}}(s_{t+1}, \mathbf{a}'; \theta^-) - Q_{\text{tot}}(s_t, \mathbf{a}_t; \theta) \right)^2 \right], \quad (25)$$

where θ^- denotes the parameters of the periodically updated target network.

2) *The Alternating Optimization Paradigm*: The holistic TPE-MARL framework is characterized by an alternating optimization paradigm between two distinct differentiable graphs: the reinforcement learning policy parameterized by θ (minimizing $\mathcal{L}_{\text{TD}}$), and the adversarial topology inference network parameterized by $\Phi = \{\phi, \theta_1, \theta_2\}$ (minimizing $\mathcal{L}_{\text{TopoNet}}$). The unified procedure is formalized in Algorithm 1.

This concludes the rigorous formulation of the TPE-MARL framework. By integrating the Game Topology Tensor construction, the information-theoretic dual intrinsic reward mechanism, and the adversarial variational inference within an alternating optimization CTDE paradigm, the algorithm systematically transforms the intractable multi-agent exploration problem into a structured, information-guided exploration of a compact quotient space.

V. SIMULATION EXPERIMENTS AND EVALUATION

This section presents a comprehensive empirical evaluation designed to interrogate the efficacy and theoretical properties of the proposed TPE-MARL framework. Moving beyond conventional task-specific metrics, our experimental protocol is structured to answer three fundamental artificial intelligence (AI) research questions: (i) **Exploration Dynamics**: Does the proposed dual intrinsic reward mechanism efficiently navigate the discrete quotient space to yield superior sample efficiency and policy convergence compared to state-of-the-art MARL algorithms? (ii) **Theoretical Optimality**: To what extent does the learned decentralized policy approximate the theoretical upper bound established by an exact, full-state search oracle (MCTS)? (iii) **Emergent Intelligence**: How do the topological exploration incentives manifest as emergent multi-agent

Algorithm 1 Topology-Enhanced MARL with Adversarial IB

Require: Horizons $E_{\text{max}}, T_{\text{max}}$, params $\beta_1, \beta_2, \lambda_{\text{GF}}$
Ensure: Policy θ^* and TopoNet $\Phi^* = \{\phi^*, \theta_1^*, \theta_2^*\}$

- 1: Init buffer $\mathcal{D} \leftarrow \emptyset$, params θ, Φ , targets θ^-, Φ^-
- 2: **for** episode $e = 1$ **to** E_{max} **do**
- 3: Reset $s_0 \sim \mathcal{P}_0(\cdot)$, hidden states $\mathbf{h}_0^{(i)}$ for $i \in \mathcal{N}_0$
- 4: **for** $t = 0$ **to** $T_{\text{max}} - 1$ **do**
- 5: # 1. Quotient Space Projection
- 6: Extract meta-observations $\mathbf{o}_{t,\text{meta}}^{(i)}$ for all CAVs
- 7: $\mathbf{h}_t^{ij} \leftarrow \text{sign}(\mathbf{V}_{\text{hash}} \cdot \Delta \mathbf{o}_{t,\text{meta}}^{(ij)})$ // LSH hash
- 8: Assemble topology tensor $\mathcal{T}_t \in \mathcal{S}_e / \sim_{\mathcal{T}}$
- 9: # 2. Decentralized Execution
- 10: $\mathbf{a}_t \sim \pi_{\theta}(\cdot | \mathbf{o}_t, \mathbf{h}_t)$; observe $r_{\text{env},t}, s_{t+1}$
- 11: # 3. Intrinsic Bonuses
- 12: **Breadth:** $N(\tilde{\mathcal{T}}_t) \leftarrow N(\tilde{\mathcal{T}}_t) + 1$; $r_{\text{vis},t} = 1/\sqrt{N(\tilde{\mathcal{T}}_t)}$
- 13: **Depth (ELBO):**
- 14: $\mathbf{z}_t^{(j)} \sim q_{\phi^-}(\cdot | \tau_t^{(j)})$ // Latent state
- 15: $\mathcal{L}_{\text{full}} \leftarrow \log p_{\theta_1^-}(\mathcal{T}_{t+1} | \mathbf{z}_t^{\text{c}+})$
- 16: $\mathcal{L}_{\text{partial}} \leftarrow \log p_{\theta_1^-}(\mathcal{T}_{t+1} | \mathbf{z}_t^{\text{partial}})$
- 17: $\hat{\mathcal{I}} \leftarrow \mathcal{L}_{\text{full}} - \mathcal{L}_{\text{partial}} \implies r_{\text{topo},t}$ // MI
- 18: $r_{\text{tot},t} \leftarrow r_{\text{env},t} + \beta_1 r_{\text{vis},t} + \beta_2 r_{\text{topo},t}$
- 19: Store $(s_t, \mathbf{o}_t, \mathbf{a}_t, r_{\text{tot},t}, \mathcal{T}_t, s_{t+1})$ in $\mathcal{D}$
- 20: **end for**
- 21: # 4. Alternating Optimization
- 22: **if** $|\mathcal{D}| \geq \text{Batch Size}$ **then**
- 23: Sample transition batch $\mathcal{B} \sim \mathcal{D}$
- 24: **RL:** $\theta \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}_{\text{TD}}(\theta; \mathcal{B})$
- 25: **TopoNet (Adversarial IB):**
- 26: $\theta_{1,2} \leftarrow \theta_{1,2} - \alpha_{\text{topo}} \nabla_{\theta_{1,2}} (\mathcal{L}_{\text{TP}} + \mathcal{L}_{\text{RG}})$
- 27: $\phi \leftarrow \phi - \alpha_{\text{topo}} \nabla_{\phi} (\mathcal{L}_{\text{TP}} + \mathcal{L}_{\text{KL}} - \lambda_{\text{GF}} \mathcal{L}_{\text{RG}})$
- 28: Soft-update targets θ^-, Φ^-
- 29: **end if**
- 30: **end for**
- 31: **return** θ^*, Φ^*

cooperative capabilities, particularly in terms of long-horizon credit assignment and robustness to OOD environmental perturbations?

The evaluation is systematically structured as follows. Section V-A establishes the experimental protocol, defining the stochastic physics engine, baseline algorithms, and AI-centric evaluation metrics. Section V-B analyzes the exploration dynamics within the quotient space. Section V-C rigorously evaluates the policy optimality via the exact MCTS oracle. Section V-D provides a spatiotemporal analysis of the emergent cooperative intelligence. Finally, Section V-E conducts ablation and sensitivity analyses to dissect the information-theoretic components of the framework.

A. Experimental Protocol and Benchmarks

1) *The Stochastic Physics Engine and Protocols*: To rigorously evaluate the multi-agent policies under high non-stationarity, we employ the SUMO (Simulation of Urban MObility) framework, augmented with the FLOW interface [72], functioning as a highly non-stationary and partially observable stochastic physics engine. The environment simulates a continuous three-lane spatial domain, where the complexity

of the Markov game is explicitly controlled through two orthogonal parameters: the macroscopic density λ_{flow} and the controllable penetration rate ρ .

The system density spans three levels ($\lambda_{\text{flow}} \in \{400, 600, 700\}$ pcu/h/lane), determining the combinatorial explosion rate of the joint action space. Concurrently, the penetration rate ($\rho \in \{0.25, 0.50, 0.75, 1.00\}$) directly modulates the ratio of controllable agents to uncontrollable stochastic entities. The uncontrollable entities (HDVs) are driven by established behavioral models (e.g., Krauss and LC2013) equipped with randomized imperfection parameters. In the context of our Dec-POMDP formulation, these entities act as confounding stochastic processes with bounded rationality, constantly injecting OOD perturbations and non-stationarity into the environment’s transition kernel $\mathcal{P}$. This full factorial design yields 12 distinct configurations with varying degrees of game-theoretic complexity, each evaluated across five independent random seeds to ensure statistical significance.

2) *Benchmarks and Theoretical Oracles*: We benchmark TPE-MARL against a spectrum of state-of-the-art algorithms, carefully selected to represent distinct exploration paradigms in MARL:

RepNet-QMIX (Ablation Control, based on [24]): Serving as the foundational representation baseline, this model incorporates the graph-based interaction representation network originating from our prior published work without the explicit topological exploration mechanism. It isolates the exact performance gain attributable to the proposed dual intrinsic rewards.

SI2E [14]: Represents *domain-agnostic structural exploration*, utilizing hierarchical state-action space modeling.

SPIE [15]: Represents *retrospective information-guided exploration*, generating structure-aware behaviors from historical trajectory aggregations.

MASER [8]: Represents *subgoal-directed exploration*, utilizing heuristic intrinsic rewards to navigate towards high-value states in the replay buffer.

CDS [13]: Represents *undirected diversity-seeking exploration*, maximizing the mutual information between agent states and trajectories to enforce behavioral diversity.

Crucially, to establish a rigorous theoretical upper bound, we introduce the **PE-MCTS Oracle**. Anchored in the value-based parallel update tree search methodology established in previous literature, this centralized solver operates with privileged full-state observability and unbounded computational horizons. It provides the exact optimal joint Q-value distributions, serving as the “microscope” for our policy optimality analysis in Section V-C.

3) *AI-Centric Evaluation Metrics*: While the environment natively outputs physical measurements, we systematically map these domain-specific indicators to core AI capabilities to evaluate the fundamental quality of the learned representations:

Global Reward Optimization Convergence: *Task Success Rate (Succ.Rate)* and *Instantaneous Traffic Flow (Inst.Flow)*. These metrics quantify the policy’s ability to overcome the sparse gradient problem and converge to a globally coherent Nash equilibrium that maximizes the collective return.

OOD Robustness: *Collision Rate (Coll.)* and *Time-to-Collision (TTC)*. Rather than mere safety checks, these metrics evaluate the policy’s resilience against the severe non-stationarity and catastrophic covariate shifts induced by the uncontrollable stochastic agents.

Action Space Stability: *Mean Absolute Jerk*, *Velocity Variance*, and *Lane-Change Interval*. These metrics assess the temporal smoothness of the policy, ensuring the neural network has learned a continuous, physically feasible action manifold rather than outputting high-frequency, jittering control signals.

To encapsulate the overall alignment with the optimization objective, we report the *Average Traffic Score (ATS)* as the composite macroscopic metric.

B. Exploration Dynamics in the Quotient Space

The core theoretical premise of TPE-MARL is that optimal multi-agent coordination can be achieved by shifting the exploration from the intractable continuous state manifold to a structured discrete quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$. This subsection empirically validates this claim by tracing the exploration trajectories, analyzing the resulting learning dynamics, and quantifying the converged macroscopic performance.

1) *Visualizing Topological Exploration Trajectories*: To provide direct empirical evidence of the intrinsic exploration mechanisms, Figure 6 visualizes the exploration heatmaps for CDS, SI2E, and TPE-MARL across the training progression (20% to 100%). The heatmaps aggregate the visit counts of the 2D game relation vectors $\mathbf{D}^{ij} = (d_{\text{mag}}^{ij}, h_{\text{int}}^{ij})$ projected into the quotient space, effectively illustrating the algorithmic attention over different topological equivalence classes.

Undirected vs. Static vs. Dynamic Exploration. The exploration pattern of CDS (Figure 6, top row) exhibits a nearly uniform, high-entropy distribution throughout the entire training lifecycle. This visualization confirms its theoretical limitation: by unconditionally maximizing behavioral diversity, CDS scatters computational resources across strategically trivial topologies, failing to exploit critical coordination patterns. SI2E (Figure 6, middle row) demonstrates structural striations but remains largely static after the 60% checkpoint, indicating a premature convergence to a fixed set of state-action hierarchical structures without adapting to the dynamic game geometry.

In stark contrast, TPE-MARL (Figure 6, bottom row) displays a highly structured and evolving phase transition that perfectly corroborates our theoretical design in Section IV. At the early training stage (20%), the topology novelty reward $r_{\text{vis},t}$ dominates, driving a uniform grid-like scanning process that maximizes the marginal entropy of the visited topologies. As training progresses (40%–60%), the collaboration reward $r_{\text{topo},t}$ begins to heavily weight the ELBO gradients, transitioning the exploration into distinct high-intensity bands. By convergence (100%), the exploration precisely locks onto sharp, high-intensity hot-spots. This confirms that the dual intrinsic mechanism successfully resolves the exploration-exploitation dilemma: systematically sweeping the quotient space for breadth, and strictly exploiting the information-rich topological structures for depth.

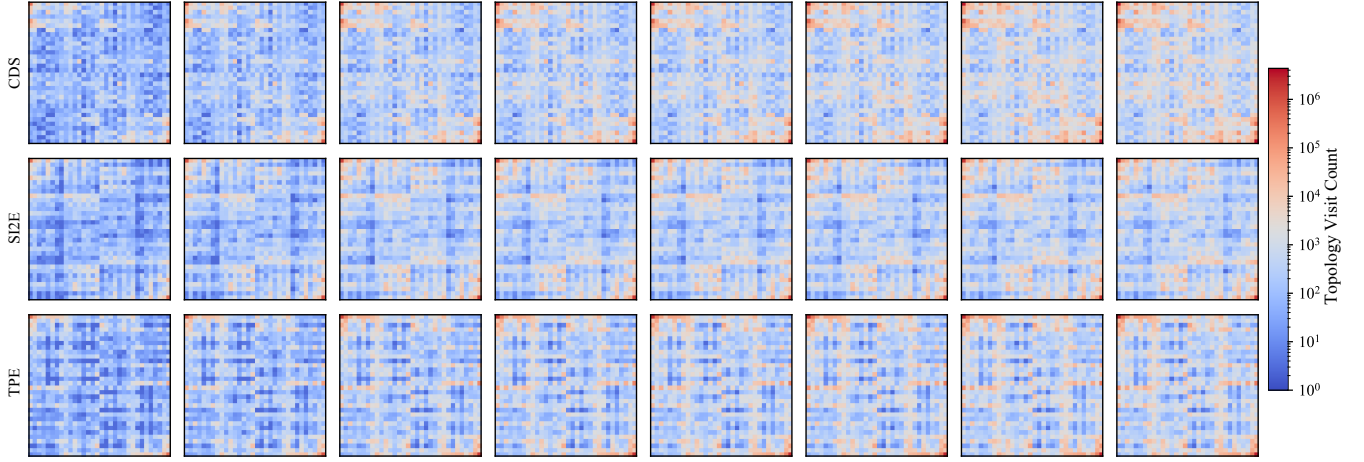


Fig. 6: Exploration heatmaps in the topological space for CDS, SI2E, and TPE-MARL at four training checkpoints. The horizontal axis represents the discretized magnitude d_{mag}^{ij} of the game relation vector, and the vertical axis represents the SimHash code h_{int}^{ij} . TPE-MARL exhibits a characteristic transition from uniform coverage (20%) to focused exploitation (100%), validating the dual intrinsic reward mechanism.

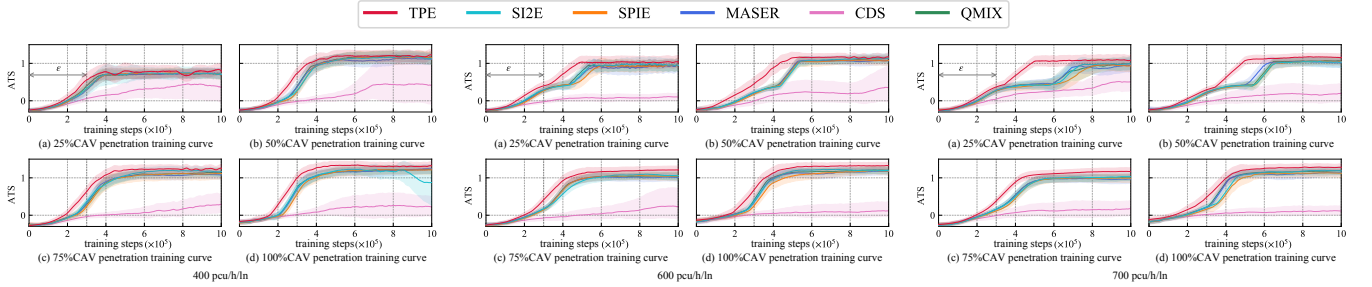


Fig. 7: Training curves under 400, 600, 700 pcu/h/ln traffic flow rate. Each subplot corresponds to a CAV penetration rate ρ : (a) 25%, (b) 50%, (c) 75%, (d) 100%. Solid lines denote the mean ATS across 5 seeds, shaded regions denote the 95% confidence interval.

2) *Sample Efficiency and Convergence Dynamics*: The superiority of the topological exploration directly translates into accelerated learning dynamics. Figure 7 present the Average Traffic Score (ATS) learning curves under varying densities (400, 600, 700 pcu/h/ln) and CAV penetration rates ρ . Shaded regions indicate the 95% confidence intervals across five independent random seeds, rigorously bounding the statistical variance.

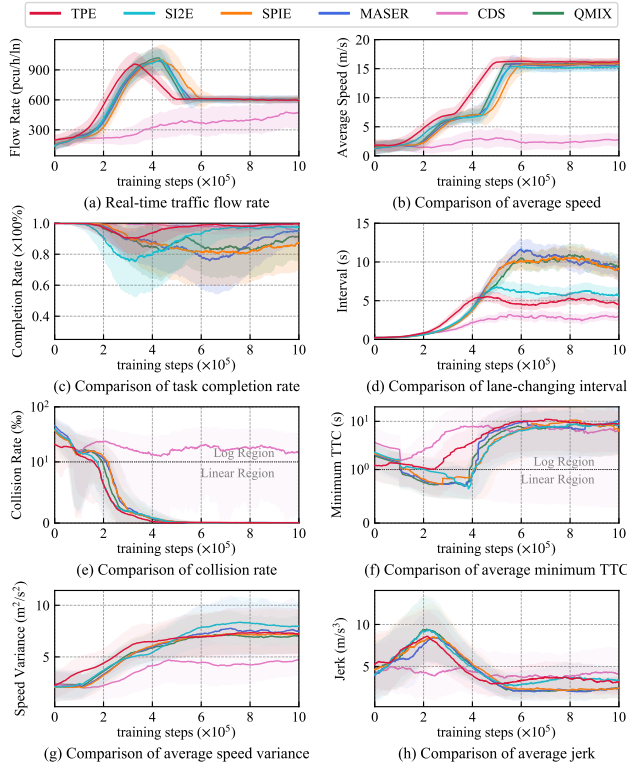
Breaking local optima via topological guidance. A prominent phenomenon is the “staircase” learning dynamics observed in high-density, low-penetration configurations (e.g., Figures 7(600)a and 7(700)a). In such highly non-stationary environments, most baselines (e.g., RepNet-QMIX, MASER) rapidly converge to a conservative safety plateau and fail to progress. TPE-MARL, however, reliably executes a secondary performance leap. This empirically demonstrates that while traditional MARL is trapped in the local optimum of passive avoidance, the information-theoretic collaboration reward actively incentivizes agents to negotiate complex topologies, eventually discovering globally optimal coordination strategies. Consequently, TPE-MARL achieves the steepest ascent and the highest converged ATS envelope across nearly all 12

configurations.

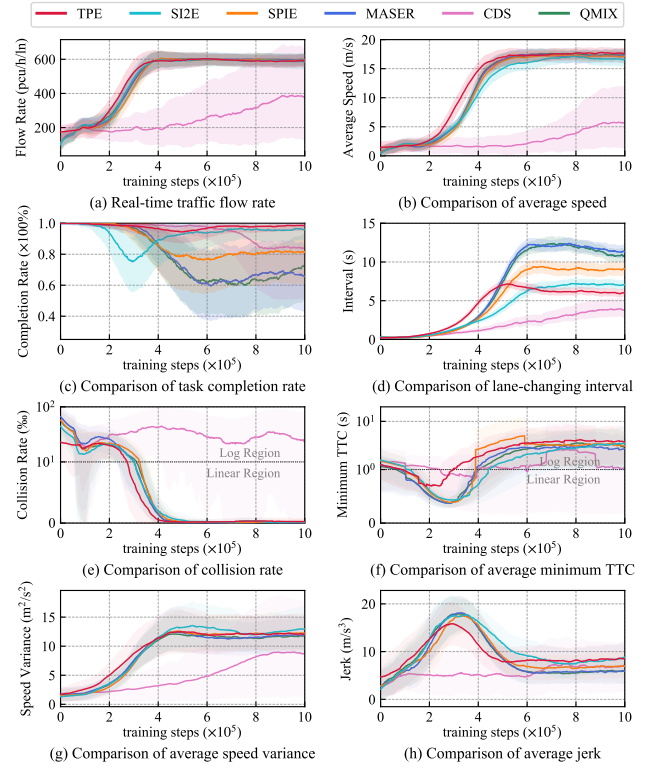
Modulating non-stationarity via penetration rate. The confidence intervals across all algorithms significantly narrow as the CAV penetration rate ρ increases from 0.25 to 1.00. This confirms that the primary source of environmental non-stationarity is the stochastic, unobservable intents of the HDVs, rather than the scale of the controllable agent set. TPE-MARL maintains significantly tighter variance bands than baselines even at $\rho = 0.25$, demonstrating exceptional OOD robustness against the unpredictable dynamics of human-driven entities.

3) *Converged Macroscopic Policy Performance*: To map the learning dynamics to concrete AI capabilities, we thoroughly evaluate the converged policies. Figures 8a to 8d and Table I detail the multi-dimensional performance metrics.

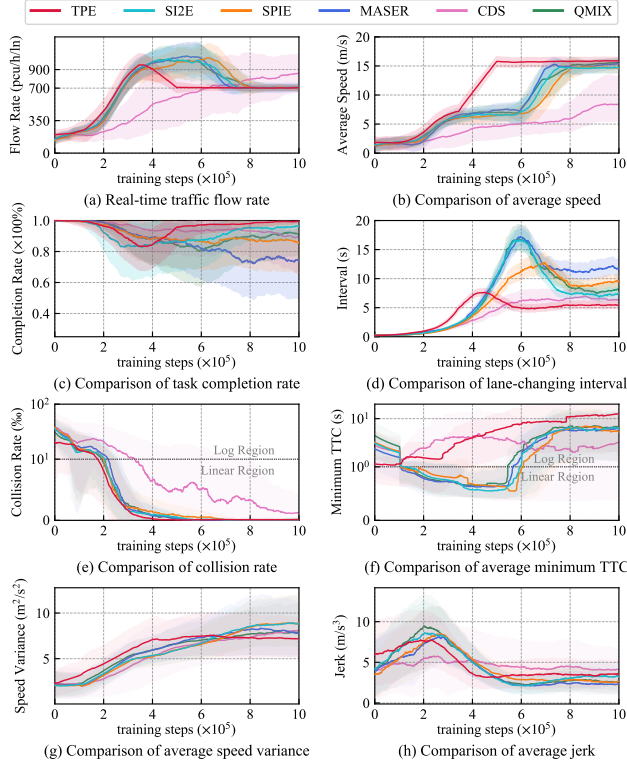
Global Reward Optimization Convergence. The task completion rate serves as the ultimate metric for measuring the successful resolution of the sparse gradient challenge. TPE-MARL establishes an overwhelming dominance, maintaining a success rate above 96% across all tested configurations, with a global average of 98.43% (Table I). This represents a 16.6% absolute improvement over the ablation control,



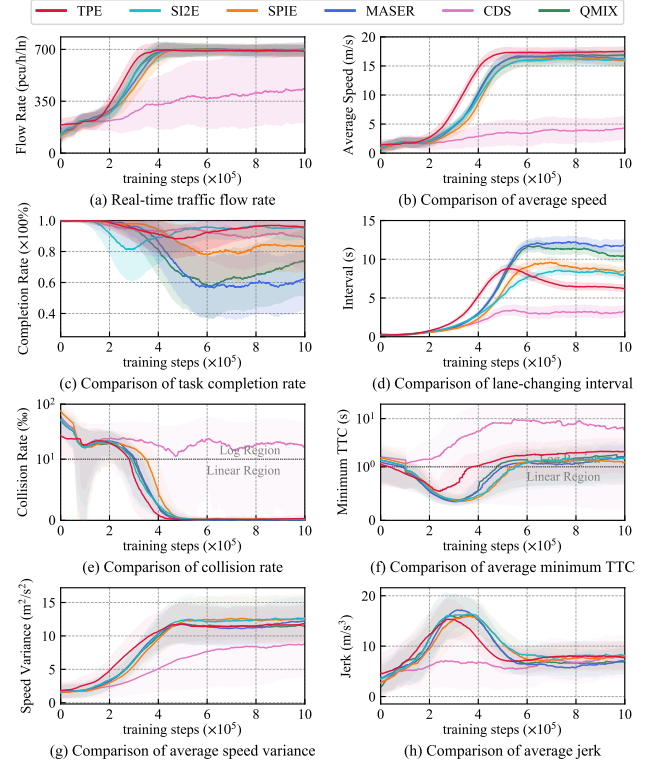
(a) Performance under 600 pcu/h/ln and 25% penetration.



(b) Performance under 600 pcu/h/ln and 75% penetration.



(c) Performance under 700 pcu/h/ln and 25% penetration.



(d) Performance under 700 pcu/h/ln and 75% penetration.

Fig. 8: Comprehensive performance evaluations under varying traffic flow rates and penetration levels.

RepNet-QMIX (84.42%). Furthermore, TPE-MARL achieves the highest global average speed (17.39 m/s), significantly outperforming the undirected exploration of CDS, which suffers catastrophic task failure (5.45 m/s).

TABLE I: Global average performance across all 12 configurations. TPE-MARL achieves the highest task completion rate (98.43%) and average speed (17.39 m/s) while maintaining competitive safety and comfort.

Algorithm	Avg. Velo. (m/s)	Succ. Rate (%)	TTC (s)	Coll. (%)	LC. Intvl. (s)
RepNet-QMIX	16.94	84.42	6.81	1.66	9.17
CDS	5.45	91.47	6.92	17.86	3.60
MASER	16.88	80.14	6.32	0.54	10.14
SPIE	16.66	86.06	6.60	1.04	7.95
SI2E	16.29	94.64	6.33	1.33	6.79
TPE-MARL	17.39	98.43	7.58	1.18	5.56

Action Space Stability and OOD Robustness. The aggressive, high-speed cooperative maneuvering induced by TPE-MARL is achieved without violating physical safety constraints. As shown in Table I, TPE-MARL exhibits the most proactive spatial negotiation, characterized by the shortest average lane-change interval (5.56 s). Despite this active topological restructuring, it maintains an exceptional Time-to-Collision (TTC) margin of 7.58 s and a nominal collision rate of 1.18%, outperforming the topological unawareness of CDS (17.86% collision rate).

These macroscopic outcomes conclusively validate that navigating the discrete quotient space via mutual information and visitation bounds not only accelerates convergence but produces policies that are mathematically efficient, operationally smooth, and highly robust to exogenous stochasticity.

C. Theoretical Optimality via Exact Oracle

While Section V-B established the macroscopic superiority of TPE-MARL, aggregate metrics alone cannot definitively confirm whether the learned policy has truly internalized the optimal game-theoretic structure, or merely discovered a robust heuristic. To rigorously validate Conjecture 1 (Optimal Policy Invariance), we evaluate the learned policy against an exact theoretical upper bound.

We deploy the PE-MCTS oracle (Section V-A2) across a fixed test set of 1,000 representative traffic scenarios. Operating with privileged full-state observability and unbounded computation, the MCTS solver provides the ground-truth optimal joint Q-value distributions, serving as a precise analytical "microscope" to dissect the optimality of the TPE-MARL policy at the granular decision level.

1) *Microscopic Optimality and Sufficient Statistics:* To quantify the alignment between the learned decentralized policy and the centralized oracle, we evaluate the two-dimensional optimality metric $\mathcal{M}_{\text{opt}}$ (defined in Section V-A2), which decouples the decision quality into lateral ($\mathcal{M}_{\text{opt, lat}}$) and longitudinal ($\mathcal{M}_{\text{opt, lon}}$) components. A score of 1.0 indicates exact alignment with the MCTS Q-value landscape.

Figure 9 presents the optimality distributions for all evaluated methods. The results provide compelling empirical val-

idation for the Information Bottleneck theory underlying our framework:

TPE-MARL achieves near-perfect, symmetric optimality. As illustrated in Figure 9a, the TPE-MARL sample points are intensely clustered in the extreme upper-right quadrant, with a mean center of (0.97, 0.99) and minimal variance. This confirms that the discrete quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$ effectively filters out redundant physical noise (e.g., micro-positional deviations) while strictly preserving the *minimal sufficient statistic* required for optimal decision-making.

Baselines suffer from asymmetric spatial blind spots. The SI2E distribution (Figure 9c) exhibits strong lateral but weak longitudinal alignment (mean 0.93, 0.84), reflecting its hierarchical state-space modeling’s inability to capture continuous temporal dynamics. Conversely, SPIE (Figure 9d) shows the exact opposite asymmetry (mean 0.84, 0.95), as its retrospective aggregation excels at longitudinal temporal coherence but fails at discrete lateral planning. By explicitly mapping both the categorical direction (via SimHash) and the continuous magnitude of game relations, TPE-MARL flawlessly resolves this joint discrete-continuous exploration dilemma.

2) *Temporal Policy Evolution and Value Manifold Alignment:* Optimal policy invariance requires that high decision quality is sustained continuously across the temporal horizon, not merely in isolated states. Figures 10 and 11 trace the temporal evolution of the single-agent and joint actions over a 100-step episode.

At the single-agent level, TPE-MARL achieves a 69.49% exact action matching rate with the MCTS oracle, representing a 2.05-fold improvement over the representation-only ablation baseline (RepNet-QMIX, 33.94%). More critically, within the exponentially large joint action space ($|\mathcal{A}_t| = 9^{N_t}$), TPE-MARL achieves a 15.73% exact joint action matching rate (compared to a random probability of $\sim 1.5 \times 10^{-4}$ for $N_t = 4$).

The theoretical significance of these results lies in the analysis of the *mismatched* actions. For the 84.27% of instances where the TPE-MARL joint action deviates from the MCTS absolute argmax, the mean Q-value gap is a negligible 0.23. This indicates that TPE-MARL is not rigidly memorizing deterministic action sequences, but has accurately learned the flat, optimal *value manifold*. The topological intrinsic rewards successfully guide the neural network to identify the correct equivalence classes of near-optimal actions, ensuring robust performance even when the exact MCTS action is not selected.

3) *Macroscopic Bounds and Optimization Landscapes:* The microscopic alignment translates directly into bounded macroscopic performance. As detailed in Table II, TPE-MARL achieves a task completion rate exceeding 96% across all tested penetration rates, with a global mean of 98.83%. This operates within a remarkably tight 1.8% margin of the theoretical upper bound established by the MCTS oracle (99.50% mean), proving that the decentralized execution incurs negligible coordination loss.

Intriguingly, TPE-MARL surpasses the MCTS oracle in average speed under low penetration regimes ($\rho = 0.25$), achieving 16.70 m/s versus the oracle’s 15.24 m/s (a 9.6%

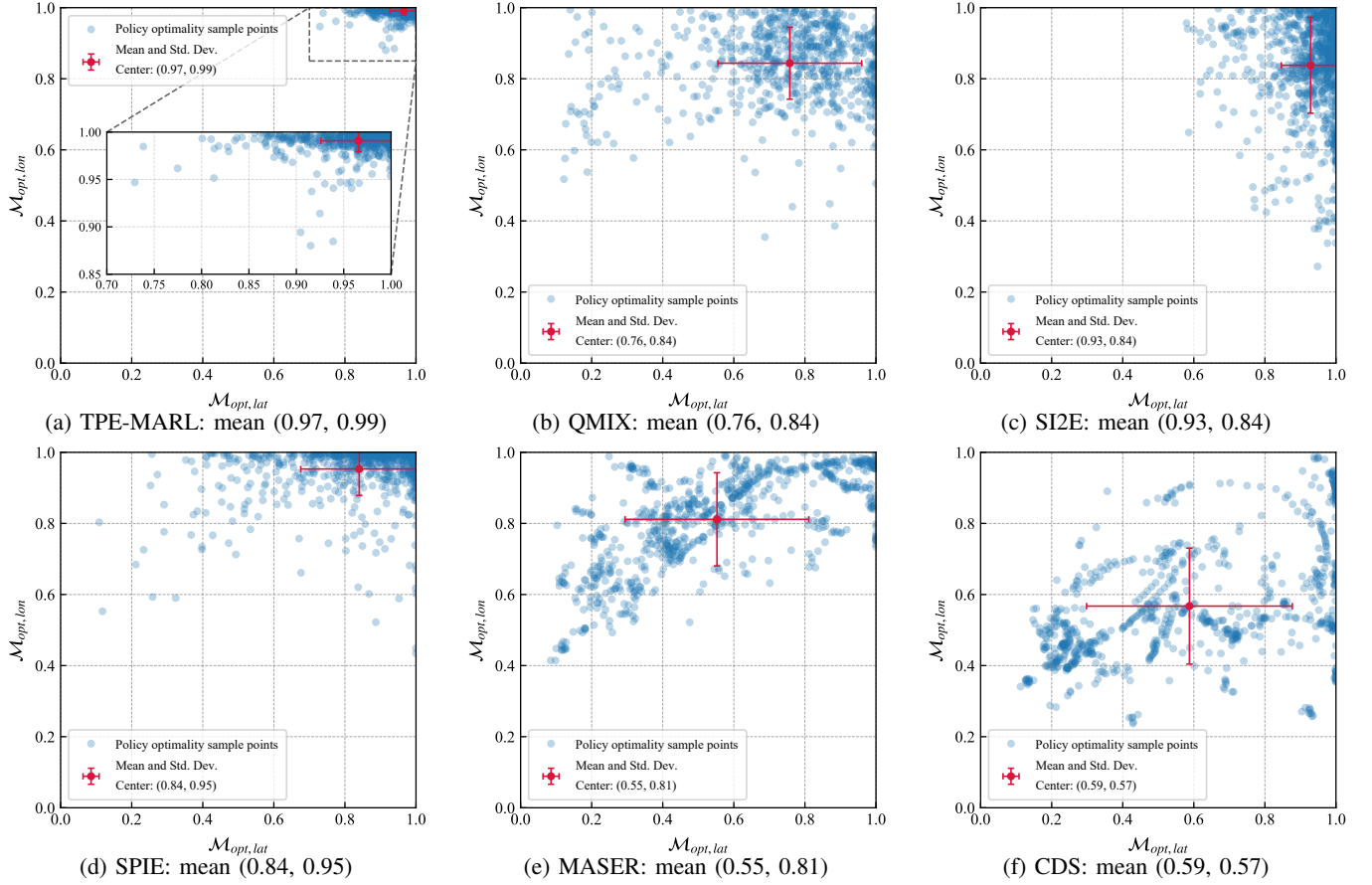


Fig. 9: Optimality metric $\mathcal{M}_{\text{opt}}$ distributions for all methods. Each point represents a single traffic scenario. The red cross marks the mean. TPE-MARL exhibits the most concentrated distribution in the upper-right quadrant, indicating consistent near-optimal decision quality across the test set.

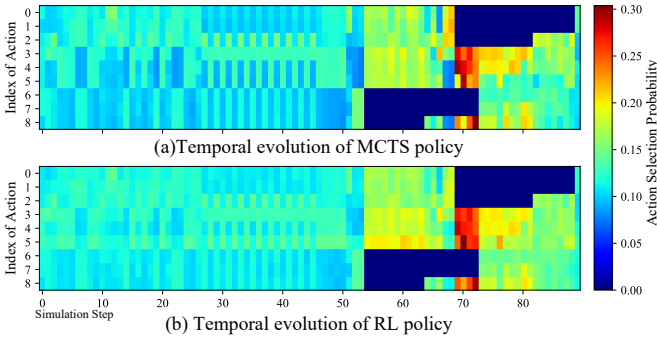


Fig. 10: Single-agent policy evolution for a representative episode. The top row shows the MCTS optimal action at each time step; the bottom row shows the TPE-MARL action. The action matching rate is 69.49%, with the TPE-MARL policy closely tracking the MCTS baseline across the decision horizon.

improvement). Rather than indicating a violation of optimality, this highlights the difference in *optimization landscapes*. The MCTS solver is constrained by a finite planning horizon, rendering it inherently conservative in the presence of stochastic HDVs. In contrast, the RL policy optimizes the

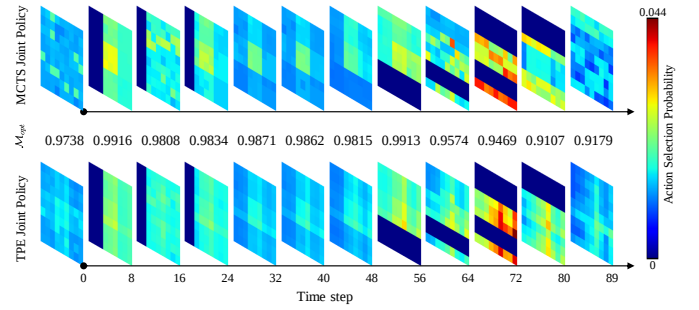


Fig. 11: Joint policy evolution for the same episode. The optimality metric $\mathcal{M}_{\text{opt}}$ remains above 0.91 throughout the episode, demonstrating consistent near-optimal joint decision quality. The joint action matching rate is 15.73%, which is 1,050 times higher than random.

infinite-horizon discounted return. Driven by the collaboration exploration reward, TPE-MARL adopts a substantially more proactive spatial negotiation strategy (reducing lane-change intervals by 69.3% compared to MCTS), demonstrating that the information-theoretic formulation discovers highly efficient cooperative paradigms that elude bounded-horizon tree search.

TABLE II: Performance comparison between TPE-MARL and the MCTS optimal baseline across five CAV penetration rates. The Δ column reports the relative difference. TPE-MARL achieves task completion rates within 1.8% of the theoretical optimum.

Config.	Policy	Velo. (m/s)	Succ. Rate (%)	Coll. (%)	TTC (s)	L.C. Intvl. (s)
$\rho = 0.25$	MCTS	15.24	99.81	0.18	7.62	19.24
	TPE-MARL	16.70	99.30	1.03	5.03	7.12
	$\Delta(\%)$	($\uparrow 9.6\%$)	($\sim 0.5\%$)	($\downarrow 0.85$)	($\downarrow 34.0\%$)	($\downarrow 62.3\%$)
$\rho = 0.40$	MCTS	16.33	99.69	0.25	7.91	18.01
	TPE-MARL	16.76	99.06	1.07	6.85	8.54
	$\Delta(\%)$	($\uparrow 2.6\%$)	($\sim 0.6\%$)	($\downarrow 0.82$)	($\downarrow 13.4\%$)	($\downarrow 52.6\%$)
$\rho = 0.50$	MCTS	17.03	99.78	0.29	9.08	18.13
	TPE-MARL	17.06	98.92	1.11	8.55	9.75
	$\Delta(\%)$	($\uparrow 0.2\%$)	($\sim 0.9\%$)	($\downarrow 0.82$)	($\downarrow 5.8\%$)	($\downarrow 46.2\%$)
$\rho = 0.67$	MCTS	17.16	99.66	0.37	9.21	16.76
	TPE-MARL	17.47	99.09	1.63	7.43	8.94
	$\Delta(\%)$	($\uparrow 1.8\%$)	($\sim 0.6\%$)	($\downarrow 1.26$)	($\downarrow 19.4\%$)	($\downarrow 46.6\%$)
$\rho = 1.00$	MCTS	18.13	99.56	0.14	11.27	18.09
	TPE-MARL	18.80	97.78	0.30	8.93	8.85
	$\Delta(\%)$	($\uparrow 3.7\%$)	($\sim 1.8\%$)	($\downarrow 0.16$)	($\downarrow 20.8\%$)	($\downarrow 51.1\%$)

D. Emergence of Multi-Agent Cooperative Intelligence

The quantitative evaluations in Section V-C definitively established that TPE-MARL strictly approximates the theoretical optimality bound. However, aggregate metrics cannot elucidate the underlying behavioral mechanisms driving this success. In this subsection, we inversely map the physical spatiotemporal trajectories to fundamental MARL capabilities. By analyzing four representative traffic scenarios (Figures 12 to 15), we demonstrate how the topology-enhanced exploration and dual intrinsic rewards explicitly foster advanced cooperative intelligence, elevating the policy from reactive heuristics to strategic multi-agent game-playing.

1) *Long-Horizon Credit Assignment (e.g., Proactive Overtaking)*: One of the most persistent challenges in MARL is temporal myopia—the tendency of agents to greedily optimize for immediate safety or local rewards at the expense of long-term efficiency. The first case (Figure 12) demonstrates how TPE-MARL overcomes this through *long-horizon credit assignment*.

Scenario & Behavior: CAV 1 intends to merge right but is obstructed by two slower HDVs. A myopic, reactive policy would immediately decelerate to merge behind the HDVs, securing an instant safety reward but suffering severe longitudinal delays. Instead, TPE-MARL executes a proactive overtaking maneuver. As evidenced by the steep slope of CAV 1’s trajectory in the time-space diagram (Figure 12b), the agent maintains high speed in the left lane for nearly 30 seconds, entirely bypassing the obstacles before executing the merge.

AI Mechanism: This behavior physically manifests the efficacy of the collaboration reward $r_{\text{topo},t}$. By explicitly valuing future topological information gains, the reward structure penalizes the low-information, highly entangled state of trailing unpredictable HDVs. Consequently, the agent learns to sacrifice the short-term, myopic convenience of an immediate merge in favor of discovering a high-mutual-information topological state (clear space ahead), successfully resolving the long-horizon credit assignment problem.

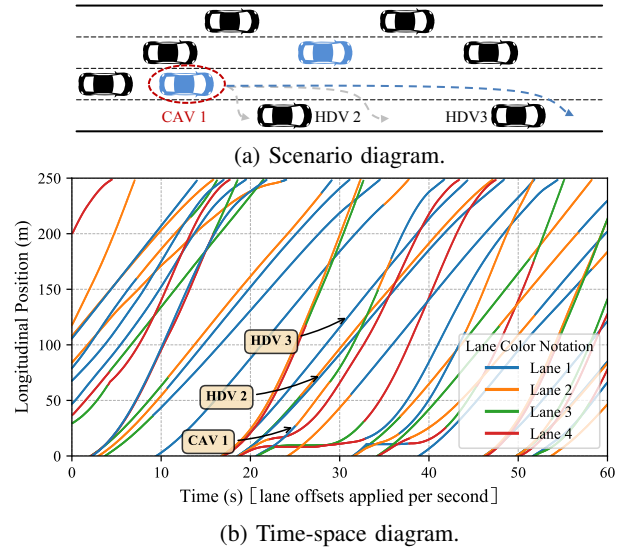


Fig. 12: Proactive overtaking and lane changing scenario: (a) spatial trajectory, where CAV 1 maintains high speed to overtake the obstructing HDVs before merging; (b) time-space diagram, where the steep slope of CAV 1’s trajectory indicates sustained high speed during the overtaking phase.

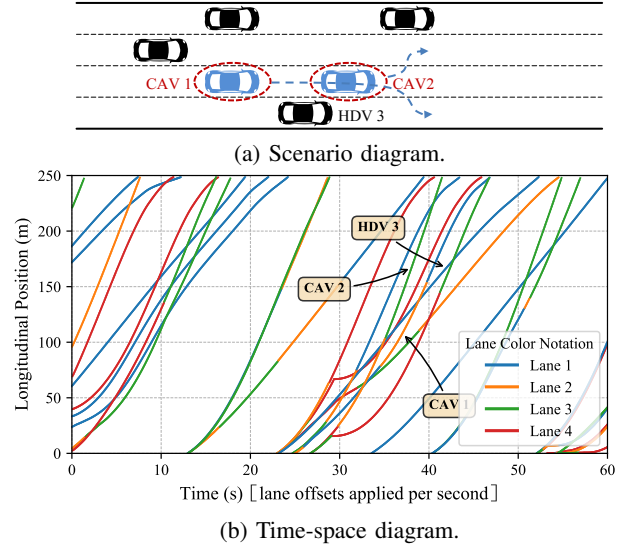


Fig. 13: Cooperative lane allocation scenario: (a) spatial trajectory, where CAV 2 yields lane 2 to CAV 1 by changing to lane 3; (b) time-space diagram, where the lane change of CAV 2 at $t \approx 28$ s creates space for CAV 1’s merge at $t \approx 38$ s.

2) *Emergence of Altruistic Policies (e.g., Cooperative Lane Allocation)*: Achieving true cooperation requires individual agents to execute locally suboptimal actions if they disproportionately benefit the global collective. Figure 13 illustrates the *emergence of altruistic policies* within the learned representations.

Scenario & Behavior: CAV 1 (lane 2) needs to merge into lane 1, which is occupied by a slow HDV. CAV 2 is trailing CAV 1 in lane 2. A selfish policy for CAV 2 would maintain its current optimal lane, effectively blocking CAV 1’s

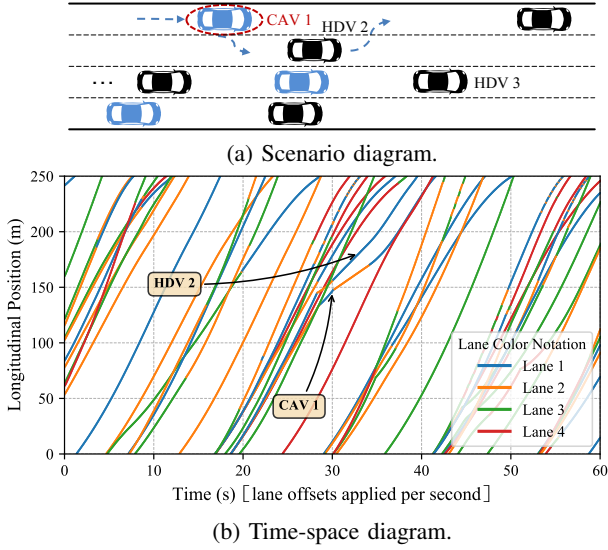


Fig. 14: Defensive evasion scenario: (a) spatial trajectory, where CAV 1 executes a combined deceleration and lane change to evade the intrusive HDV 2; (b) time-space diagram, where the double-peak structure of CAV 1’s lateral position shows the evade-and-return pattern.

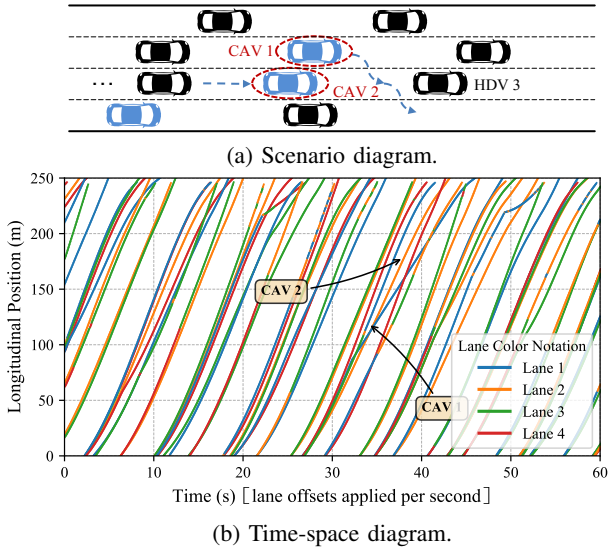


Fig. 15: Zipper-merge scenario: (a) spatial trajectory, where CAV 2 decelerates to create a gap for CAV 1’s merge; (b) time-space diagram, where the gap created by CAV 2’s deceleration at $t \approx 35$ s enables CAV 1’s merge at $t \approx 37$ s.

backward merge and forcing CAV 1 to brake. TPE-MARL, however, orchestrates a sacrificial maneuver. As seen at $t \approx 28$ s (Figure 13b), CAV 2 actively initiates a leftward lane change into lane 3, yielding its right-of-way entirely to CAV 1. This altruistic concession allows CAV 1 to maintain speed and complete a seamless merge at $t \approx 38$ s.

AI Mechanism: This emergent altruism is a direct theoretical consequence of the monotonic value factorization under the CTDE paradigm. Driven by the centralized mixing

network (QMIX), the agents are not optimizing independent utilities, but rather striving to maximize the global Nash equilibrium. The collaboration intrinsic reward further amplifies this by identifying that the coordinated yielding generates a highly predictable, maximum-information joint state, forcing the agents to converge upon this globally optimal cooperative paradigm.

3) *Robustness to Out-of-Distribution Perturbations (e.g., Defensive Evasion):* In open mixed-autonomy environments, uncontrollable HDVs constantly inject extreme non-stationary transitions. Figure 14 illustrates the policy’s *robustness to OOD perturbations*.

Scenario & Behavior: CAV 1 is traveling steadily in lane 4 when an adjacent HDV suddenly executes a forced, aggressive cut-in maneuver. This represents a severe covariate shift, introducing an OOD perturbation rarely densely populated in the training distribution. Rather than a brittle failure or simplistic emergency braking, the TPE-MARL policy executes a graceful degradation. The time-space diagram (Figure 14b) reveals a sophisticated evade-and-return pattern: CAV 1 simultaneously decelerates and changes to lane 3 at $t \approx 30$ s, and after the HDV stabilizes, it smoothly reclaims its original lane at $t \approx 33$ s.

AI Mechanism: The ability to maintain survival and degrade gracefully under unexpected stochasticity is heavily attributed to the topology novelty reward $r_{vis,t}$. By systematically maximizing the marginal entropy of the visited topologies during training, the policy was extensively exposed to extreme topological boundaries. This broad quotient space coverage guarantees that even when facing catastrophic physical OOD shifts, the policy can swiftly map the physical state to a safe, previously explored topological equivalence class, executing a mathematically robust evasion.

4) *Dynamic Negotiation in High-Dimensional Joint Spaces (e.g., Zipper-Merge):* The pinnacle of multi-agent coordination involves continuous, synchronous state modifications across multiple entities. Figure 15 showcase *dynamic negotiation in high-dimensional joint spaces* through the classic zipper-merge.

Scenario & Behavior: CAV 1 (lane 3) must merge into lane 2, inserting itself between CAV 2 and a leading HDV. This maneuver requires sub-second spatiotemporal synchronization. At $t \approx 35$ s (Figure 15b), CAV 2 proactively decelerates to sculpt an exact spatial gap. Simultaneously, CAV 1 aligns its velocity and merges perfectly into the newly created pocket at $t \approx 37$ s, achieving a flawless, high-speed interleaving pattern.

AI Mechanism: The zipper-merge represents the ultimate manifestation of mutual information maximization. In a joint action space scaling exponentially (9^{N_t}), discovering this delicate synchronization through random exploration is near impossible. However, the variational Evidence Lower Bound (ELBO) explicitly evaluates and rewards the predictability of joint topological configurations. By opening the gap, CAV 2 drastically minimizes the conditional entropy $\mathcal{H}(\mathcal{T}_{t+1}^{(1)} | C_{12})$ for CAV 1. The agents collaboratively navigate the high-dimensional action space precisely by seeking these mutual-information-maximizing topological geometries, synthesizing fluid, human-like dynamic negotiation.

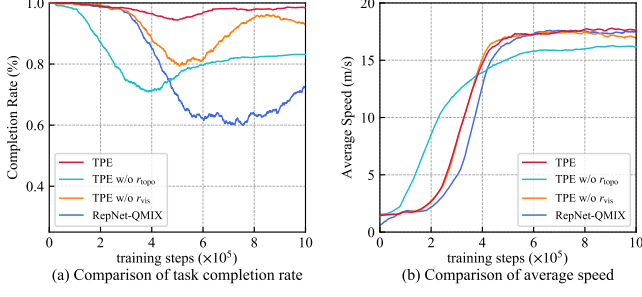


Fig. 16: Component ablation results under 600 pcu/h/ln and 75% CAV penetration. The full TPE-MARL algorithm achieves the highest task completion rate and average speed. The ablation of either intrinsic reward component results in measurable degradation. (a) Task completion rate over training steps. (b) Average speed over training steps.

E. Ablation and Sensitivity Analysis

The preceding evaluations established the system-level superiority of TPE-MARL. In this final subsection, we dissect the internal information-theoretic mechanics of the framework. By ablating individual components and analyzing hyperparameter sensitivities under the challenging 600 pcu/h/ln, 75% CAV penetration configuration, we empirically validate the theoretical design choices formulated in Section IV.

1) *Component Ablation: The Synergy of Breadth and Depth*: To isolate the contributions of the dual intrinsic reward mechanism, we compare the full TPE-MARL architecture against three ablated variants: (i) **TPE w/o r_{topo}** (relying solely on novelty breadth); (ii) **TPE w/o r_{vis}** (relying solely on collaboration depth); and (iii) **RepNet-QMIX** (the purely extrinsic baseline).

Figure 16 presents the learning trajectories of these variants. The full TPE-MARL algorithm unequivocally establishes the highest converged task completion rate and average speed. Removing either intrinsic component results in a severe performance degradation, confirming their theoretical non-redundancy.

Crucially, the ablation exposes the underlying mathematical roles of the two rewards. The variant lacking the novelty reward (TPE w/o r_{vis}) suffers a pronounced mid-training valley (dropping to 80% task success). Without the explicit maximization of marginal state entropy $\mathcal{H}(\mathcal{T})$ provided by r_{vis} , the policy struggles to escape local topological optima. Conversely, the variant lacking the collaboration reward (TPE w/o r_{topo}) initially attains high speed but experiences catastrophic task failure (dropping to 70% success). This confirms that without the conditional entropy minimization (ELBO) provided by r_{topo} to ground the exploration in strategically informative interactions, the novelty-seeking behavior devolves into undirected divergence. The synergy of both rewards is thus strictly necessary to navigate the quotient space stably.

2) *The Information-Theoretic Exploration-Exploitation Trade-off*: The hyperparameters β_1 and β_2 govern the intensity of the intrinsic exploration bonuses. Figures 17 to 20 map the sensitivity of the policy to these parameters. Rather than a monotonic relationship, both parameters

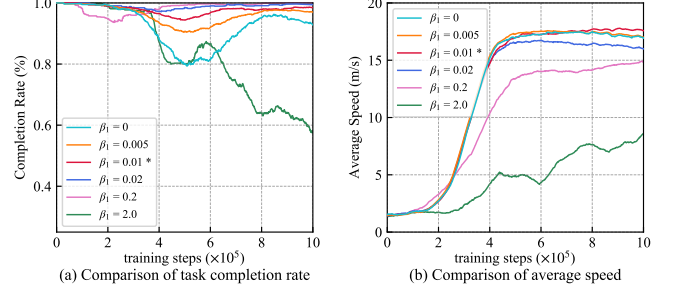


Fig. 17: Sensitivity analysis of the novelty reward weight β_1 . The optimal value is $\beta_1 = 0.01$, achieving the highest task completion rate and average speed.

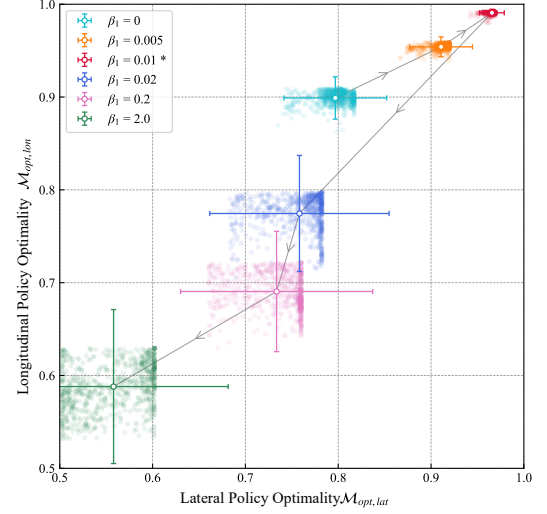


Fig. 18: Effect of β_1 on the optimality metric $\mathcal{M}_{\text{opt}}$ distribution. The distribution shifts toward the upper-right corner at $\beta_1 = 0.01$, indicating improved decision quality.

exhibit a stark *inverted-U shape* in the performance metrics, perfectly corroborating the fundamental information-theoretic exploration-exploitation trade-off.

Sensitivity to β_1 (Novelty Weight): As illustrated in Figures 17 and 18, increasing β_1 from 0 to the optimal value of 0.01 shifts the optimality metric $\mathcal{M}_{\text{opt}}$ distribution firmly into the upper-right quadrant, achieving maximal task success. However, excessively high values ($\beta_1 \geq 0.2$) trigger a dramatic collapse in both speed and success rate. Theoretically, an over-weighted r_{vis} forces the policy to prioritize traversing the quotient space indiscriminately over the actual driving task. This induces a *dimension-cursed undirected divergence*, where agents continuously destabilize their policies merely to visit novel but irrelevant topological extremities.

Sensitivity to β_2 (Collaboration Weight): The response to β_2 (Figures 19 and 20) reveals an even more profound phenomenon. The optimal balance is struck at $\beta_2 = 0.02$. When β_2 is pushed to extreme values (e.g., $\beta_2 = 0.2$), the task completion rate remains artificially high, but the average speed plummets to near-zero. This mathematical pathology can be diagnosed as an *information-locked state*. The mu-

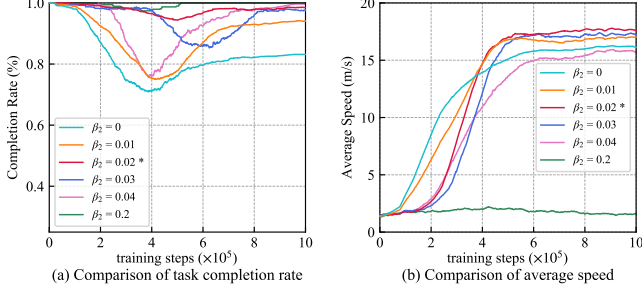


Fig. 19: Sensitivity analysis of the collaboration reward weight β_2 . The optimal value is $\beta_2 = 0.02$, achieving the best balance between task completion and speed.

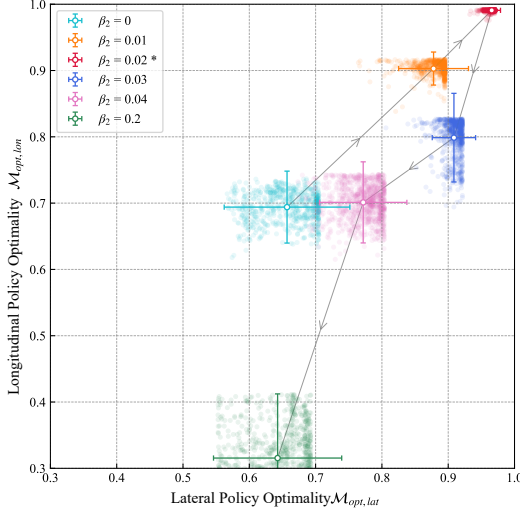


Fig. 20: Effect of β_2 on the optimality metric $\mathcal{M}_{\text{opt}}$ distribution. The distribution shifts toward the upper-right corner at $\beta_2 = 0.02$.

tual information estimator heavily rewards predictable joint representations. If β_2 dominates the extrinsic reward, the agents discover that the easiest way to maximize predictability (and thus r_{topo}) is to simply stop moving or maintain static, rigid formations. They become “locked” in low-speed, high-information topological configurations, refusing to execute dynamic maneuvers (like lane changes) that temporarily increase conditional entropy. This validates that the mutual information reward must strictly act as a shaping bonus, bounded by the extrinsic task objective.

3) *Information Fidelity vs. Topological Compactness*: The final analysis investigates the SimHash code length m , which dictates the resolution of the angular mapping operator (Equation 4). Figure 21 demonstrates the algorithm’s sensitivity to $m \in \{6, 8, 9, 10, 12\}$.

The parameter m explicitly controls the capacity of the *Information Bottleneck*. At $m = 6$ (64 hash codes), the bottleneck is excessively narrow. The performance severely degrades because the quotient space is too compact, leading to severe *topological aliasing*—fundamentally distinct game-theoretic states collide into the same equivalence class, hopelessly corrupting both the novelty visit counts and the mutual

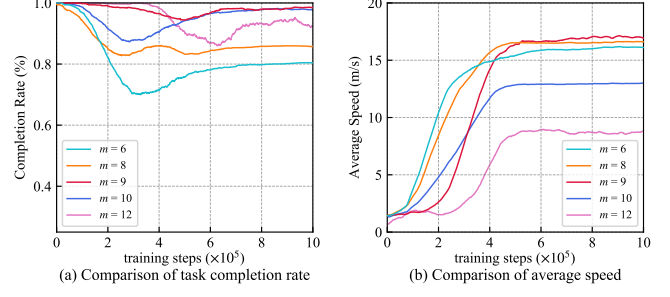


Fig. 21: Sensitivity analysis of SimHash code length m . The optimal value is $m = 9$, balancing information fidelity and topological space compactness.

information estimates.

Conversely, as m increases beyond 9 (e.g., $m = 12$, yielding 4,096 hash codes), the performance gains saturate, but the convergence rate noticeably deteriorates, requiring approximately 20% more training steps. While a wider bottleneck preserves maximum *game-theoretic information fidelity*, the quotient space expands exponentially, diluting the visitation densities and weakening the UCB exploration bounds.

The empirical optimum at $m = 9$ (512 distinct directional topologies) confirms our theoretical derivation: it represents the optimal Information Bottleneck. It flawlessly balances *topological space compactness* (ensuring statistically significant visitation densities for UCB exploration) with *information fidelity* (preserving sufficient geometric granularity for the optimal policy invariance conjecture to hold).

VI. PHYSICAL SCALE-MODEL EXPERIMENTS

The preceding simulation evaluations (Section V) rigorously established the theoretical optimality and sample efficiency of the TPE-MARL framework. However, the ultimate validation of any representation learning algorithm lies in its OOD generalization capabilities. The transition from simulation to a physical cyber-physical system (CPS) introduces severe covariate shifts—specifically, unmodeled physical dynamics, sensor noise, and communication latency.

This section elevates the “sim-to-real” gap from an engineering hurdle to a fundamental test of representation robustness. We aim to demonstrate that the discrete quotient space induced by the Game Topology Tensor acts as a mathematically rigorous Information Bottleneck, effectively filtering out continuous high-frequency physical noise while preserving the low-frequency topological intents necessary for multi-agent coordination.

A. Cyber-Physical Testbed and Topological Homomorphism

To physically validate the policy without losing theoretical stringency, the experimental platform must faithfully reproduce the multi-agent interaction structure while deliberately injecting realistic sources of domain shift.

1) The Physical Layer as an OOD Noise Generator:

Figure 22 illustrates the layered architecture of the physical scale-model platform. Crucially, the *Physical environment layer* is not merely a hardware execution space; it functions mathematically as an OOD noise generator that constantly perturbs the idealized Dec-POMDP transition kernel.

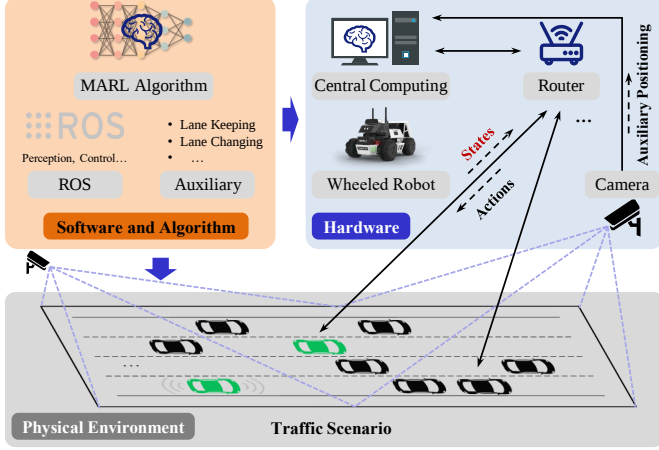


Fig. 22: Layered architecture of the cyber-physical testbed. Rather than a simple execution environment, the physical and hardware layers act as OOD noise generators, injecting non-stationary transition kernels (PID errors) and unobservable state latencies (SLAM delays) to rigorously test the policy’s representation robustness.

The execution entities are Ackermann-steering differential robots equipped with Jetson Nano onboard computers (67 TOPS), orchestrated by a central Intel i7-12700K edge node. The hardware constraints, summarized in Table III, explicitly introduce two critical forms of non-stationarity:

Non-stationary Transition Kernel: The commanded joint actions $\mathbf{a}_t$ are executed by low-level PID controllers. The inevitable tracking errors and friction variations introduce a stochastic perturbation, meaning the physical transition $\mathcal{P}_{\text{real}}(\cdot \mid s, \mathbf{a}_t)$ systematically deviates from the simulation kernel.

Unobservable State Latency: The laser SLAM and IMU localization systems operate at 30 Hz, introducing up to 33 ms of temporal lag. This creates a severe partial observability challenge where the policy must act on delayed, noisy state projections.

TABLE III: Hardware specifications of the scale-model platform. These physical constraints explicitly define the boundaries of the covariate shift introduced into the real-world deployment.

Category	Item	Specification
Robots	Dimensions	450×220×150 mm
	Max speed	1.0 m/s
	Compute	Jetson Nano (67 TOPS)
Control	Localization	Laser SLAM + IMU (30 Hz, ±5 mm)

By maintaining a strict geometric scale (e.g., 0.36 m lane width corresponding to the robot dimensions), the platform ensures that the underlying game-theoretic geometries remain consistent, isolating the evaluation strictly to the algorithm’s robustness against these injected physical perturbations.

2) *Topological Homomorphism via Spatial Relaxation:* A profound theoretical challenge arises during the continuous physical execution of lateral maneuvers. In simulation, lane changes are often discretized as instantaneous state transitions. However, in the physical continuous manifold, a vehicle executing a lane change simultaneously occupies both the source and target lanes for an extended duration.

If we impose a hard discretization threshold (e.g., snapping the agent to the nearest lane centerline), the continuous-discrete mapping $\mathcal{T}(s_e)$ collapses. This discrete-continuous mismatch destroys the topological equivalence relation ($\sim_{\mathcal{T}}$) established in Conjecture 1, leading to severe topological aliasing where a blocking vehicle is falsely represented as safely occupying only a single lane.

To resolve this, we introduce the *Companion State Projection* mechanism, theoretically formalized as a *spatial relaxation* to maintain topological homomorphism. As illustrated in Figure 23, when the physical footprint $\mathcal{R}^{(i)}(t)$ intersects two adjacent lanes, the mechanism projects a virtual companion contour into the intersected lane. This projection is directly integrated into the neighbor meta-observation tensor $\mathbf{s}_{t,\text{nbr}}^{(i)}$.

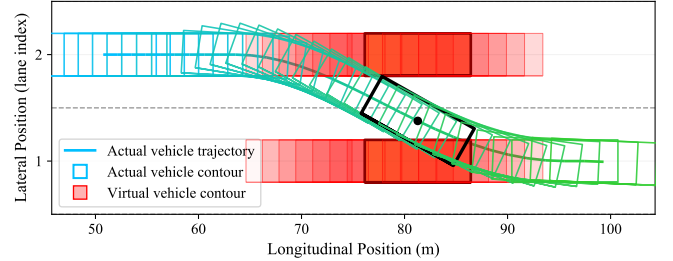


Fig. 23: Topological Homomorphism via Spatial Relaxation. During continuous physical lane changes, hard boundary thresholds destroy the topological state structure. The companion state projection acts as a mathematical relaxation, generating virtual contours to ensure the discrete quotient space representation remains strictly consistent and equivalent during physical transients.

Rather than a simple engineering patch to avoid collisions, this spatial relaxation mathematically guarantees that the surjective mapping from the continuous physical space to the discrete quotient space remains smooth and homomorphic during transient executions. It inherently absorbs the spatial stochasticity of the PID-controlled lane change, ensuring that the optimal policy invariance (Conjecture 1) holds strictly true even amidst the physical domain shift.

B. OOD Generalization in High-Dimensional Joint Action Spaces

To rigorously evaluate the policy’s capacity for multi-agent coordination and out-of-distribution generalization, we

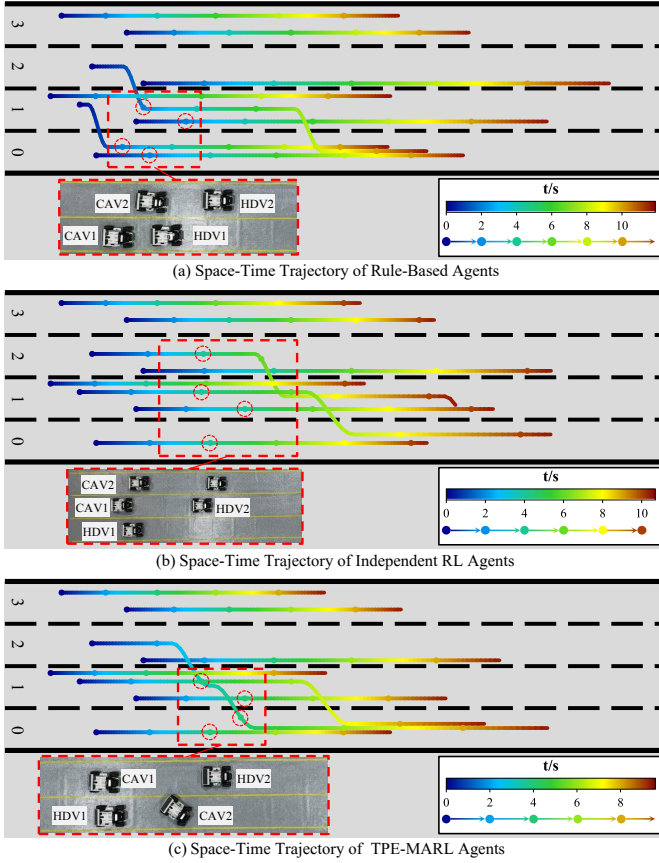


Fig. 24: Space-time trajectory comparison demonstrating OOD generalization capabilities. (a) Rule-based policies fall into conservative local optima. (b) Independent RL suffers from severe covariate shift and overfitting, leading to aggressive near-collision behaviors when exposed to physical latency. (c) TPE-MARL exhibits profound OOD robustness, flawlessly executing a cooperative zipper-merge through precise long-horizon credit assignment despite physical execution noise.

deploy the algorithms in a physical zipper-merge scenario. By deliberately focusing on the zipper-merge rather than geometrically constrained configurations such as ramp merging, the experimental design elegantly isolates pure multi-agent game interactions. As illustrated in Figure 15a, this creates a highly coupled joint action space where any individual agent’s selfish optimum (e.g., forcing a merge) inevitably leads to a globally sub-optimal Nash equilibrium. This scenario serves as the ultimate crucible for testing whether the learned representations can synthesize cooperative intelligence amidst physical uncertainties.

1) *Trajectory Analysis and Covariate Shift*: We contrast the TPE-MARL policy against a reactive rule-based baseline (representing non-cooperative traffic) and an Independent RL baseline (representing decentralized execution without topological information bottlenecks). The space-time trajectories in Figure 24 reveal starkly different responses to the physical domain shift.

Rule-Based: Conservative Local Optima. As shown in Figure 24a, the reactive policy exhibits fundamentally myopic

behavior. CAV-1, constrained by deterministic safety rules, immediately decelerates to yield to the HDV, forcing the trailing CAV-2 into a cascade of braking. The system is trapped in a conservative local optimum characterized by low speed and high temporal occupancy.

Independent RL: Covariate Shift and Catastrophic Overfitting. The Independent RL trajectory (Figure 24b) visually exposes the vulnerability of conventional deep reinforcement learning to the sim-to-real gap. Operating directly on the raw continuous state space, the Independent RL agent lacks an information bottleneck to filter physical noise. Consequently, it suffers from severe *covariate shift*. While it may achieve high efficiency in the idealized deterministic simulation, it catastrophically overfits to the idealized transition kernel. When exposed to the physical testbed’s 33 ms unobservable latency and underlying PID tracking errors, the policy becomes hypersensitive. This manifests as highly aggressive, near-collision spatial negotiations, completely failing to maintain safety margins.

TPE-MARL: OOD Generalization and Long-Horizon Credit Assignment. In stark contrast, the TPE-MARL trajectory (Figure 24c) demonstrates extraordinary OOD generalization. Despite the identical physical noise and latency injected into the system, the policy executes a mathematically precise zipper-merge. CAV-1 proactively decelerates, deliberately sacrificing its own immediate spatial progress to sculpt a safe merging pocket for CAV-2. This behavior perfectly executes *long-horizon credit assignment*. Because the Game Topology Tensor effectively absorbs the high-frequency physical perturbations into a stable discrete equivalence class, the policy remains confident in its learned representations, enabling the agents to seamlessly negotiate the high-dimensional joint space in the physical world.

2) *Quantitative Efficacy of Joint Mutual Information*: To quantify the systemic impact of this topological coordination, we evaluate the cooperative efficiency gain G_{coop} across 10 repeated physical trials.

TABLE IV: Quantitative validation of the zipper-merge scenario. The cooperative efficiency gain ($G_{\text{coop}} = +1.35$ s) achieved by TPE-MARL serves as the direct physical manifestation of the joint mutual information optimized by the dual intrinsic rewards.

Condition	Total task time (s)	Cooperative gain G_{coop} (s)
Reactive rule-based	11.24 ± 0.45	-0.73
Independent RL	10.51 ± 0.72	0
TPE-MARL	9.16 ± 0.15	+1.35

As detailed in Table IV, TPE-MARL achieves a total task completion time of 9.16 s, yielding a substantial cooperative gain of +1.35 s over the Independent RL baseline. Furthermore, the temporal variance is compressed to a mere ± 0.15 s, sharply contrasting with the ± 0.72 s variance of the overfitting Independent RL.

Crucially, this +1.35 s improvement is not merely an engineering metric; it represents the direct physical manifes-

tation of optimizing the joint mutual information via the dual intrinsic rewards. The collaboration reward incentivized the agents to seek maximally predictable topological interactions during training, which physically translated into a highly synchronized, low-variance dynamic negotiation that is provably resilient to the noise of the real world.

C. Executability and Computational Complexity of the Information Bottleneck

The final dimension of our physical validation investigates the real-time executability of the proposed architecture. Rather than merely reporting engineering metrics, we translate trajectory tracking accuracy and system latency into direct empirical evidence for the *smoothness of the learned action manifold* and the *computational scalability of the topological dimensionality reduction*.

1) Smooth Action Manifolds and Behavioral Consistency:

The trajectory tracking error $\mathcal{E}_{\text{RMSE}}$ measures the deviation between the neural network’s commanded spatial path and the physical robot’s executed trajectory. As quantified in Table V, TPE-MARL achieves a remarkable tracking accuracy of 3.2 cm, fundamentally outperforming the Independent RL baseline (4.3 cm).

Crucially, this 3.2 cm accuracy is not merely an artifact of well-tuned PID controllers; it is a mathematical manifestation of the topology novelty reward $r_{\text{vis},t}$. By mathematically bounding the exploration and ensuring exhaustive visitation of the quotient space during training, the neural network avoids outputting highly non-linear or discontinuous control signals at the boundaries of its state distribution. The policy learns a globally *smooth action manifold*. This inherent stability dramatically reduces the actuation burden on the low-level controllers, leading to an exceptionally high behavioral consistency ($C_{\text{behav}} = 0.011$), an order of magnitude lower in variance compared to the jitter-prone Independent RL ($C_{\text{behav}} = 0.053$).

TABLE V: Physical execution accuracy and robustness metrics. The low variance in behavioral consistency ($C_{\text{behav}} = 0.011$) underscores the stability of the learned representations against physical execution noise.

Condition	Tracking Error	Behavioral Consistency
	$\mathcal{E}_{\text{RMSE}}$ (cm)	C_{behav}
Reactive rule-based	3.1	0.010
Independent RL	4.3	0.053
TPE-MARL	3.2	0.011

2) *Algorithmic Complexity and Real-Time Scalability:* To evaluate the algorithmic complexity in practice, we dissect the end-to-end latency (Δt_{E2E}) of the complete decision-execution loop. Operating on a standard edge-computing node, the TPE-MARL architecture realizes a single-step decision time ($\Delta t_{\text{decision}}$) of merely 18.5 ms, culminating in a total end-to-end latency of 45.2 ms for the entire perception-to-actuation pipeline. This ultra-low latency directly corroborates the theoretical computational advantages of our Information Bottleneck design introduced in Section IV.

Because the Game Topology Tensor projects continuous multi-agent states into a discrete quotient space via LSH (SimHash), the construction complexity is strictly bounded by $\mathcal{O}(N_t \cdot K \cdot m)$. Furthermore, the variational inference for the dual intrinsic reward utilizes lightweight MLP decoders, bounding the estimation complexity to $\mathcal{O}(N_t \cdot K \cdot d_z)$. By bypassing complex, computationally intractable geometric operators in the fully continuous joint state space, the algorithm achieves strict linear time complexity.

This mathematical scalability translates to immense practical feasibility. The sub-50 ms end-to-end latency comfortably satisfies the rigorous 100 ms control interval constraint of the Dec-POMDP framework. Moreover, it incurs minimal computational overhead—consuming only a fraction of the processing capacity on resource-constrained execution edges (e.g., 15% CPU utilization on a low-power Jetson Nano). This proves that the topology-enhanced coordination intelligence can be effortlessly scaled to decentralized real-world fleets without succumbing to the curse of dimensionality in computation.

VII. DISCUSSION

The preceding sections have theoretically formulated, methodologically constructed, and empirically validated the TPE-MARL framework. While the experimental validation is grounded in the complex, non-stationary dynamics of multi-vehicle collaborative decision-making, the theoretical implications of this work extend beyond the specific physical testbed. This section critically reflects on the universality of the proposed framework, delineates its fundamental theoretical boundaries, and outlines trajectories for future algorithmic evolution.

A. Universality of the Topological Quotient Space

The core theoretical contribution of this work is the abstraction of continuous multi-agent interactions into a discrete quotient space $\mathcal{S}_e / \sim_{\mathcal{T}}$. Historically, MARL architectures applied to spatial domains have heavily coupled their representations with the specific kinematic and physical constraints of the entities involved.

The Game Topology Tensor intentionally breaks this coupling. By formalizing the interactions purely through pairwise relational vectors and Locality-Sensitive Hashing (LSH), the framework becomes *domain-agnostic*. The dual intrinsic reward mechanism—maximizing marginal state entropy for breadth and conditional mutual information for depth—does not rely on the semantics of vehicular acceleration or steering. Consequently, the TPE-MARL paradigm is theoretically universally applicable to any decentralized multi-agent system characterized by spatial geometries and sparse coordination rewards, such as autonomous UAV swarms navigating adversarial environments or multi-robot fleets in dynamic warehouses. The physical zipper-merge verified in Section VI is merely one specific physical projection of a generalized high-dimensional dynamic negotiation.

B. Theoretical Boundaries: Topological Aliasing and Hash Collisions

Despite its empirical success, it is imperative to acknowledge the mathematical boundaries of the topological dimensionality reduction. The optimal policy invariance established in Conjecture 1 relies on the assumption that the equivalence relation $\sim_{\mathcal{T}}$ perfectly separates distinct game-theoretic intents.

As analyzed in Section V-E3, the SimHash code length m acts as an Information Bottleneck. While $m = 9$ proved optimal for the evaluated traffic densities, a theoretical limitation emerges as the localized agent density approaches infinity. Under extreme swarm congestion, the angular discretization inherent in LSH will inevitably force structurally distinct, critical game states to collide into identical hash representations. We define this pathology as *Topological Aliasing*. When severe aliasing occurs, the topology novelty reward $r_{\text{vis},t}$ will falsely truncate the exploration of a novel state, mistakenly identifying it as a highly visited equivalence class. Mitigating topological aliasing without exponentially expanding the quotient space dimensionality remains a fundamental challenge for hash-based MARL exploration.

C. Robustness Limits: Adversarial Topological Perturbations

The physical experiments in Section VI demonstrated the framework’s formidable OOD generalization against high-frequency physical noise, such as PID tracking errors and localization latencies. The spatial relaxation mechanism effectively preserves topological homomorphism during these transient continuous disturbances.

However, the current framework implicitly assumes that the foundational inputs to the topology tensor—the meta-observations—are structurally intact. If the system is subjected to structured, low-frequency perturbations—such as adversarial sensor spoofing or catastrophic occlusion where an entire interacting agent is deleted from the local field of view—the resulting topology tensor will be fundamentally corrupted. Under such adversarial topological perturbations, the variational Evidence Lower Bound (ELBO) utilized in the collaboration reward may generate erroneously high mutual information estimates, misleading the exploration policy.

Future extensions of this framework must transition from deterministic topological encoding to probabilistic topology graphs. By incorporating epistemic uncertainty quantification into the variational decoder, the dual intrinsic reward mechanism could be calibrated to down-weight the exploration bonuses in regions of the quotient space where the topological structure itself is mathematically uncertain, thereby achieving a mathematically provable resilience against adversarial structural perturbations.

VIII. CONCLUSION

In this paper, we presented TPE-MARL, a novel information-theoretic exploration framework designed to resolve the curse of dimensionality in continuous multi-agent cooperative decision-making. By abstracting the intractable continuous kinematic manifold into a mathematically rigorous discrete quotient space—via the Game Topology Tensor and

Locality-Sensitive Hashing—we systematically transformed an undirected search problem into a structured topological traversal.

The core methodological contribution lies in formulating multi-agent exploration as a joint information capacity optimization problem. Driven by a dual intrinsic reward mechanism, the framework seamlessly balances exploration breadth and depth: the topology novelty reward maximizes the marginal entropy of visited topological states, while the collaboration reward, anchored by a tractable variational Evidence Lower Bound (ELBO), minimizes conditional entropy to aggressively exploit strategically profound interactions. The integration of an adversarial Information Bottleneck rigorously guarantees that the learned representations encode only critical game-theoretic intents, actively decoupling them from redundant physical noise.

Extensive empirical evaluations corroborated the theoretical soundness of our approach. In highly non-stationary simulations, TPE-MARL achieved asymptotic policy invariance, converging to near-optimal decision distributions that tightly approximated the theoretical upper bounds established by an exact MCTS oracle. More crucially, validation on a physical cyber-physical testbed demonstrated exceptional OOD generalization. Empowered by a spatial relaxation mechanism that ensures continuous-discrete topological homomorphism, the learned decentralized policy successfully synthesized high-dimensional dynamic negotiations (i.e., the cooperative zipper-merge) entirely zero-shot, proving resilient against severe physical covariate shifts such as communication latency and actuation tracking errors.

Ultimately, this work transcends the specific application domain of vehicular coordination. By formally defining spatial multi-agent coordination as the information-guided exploration of topological equivalence classes, TPE-MARL provides a highly scalable, mathematically grounded paradigm for achieving emergent cooperative intelligence in general open and dynamic environments.

REFERENCES

- [1] J. Hao, T. Yang, H. Tang, C. Bai, J. Liu, Z. Meng, P. Liu, and Z. Wang, “Exploration in deep reinforcement learning: From single-agent to multiagent domain,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 8762–8782, 2024.
- [2] F. A. Oliehoek, C. Amato, et al., *A concise introduction to decentralized POMDPs*. Springer, 2016, vol. 1.
- [3] S. Zhu, J. Zhou, A. Chen, M. Bai, J. Chen, and J. Xu, “Maexp: A generic platform for rl-based multi-agent exploration,” in *Proc. IEEE Int. Conf. Robot. Autom.*, 2024, pp. 5155–5161.
- [4] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, “Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning,” in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 4295–4304.
- [5] S. Liu, D. Xing, P. Gu, X. Wang, B. An, and Y. Liu, “Solving homogeneous and heterogeneous cooperative tasks with greedy sequential execution,” in *Int. Conf. Learn. Represent.*, vol. 2024, 2024, pp. 1875–1891.
- [6] 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 *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24 611–24 624.
- [7] Q. Long, R. Li, M. Zhao, T. Gao, and D. Terzopoulos, “Inverse attention agents for multi-agent systems,” in *Int. Conf. Learn. Represent.*, vol. 2025, 2025, pp. 2776–2793.

- [8] J. Jeon, W. Kim, W. Jung, and Y. Sung, "MASER: Multi-agent reinforcement learning with subgoals generated from experience replay buffer," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, 17–23 Jul 2022, pp. 10041–10052.
- [9] I.-J. Liu, U. Jain, R. A. Yeh, and A. Schwing, "Cooperative exploration for multi-agent deep reinforcement learning," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 18–24 Jul 2021, pp. 6826–6836.
- [10] L. Zheng, J. Chen, J. Wang, J. He, Y. Hu, Y. Chen, C. Fan, Y. Gao, and C. Zhang, "Episodic multi-agent reinforcement learning with curiosity-driven exploration," in *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3757–3769.
- [11] D. Jarrett, C. Tallec, F. Althé, T. Mesnard, R. Munos, and M. Valko, "Curiosity in hindsight: Intrinsic exploration in stochastic environments," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, 23–29 Jul 2023, pp. 14780–14816.
- [12] Y. Jiang, Q. Liu, Y. Yang, X. Ma, D. Zhong, H. Hu, J. Yang, B. Liang, B. Xu, C. Zhang, and Q. Zhao, "Episodic novelty through temporal distance," in *Int. Conf. Learn. Represent.*, 2025.
- [13] C. Li, T. Wang, C. Wu, Q. Zhao, J. Yang, and C. Zhang, "Celebrating diversity in shared multi-agent reinforcement learning," in *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3991–4002.
- [14] X. Zeng, H. Peng, and A. Li, "Effective exploration based on the structural information principles," in *Adv. Neural Inf. Process. Syst.*, 2024.
- [15] C. Yu, N. Burgess, M. Sahani, and S. J. Gershman, "Successor-predecessor intrinsic exploration," in *Adv. Neural Inf. Process. Syst.*, 2023.
- [16] Z. Hu, Z. Zhang, H. Li, C. Chen, H. Ding, and Z. Wang, "Attention-guided contrastive role representations for multi-agent reinforcement learning," in *Int. Conf. Learn. Represent.*, vol. 2024, 2024, pp. 23 627–23 649.
- [17] Y. Du, L. Han, M. Fang, J. Liu, T. Dai, and D. Tao, "Liir: Learning individual intrinsic reward in multi-agent reinforcement learning," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
- [18] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proc. 34th Annu. ACM Symp. Theory Comput.*, 2002, pp. 380–388.
- [19] Y. Hu, J. Fu, and G. Wen, "Graph soft actor-critic reinforcement learning for large-scale distributed multirobot coordination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 665–676, 2025.
- [20] Z. Pu, H. Wang, Z. Liu, J. Yi, and S. Wu, "Attention enhanced reinforcement learning for multi agent cooperation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8235–8249, 2023.
- [21] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 09–15 Jun 2019, pp. 2961–2970.
- [22] Y. s. Liu, W. Wang, Y. Hu, J. Hao, X. Chen, and Y. Gao, "Multi-agent game abstraction via graph attention neural network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 05, 2020, pp. 7211–7218.
- [23] H. Zhang, J. Cheng, L. Zhang, Y. Li, and W. Zhang, "H2gcn: Hierarchical-hops graph neural networks for multi-robot exploration in unknown environments," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3435–3442, 2022.
- [24] Y. Han, L. Zhang, D. Meng, X. Hu, and Y. Lu, "Spformer: A transformer based drl decision making method for connected automated vehicles," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2024, pp. 1223–1230.
- [25] W. Du, S. Ding, C. Zhang, and Z. Shi, "Multiagent reinforcement learning with heterogeneous graph attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 6851–6860, 2023.
- [26] G. Bernádez, J. Suárez-Varela, A. López, X. Shi, S. Xiao, X. Cheng, P. Barlet-Ros, and A. Cabellos-Aparicio, "Magneto: A graph neural network-based multi-agent system for traffic engineering," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 2, pp. 494–506, 2023.
- [27] Y. Li, X. Zhang, T. Zeng, J. Duan, C. Wu, D. Wu, and X. Chen, "Task placement and resource allocation for edge machine learning: A gnn-based multi-agent reinforcement learning paradigm," *IEEE Trans. Parallel Distrib. Syst.*, vol. 34, no. 12, pp. 3073–3089, 2023.
- [28] J. Wang, Y. Jin, F. Ding, and C. Wei, "Causal-inspired multi-agent decision-making via graph reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–13, 2026.
- [29] T. Sun, L.-H. Fan, X.-M. Yu, and S. Ma, "Fst-gnn: Feedback space-time graph neural network model for networked multi-agent formation prediction with noise interference," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 6, pp. 5983–5994, 2024.
- [30] Z. Zhou, S. Wang, A. Huang, J. Lou, W. Tang, and D. Navarro-Alarcon, "Kitnet: A region-attention-activated trajectory predictor with hierarchical graph neural network in dynamic-mutant multi-agent system," *IEEE Trans. Intell. Transp. Syst.*, vol. 27, no. 2, pp. 1861–1877, 2026.
- [31] S. Khandelwal and L. Sigal, "Attentionrnn: A structured spatial attention mechanism," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, October 2019.
- [32] F. Zhang, B. Liu, K. Wang, V. Tan, Z. Yang, and Z. Wang, "Relational reasoning via set transformers: Provable efficiency and applications to marl," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 35 825–35 838.
- [33] L. Kong, J. Feng, H. Liu, D. Tao, Y. Chen, and M. Zhang, "Mag-gnn: Reinforcement learning boosted graph neural network," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 12000–12021.
- [34] H. KANG, T. Kwon, J. Park, and J. R. Morrison, "Learning np-hard multi-agent assignment planning using gnn: Inference on a random graph and provable auction-fitted q-learning," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 16 098–16 109.
- [35] M. Morris, T. D. Barrett, and A. Pretorius, "Universally expressive communication in multi-agent reinforcement learning," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 33 508–33 522.
- [36] Y. Zhang, H. Tong, Y. Xia, Y. Zhu, Y. Chi, and L. Ying, "Batch active learning with graph neural networks via multi-agent deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 8, 2022, pp. 9118–9126.
- [37] M. Lopes, T. Lang, M. Toussaint, and P.-y. Oudeyer, "Exploration in model-based reinforcement learning by empirically estimating learning progress," in *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
- [38] S. Tiwari, O. Gottesman, and G. D. Konidaris, "Geometry of neural reinforcement learning in continuous state and action spaces," in *Int. Conf. Learn. Represent.*, 2025.
- [39] R. Creus Castanyer, J. Romoff, and G. Berseth, "Improving intrinsic exploration by creating stationary objectives," in *Int. Conf. Learn. Represent.*, vol. 2024, 2024, pp. 26 379–26 409.
- [40] M. Yuan, B. Li, X. Jin, and W. Zeng, "Automatic intrinsic reward shaping for exploration in deep reinforcement learning," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, 23–29 Jul 2023, pp. 40 531–40 554.
- [41] S. P. Patankar, M. Ouellet, J. Cervino, A. Ribeiro, K. A. Murphy, and D. Bassett, "Intrinsically motivated graph exploration using network theories of human curiosity," in *Proc. 2nd Learn. Graphs Conf.*, vol. 231, 27–30 Nov 2024, pp. 23:1–23:15.
- [42] T. Li, K. Zhu, J. Li, and Y. Zhang, "Learning distinguishable trajectory representation with contrastive loss," in *Adv. Neural Inf. Process. Syst.*, 2024.
- [43] J. Li, K. Kuang, B. Wang, X. Li, F. Wu, J. Xiao, and L. Chen, "Two heads are better than one: A simple exploration framework for efficient multi-agent reinforcement learning," in *Adv. Neural Inf. Process. Syst.*, 2023.
- [44] J. Fan and W. Li, "DRIBO: Robust deep reinforcement learning via multi-view information bottleneck," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, 17–23 Jul 2022, pp. 6074–6102.
- [45] A. DiGiovanni and A. Tewari, "Balancing adaptability and non-exploitability in repeated games," in *Proc. 38th Conf. Uncertainty Artif. Intell.*, vol. 180, 01–05 Aug 2022, pp. 559–568.
- [46] E. Nisioti, E. Masquil, G. Hamon, and C. Moulin-Frier, "Autotelic reinforcement learning in multi-agent environments," in *Proc. 2nd Conf. Lifelong Learn. Agents*, vol. 232, 22–25 Aug 2023, pp. 137–161.
- [47] P. Li, J. Hao, H. Tang, Y. Zheng, and X. Fu, "RACE: Improve multi-agent reinforcement learning with representation asymmetry and collaborative evolution," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, 23–29 Jul 2023, pp. 19 490–19 503.
- [48] W. Wang, D. Ye, and Z. Lu, "Mutual-information regularized multi-agent policy iteration," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 2617–2635.
- [49] W. Z. Wang, A. Shih, A. Xie, and D. Sadigh, "Influencing towards stable multi-agent interactions," in *Proc. 5th Conf. Robot Learn.*, vol. 164, 08–11 Nov 2022, pp. 1132–1143.
- [50] Z. Ma, R. Wang, F.-F. Li, M. Bernstein, and R. Krishna, "Elign: Expectation alignment as a multi-agent intrinsic reward," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 8304–8317.
- [51] B. Dai, C. Zhu, B. Guo, and D. Wipf, "Compressing neural networks using the variational information bottleneck," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, 10–15 Jul 2018, pp. 1135–1144.
- [52] L. Ardizzone, R. Mackowiak, C. Rother, and U. Köthe, "Training normalizing flows with the information bottleneck for competitive generative classification," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7828–7840.
- [53] E. Z. Liu, A. Raghunathan, P. Liang, and C. Finn, "Decoupling exploration and exploitation for meta-reinforcement learning without

sacrifices,” in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 18–24 Jul 2021, pp. 6925–6935.

- [54] Y. L. Lo, B. Sengupta, J. Foerster, and M. Noukhovitch, “Learning multi-agent communication with contrastive learning,” in *Int. Conf. Learn. Represent.*, vol. 2024, 2024, pp. 51 029–51 046.
- [55] G. Yang, T. Yang, J. Qiao, Y. Wu, J. Huo, X. Chen, and Y. Gao, “Multi-agent reinforcement learning with communication-constrained priors,” in *Adv. Neural Inf. Process. Syst.*, 2025.
- [56] Y. Pan, Z. Liu, and H. Wang, “Wonder wins ways: Curiosity-driven exploration through multi-agent contextual calibration,” in *Adv. Neural Inf. Process. Syst.*, 2025.
- [57] M. Klissarov, P. D’Oro, S. Sodhani, R. Raileanu, P.-L. Bacon, P. Vincent, A. Zhang, and M. Henaff, “Motif: Intrinsic motivation from artificial intelligence feedback,” *arXiv preprint arXiv:2310.00166*, 9 2023.
- [58] H. Ma, Z. Luo, T. V. Vo, K. Sima, and T.-Y. Leong, “Highly efficient self-adaptive reward shaping for reinforcement learning,” in *Int. Conf. Learn. Represent.*, 2025.
- [59] N. Lichtlé, E. Vinitsky, M. Nice, R. Bhadani, M. Bunting, F. Wu, B. Piccoli, B. Seibold, D. B. Work, J. W. Lee, J. Sprinkle, and A. M. Bayen, “From sim to real: A pipeline for training and deploying traffic smoothing cruise controllers,” *IEEE Trans. Rob.*, vol. 40, pp. 4490–4505, 2024.
- [60] S. E. Bachoumas and P. Artemiadis, “Learning adversarial policies for swarm leader identification using a probing agent,” in *Proc. IEEE Int. Conf. Robot. Autom.*, 2025, pp. 1–7.
- [61] S. Li, J. Lin, X. Yang, H. Zhang, and P. Zhao, “Enhancing the policy generalization on ood tasks via latent variable distribution enhancement sampler,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 37, no. 3, pp. 1477–1491, 2026.
- [62] S. Huang, J. Hu, H. Chen, P. Cui, H. Piao, L. Sun, and B. Yang, “Generalizable causal reinforcement learning for out-of-distribution environments,” *IEEE Trans. Ind. Informat.*, vol. 21, no. 7, pp. 5514–5524, 2025.
- [63] Y. Han, L. Zhang, D. Meng, and Z. Zhang, “A differentiated reward method for reinforcement learning based multi-vehicle cooperative decision-making algorithms,” *arXiv preprint arXiv:2502.00352*, 2025.
- [64] —, “Hybrid differential reward: Combining temporal difference and action gradients for efficient multi-agent reinforcement learning in cooperative driving,” *arXiv preprint arXiv:2511.16916*, 2025.
- [65] A. Y. Ng, D. Harada, and S. Russell, “Policy invariance under reward transformations: Theory and application to reward shaping,” in *Proc. Int. Conf. Mach. Learn.*, vol. 99, 1999, pp. 278–287.
- [66] H. Tang, R. Houthoofd, D. Foote, A. Stooke, O. Xi Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel, “# exploration: A study of count-based exploration for deep reinforcement learning,” in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [67] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *arXiv preprint physics/0004057*, 2000.
- [68] Y. Han, L. Zhang, D. Meng, Z. Zhang, X. Hu, and S. Weng, “A value-based parallel update mcts method for multi-agent cooperative decision-making of connected and automated vehicles,” *IEEE Trans. Intell. Transp. Syst.*, vol. 27, no. 1, pp. 1400–1415, 2026.
- [69] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.
- [70] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [71] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.
- [72] C. Wu, A. R. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, “Flow: A modular learning framework for mixed autonomy traffic,” *IEEE Trans. Rob.*, vol. 38, no. 2, pp. 1270–1286, 2022.