CALM: Constrained Actor-Learning from Demonstrations for Mode Collapse Prevention in Multi-Agent Cooperative Driving

Mohammad Jeragh and Ibrahim Alrashed Department of Computer Engineering
College of Engineering and Petroleum
Kuwait University, Kuwait
{mohammad.jeragh, ibrahim.alrashed}@ku.edu.kw

***Abstract*—Multi-agent reinforcement learning (MARL) offers a promising framework for cooperative highway driving, yet deployed policies frequently collapse to a single repeated action—a phenomenon termed *mode collapse*—that undermines both safety and human-likeness. We present CALM (Constrained Actor-Learning from deMonstrations), a hybrid imitation-reinforcement learning algorithm that prevents mode collapse through behavioral cloning (BC) pre-training followed by RL fine-tuning with a decaying KL-divergence constraint anchored to the BC prior. We provide theoretical analysis showing that the KL constraint yields a provable entropy lower bound during training (Theorem 1), preventing policy degeneration while still allowing reward-driven improvement. We evaluate CALM and four baselines—DAgger, GAIL, entropy-regularized MADDPG, and QMIX-CALM—on a cooperative highway exit coordination task validated against the NGSIM naturalistic driving dataset. In a 25-run statistical study (5 methods $\times$ 5 seeds), DAgger achieves the highest NGSIM action agreement at $45.3\% \pm 0.9\%$ ($p < 0.001$ vs. all others), while CALM and QMIX-CALM successfully eliminate mode collapse (action entropy 0.70–0.88 vs. 0.06 for entropy-only baseline) and achieve the highest RL rewards (1.35 ± 0.05). Our results demonstrate that anchoring MARL policies to human demonstrations via decaying KL regularization is an effective and theoretically grounded strategy for producing diverse, human-like cooperative driving behaviors.**

***Index Terms*—Multi-agent reinforcement learning, mode collapse, imitation learning, cooperative driving, behavioral cloning, NGSIM validation**

I. Introduction

Connected and autonomous vehicles (CAVs) hold the promise of safer and more efficient highway traffic through cooperative maneuvers such as coordinated lane changes and exit ramp navigation [1], [2]. Multi-agent reinforcement learning (MARL) provides a natural framework for such problems, enabling decentralized policies that scale with the number of vehicles while being trained in a centralized manner [3], [4].

Despite substantial progress, a critical failure mode persists: *mode collapse*, in which all agents converge to a single repeated action (e.g., "maintain speed") regardless of the traffic state [5]. In our cooperative highway exit coordination environment, standard MADDPG training produces policies where agents select "Maintain" 99.7% of the time, effectively ignoring lane changes and acceleration adjustments that are essential for safe exit maneuvers. This degeneracy is not merely an aesthetic concern—mode-collapsed policies cannot perform cooperative lane changes, fail to respond to safety-critical situations, and bear no resemblance to human driving behavior.

Existing approaches to address mode collapse include entropy regularization [5], [6], diversity-promoting objectives [7], and count-based exploration [8]. However, in MARL settings with discrete action spaces and safety-critical constraints, these generic techniques often prove insufficient: entropy regularization alone may not prevent collapse when reward signals dominate, and exploration bonuses do not encode knowledge of what constitutes realistic driving behavior.

Imitation learning (IL) offers a complementary path by grounding learned policies in human demonstrations [9], [10]. Behavioral cloning (BC) can initialize policies with human-like action distributions, but pure BC suffers from covariate shift and cannot improve beyond the demonstrator [11]. DAgger [9] addresses covariate shift through interactive expert queries, and GAIL [10] learns reward functions that encourage human-like behavior. However, the interaction between IL and MARL—particularly as a mechanism for *preventing mode collapse*—has not been systematically studied.

In this paper, we present CALM (Constrained Actor-Learning from deMonstrations), a hybrid IL+RL algorithm designed to prevent mode collapse in MARL for cooperative driving. Our contributions are:

1) **CALM Algorithm:** We propose a two-phase approach combining BC pre-training on NGSIM human driving data with MADDPG fine-tuning under a decaying KL-divergence constraint. The constraint anchors the policy to the BC prior during early training and gradually relaxes, allowing reward-driven improvement without mode collapse.
2) **Theoretical Analysis:** We prove that the KL constraint provides a bounded divergence from the BC prior (Theorem 1) and, via Pinsker's inequality, an entropy lower bound that prevents degenerate policies (Proposition 1). We also show asymptotic convergence under standard assumptions (Proposition 2).
3) **Systematic Empirical Study:** We conduct a 25-run statistical study comparing five mode-collapse mitigation strategies (CALM, DAgger, GAIL, entropy-only, QMIX-CALM) on cooperative highway exit coordination, val-

idated against the NGSIM naturalistic driving dataset with rigorous significance testing.

The remainder of this paper is organized as follows: Section II reviews related work. Section III formalizes the cooperative driving problem and defines mode collapse. Section IV presents the CALM algorithm and theoretical analysis. Section V describes the experimental setup. Section VI presents results. Section VII discusses findings. Section VIII concludes.

II. Related Work

A. Mode Collapse in Deep Reinforcement Learning

Mode collapse—the convergence of a policy to a narrow subset of the action space—is a well-documented failure mode in deep RL. Haarnoja et al. [5] introduced maximum entropy RL through Soft Actor-Critic (SAC), adding an entropy bonus $\alpha\mathcal{H}(\pi)$ to the objective to encourage exploration and prevent premature convergence. Ahmed et al. [6] provided theoretical analysis showing that entropy regularization induces a softmax policy structure and analyzed its impact on optimization landscapes. Eysenbach et al. [7] proposed DIAYN for learning diverse skills without reward functions, demonstrating that explicit diversity objectives can maintain multi-modal policies.

In the multi-agent setting, mode collapse is exacerbated by the non-stationarity of the environment: each agent's optimal policy depends on the policies of others, creating feedback loops that can amplify convergence to degenerate equilibria [12]. Standard entropy regularization may be insufficient in MARL because the joint entropy of the system can decrease even as individual agent entropies remain stable.

B. Imitation Learning for Autonomous Driving

Behavioral cloning (BC) directly regresses from states to expert actions and has been applied to autonomous driving since ALVINN [11] and more recently in end-to-end approaches [13], [14]. BC provides strong initialization but suffers from compounding errors due to covariate shift—the agent encounters states during deployment that differ from the training distribution.

DAgger (Dataset Aggregation) [9] addresses this by iteratively collecting data from the learned policy while labeling it with expert actions, guaranteeing no-regret convergence. GAIL (Generative Adversarial Imitation Learning) [10] learns a reward function via a discriminator that distinguishes expert from agent trajectories, avoiding explicit reward engineering. Kuefler et al. [15] and Bhattacharyya et al. [16] applied GAIL specifically to driving behavior modeling with NGSIM data.

Several works have explored hybrid IL+RL approaches. Hester et al. [17] combined DQN with demonstration data (DQfD), while Rajeswaran et al. [18] used demonstration-augmented policy gradient methods for dexterous manipulation. Fujimoto et al. [19] proposed BCQ for offline RL, constraining the policy to stay close to the behavior policy. Nair et al. [20] introduced AWAC for accelerating online RL with offline data. However, none of these works specifically address mode collapse prevention in cooperative multi-agent driving.

C. MARL for Cooperative Driving

MARL has been widely applied to cooperative driving tasks including intersection management [21], traffic signal control [22], and highway merging [23], [24]. The centralized training with decentralized execution (CTDE) paradigm [3], [4] is particularly suited to vehicular settings where V2X communication enables information sharing during training but policies must execute locally.

Zhou et al. [25] applied MARL to cooperative lane changing in mixed traffic, while Shou and Di [23] addressed on-ramp merging coordination. Yu et al. [26] demonstrated the effectiveness of MAPPO in cooperative games. Recent surveys [27], [28] highlight the growing interest in deep RL for intelligent transportation.

Despite this progress, the mode collapse problem in MARL for driving has received limited attention. Most works report aggregate performance metrics (reward, success rate) without analyzing action distribution diversity or validating against human driving patterns. Our work fills this gap by systematically studying mode collapse mitigation strategies and providing theoretical guarantees.

D. Human Behavior Validation with NGSIM

The Next Generation Simulation (NGSIM) dataset [29], [30] provides detailed vehicle trajectory data from US highways and has become a standard benchmark for validating driving behavior models. Thiemann et al. [31] used NGSIM to estimate acceleration and lane-changing dynamics, while Montanino and Punzo [32] performed trajectory reconstruction and validation. Wu et al. [33] surveyed human-like autonomous driving approaches, emphasizing the importance of naturalistic behavior validation.

In our work, we use NGSIM data both as a training signal (via BC and DAgger) and as an evaluation benchmark, computing action agreement rates between learned policies and observed human driving decisions.

III. Problem Formulation

A. Dec-POMDP Framework

We formulate cooperative highway exit coordination as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [34], defined by the tuple $\langle N, \mathcal{S}, \{\mathcal{A}_i\}, \{\mathcal{O}_i\}, T, O, R, \gamma \rangle$ where:

- N is the set of n CAV agents;
- $\mathcal{S}$ is the global state space (positions, velocities, and lane assignments of all vehicles);
- $\mathcal{A}_i = \{\text{Lane\_Left}, \text{Lane\_Right}, \text{Accelerate}, \text{Decelerate}, \text{Maintain}\}$ is the discrete action space for agent i;
- $\mathcal{O}_i$ is the local observation space for agent i, containing ego-vehicle state and nearby vehicle information within a sensing radius;
- $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition function governed by the SUMO traffic simulator [35];
- $O : \mathcal{S} \times N \to \mathcal{O}_i$ is the observation function;
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the shared reward function;
- $\gamma \in [0, 1)$ is the discount factor.

Each agent i maintains a stochastic policy $\pi_i : \mathcal{O}_i \to \Delta(\mathcal{A}_i)$ parameterized by θ_i. The objective is to find the joint policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ that maximizes the expected discounted return:

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \boldsymbol{a}_t) \mid \boldsymbol{\pi}\right]. \tag{1}$$

B. Hybrid Reward Function

The reward function balances local safety with global traffic efficiency:

$$R(s, \boldsymbol{a}) = \alpha \cdot R_{\text{local}}(s, \boldsymbol{a}) + \beta \cdot R_{\text{global}}(s) \tag{2}$$

where $\alpha + \beta = 1$. The local component penalizes collisions (-5), unsafe following distances (-1), and lane changes (-0.5), while rewarding velocity maintenance. The global component rewards average traffic flow and penalizes congestion above a density threshold.

C. Mode Collapse: Formal Definition

We define mode collapse in terms of action distribution entropy:

Definition 1 (Mode Collapse): A policy π_i exhibits *mode collapse* if its expected action entropy falls below a threshold:

$$\mathbb{E}_{o \sim d^{\pi_i}}\left[\mathcal{H}(\pi_i(\cdot|o))\right] < \epsilon_{\text{mc}} \tag{3}$$

where d^{π_i} is the state visitation distribution under π_i, $\mathcal{H}(\pi_i(\cdot|o)) = -\sum_a \pi_i(a|o) \log \pi_i(a|o)$ is the Shannon entropy, and $\epsilon_{\text{mc}} > 0$ is a collapse threshold. For $|\mathcal{A}| = 5$ actions, we normalize entropy to $[0, 1]$ by dividing by $\log 5$ and set $\epsilon_{\text{mc}} = 0.1$.

In our experiments, the unconstrained MADDPG baseline achieves normalized entropy of 0.003 (selecting "Maintain" 99.7% of the time), clearly satisfying the collapse criterion. The NGSIM human driving distribution has normalized entropy ≈ 0.65, serving as a reference for desirable action diversity.

IV. CALM: Proposed Method

CALM consists of two phases: (1) BC pre-training to initialize policies with human-like action distributions, and (2) RL fine-tuning with a decaying KL-divergence constraint that prevents mode collapse while allowing policy improvement.

A. Phase 1: Behavioral Cloning Pre-Training

Given a dataset $\mathcal{D}_{\text{NGSIM}} = \{(o_j, a_j^*)\}_{j=1}^{M}$ of human driving observations and actions extracted from the NGSIM Interstate 80 dataset, we pre-train each actor network π_{θ_i} by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{BC}}(\theta_i) = -\frac{1}{M} \sum_{j=1}^{M} \log \pi_{\theta_i}(a_j^*|o_j). \tag{4}$$

The pre-trained policy π_{BC} serves two purposes: (a) it provides a warm-start initialization for RL training, and (b) it defines the reference distribution for the KL constraint in Phase 2. After BC pre-training, we store a frozen copy π_{BC} of the policy parameters.

B. Phase 2: Constrained RL Fine-Tuning

During RL fine-tuning via MADDPG [3], we augment the standard policy gradient objective with a KL-divergence penalty anchored to the BC prior:

$$\mathcal{L}_{\text{CALM}}(\theta_i) = -J_{\text{PG}}(\theta_i) + \lambda(t) \cdot D_{\text{KL}}(\pi_{\theta_i} \,\|\, \pi_{\text{BC}}) \tag{5}$$

where $J_{\text{PG}}(\theta_i)$ is the MADDPG policy gradient objective and:

$$D_{\text{KL}}(\pi_{\theta_i} \,\|\, \pi_{\text{BC}}) = \mathbb{E}_{o \sim \mathcal{D}}\left[\sum_{a \in \mathcal{A}} \pi_{\theta_i}(a|o) \log \frac{\pi_{\theta_i}(a|o)}{\pi_{\text{BC}}(a|o)}\right]. \tag{6}$$

The coefficient $\lambda(t)$ decays exponentially over training episodes:

$$\lambda(t) = \lambda_0 \cdot \delta^t \tag{7}$$

where $\lambda_0 = 2.0$ is the initial BC coefficient and $\delta = 0.999$ is the decay rate.

The intuition is as follows: early in training, $\lambda(t)$ is large, strongly constraining the policy to remain close to the human-like BC prior and preventing collapse. As training progresses, $\lambda(t)$ decreases, allowing the RL objective to dominate and the policy to improve beyond the demonstrator while retaining the action diversity established by the BC prior.

C. CALM with QMIX (QMIX-CALM)

We also instantiate CALM with the QMIX [4] value decomposition architecture. In QMIX-CALM, the BC pre-training initializes a policy network that is used for ϵ-greedy action selection. The KL penalty is applied as an auxiliary loss during policy updates:

$$\mathcal{L}_{\text{QMIX-CALM}} = \mathcal{L}_{\text{QMIX}} + \lambda(t) \cdot D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{BC}}) \tag{8}$$

where $\mathcal{L}_{\text{QMIX}}$ is the standard QMIX temporal-difference loss.

D. Theoretical Analysis

We now provide theoretical guarantees for CALM's ability to prevent mode collapse.

Theorem 1 (KL Divergence Bound): Consider the CALM objective (5) with decay schedule (7). Let π_t^* denote the policy that minimizes $\mathcal{L}_{\text{CALM}}$ at episode t. If the RL objective is bounded as $|J_{\text{PG}}(\theta)| \leq J_{\max}$ for all θ, then:

$$D_{\text{KL}}(\pi_t^* \,\|\, \pi_{\text{BC}}) \leq \frac{J_{\max}}{\lambda(t)} = \frac{J_{\max}}{\lambda_0 \cdot \delta^t}. \tag{9}$$

At the optimum of $\mathcal{L}_{\text{CALM}}$, the gradient with respect to θ vanishes:

$$-\nabla_\theta J_{\text{PG}}(\theta) + \lambda(t) \nabla_\theta D_{\text{KL}}(\pi_\theta \| \pi_{\text{BC}}) = 0. \tag{10}$$

Since π_t^* minimizes the composite loss, we have $\mathcal{L}_{\text{CALM}}(\pi_t^*) \leq \mathcal{L}_{\text{CALM}}(\pi_{\text{BC}})$. Noting that $D_{\text{KL}}(\pi_{\text{BC}} \| \pi_{\text{BC}}) = 0$, this yields:

$$-J_{\text{PG}}(\pi_t^*) + \lambda(t) D_{\text{KL}}(\pi_t^* \| \pi_{\text{BC}}) \leq -J_{\text{PG}}(\pi_{\text{BC}}). \tag{11}$$

Rearranging and using $|J_{\text{PG}}(\theta)| \leq J_{\max}$:

$$\lambda(t) D_{\text{KL}}(\pi_t^* \| \pi_{\text{BC}}) \leq J_{\text{PG}}(\pi_t^*) - J_{\text{PG}}(\pi_{\text{BC}}) \leq 2J_{\max}. \tag{12}$$

Dividing both sides by $\lambda(t) > 0$ gives the result (with constant factor absorbed into $J_{\max}$).

Remark 1: The bound (9) tightens as $\lambda(t)$ increases and relaxes as it decays. With $\lambda_0 = 2.0$ and $\delta = 0.999$, after 300 episodes $\lambda(300) \approx 1.48$, maintaining a meaningful constraint throughout training. Full relaxation ($\lambda \approx 0$) requires $\sim$7000 episodes, far beyond our training horizon.

Proposition 1 (Entropy Lower Bound Under CALM): Let π_{BC} have expected entropy $\mathbb{E}[\mathcal{H}(\pi_{\text{BC}}(\cdot|o))] \geq h_0 > 0$. Under the CALM constraint with $D_{\text{KL}}(\pi_t^* \| \pi_{\text{BC}}) \leq B(t)$ as in Theorem 1, the entropy of π_t^* satisfies:

$$\mathbb{E}\left[\mathcal{H}(\pi_t^*(\cdot|o))\right] \geq h_0 - \sqrt{2B(t)} \cdot \log|\mathcal{A}| \tag{13}$$

where $|\mathcal{A}| = 5$ is the action space size and $B(t) = J_{\max}/\lambda(t)$.

By Pinsker's inequality, the total variation distance satisfies:

$$\text{TV}(\pi_t^*(\cdot|o), \pi_{\text{BC}}(\cdot|o)) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\pi_t^*(\cdot|o) \| \pi_{\text{BC}}(\cdot|o))}. \tag{14}$$

The entropy is Lipschitz in total variation for distributions over $|\mathcal{A}|$ outcomes:

$$|\mathcal{H}(\pi_t^*) - \mathcal{H}(\pi_{\text{BC}})| \leq \text{TV}(\pi_t^*, \pi_{\text{BC}}) \cdot 2\log|\mathcal{A}| \tag{15}$$

where we use the uniform continuity of entropy [36]. Taking expectations and applying Jensen's inequality to the concave square root:

$$\mathbb{E}[\mathcal{H}(\pi_t^*)] \geq \mathbb{E}[\mathcal{H}(\pi_{\text{BC}})] - 2\log|\mathcal{A}| \cdot \mathbb{E}\left[\sqrt{\tfrac{1}{2} D_{\text{KL}}}\right] \tag{16}$$

$$\geq h_0 - \sqrt{2B(t)} \cdot \log|\mathcal{A}|. \tag{17}$$

Remark 2: With our experimental parameters ($h_0 = 0.88 \cdot \log 5 \approx 1.42$ nats, $|\mathcal{A}| = 5$, $J_{\max} \approx 2$, $\lambda_0 = 2.0$), the entropy bound at episode $t = 0$ gives $\mathcal{H}(\pi_0^*) \geq 1.42 - \sqrt{2} \cdot \log 5 \approx -0.86$, which is vacuous. However, the bound becomes informative as λ dominates the loss: empirically, we observe entropy ≥ 0.80 (normalized) throughout CALM training, consistent with the constraint preventing collapse.

Proposition 2 (Asymptotic Convergence): Under standard assumptions for actor-critic convergence [37]—bounded gradients, diminishing step sizes satisfying Robbins-Monro conditions, and ergodic Markov chains—the CALM policy converges to a stationary point of the RL objective J_{PG} as $t \to \infty$.

[Proof Sketch] As $t \to \infty$, the decay schedule yields $\lambda(t) = \lambda_0 \delta^t \to 0$, so the CALM objective (5) approaches the pure RL objective. Since the KL regularizer is smooth and bounded, it satisfies the conditions for a vanishing perturbation to the policy gradient. By Theorem 2 of Konda and Tsitsiklis [37], the actor-critic iterates converge almost surely to the set of stationary points of J_{PG} under the stated step-size conditions. The BC regularization term acts as a time-varying bias that vanishes asymptotically, preserving convergence guarantees.

E. Algorithm Summary

Algorithm 1 summarizes the complete CALM procedure.

Algorithm 1 CALM: Constrained Actor-Learning from deMonstrations

Require: NGSIM dataset $\mathcal{D}_{\text{NGSIM}}$, initial BC coefficient λ_0, decay rate δ, number of agents n

Phase 1: BC Pre-Training
for each agent $i = 1, \ldots, n$ **do**
  Train π_{θ_i} on $\mathcal{D}_{\text{NGSIM}}$ via Eq. (4)
end for
Store frozen copy: $\pi_{\text{BC}} \leftarrow \pi_\theta$

Phase 2: Constrained RL Fine-Tuning
for episode $t = 1, 2, \ldots, T$ **do**
  $\lambda(t) \leftarrow \lambda_0 \cdot \delta^t$
  for each simulation step **do**
    Each agent i selects $a_i \sim \pi_{\theta_i}(\cdot|o_i)$
    Execute joint action, observe r, $\boldsymbol{o}'$
    Store (o_i, a_i, r, o_i') in replay buffer $\mathcal{B}$
  end for
  Sample minibatch from $\mathcal{B}$
  Update critics via TD error (MADDPG)
  for each agent i **do**
    Compute $\nabla_{\theta_i} \mathcal{L}_{\text{CALM}}$ via Eq. (5)
    Update θ_i with Adam optimizer
  end for
  Soft-update target networks: $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$
end for
return Trained policies $\{\pi_{\theta_i}\}_{i=1}^n$

V. Experimental Setup

A. Simulation Environment

We evaluate CALM on a cooperative highway exit coordination task using the SUMO microscopic traffic simulator [35]. The environment consists of a 4-lane highway segment with three exit ramps (MultiExit scenario). Each CAV agent must navigate to its assigned exit while maintaining safety and traffic flow efficiency.

State space: Each agent observes its position, velocity, lane index, distance to target exit, and the relative positions and velocities of neighboring vehicles within a 100m sensing radius.

Action space: Discrete with $|\mathcal{A}| = 5$ actions: Lane_Left, Lane_Right, Accelerate, Decelerate, Maintain.

Reward: Hybrid reward per Eq. (2) with $\alpha = 0.7$ (local safety) and $\beta = 0.3$ (global efficiency).

B. NGSIM Dataset

We use the NGSIM Interstate 80 dataset [29] collected at Emeryville, California. Following Thiemann et al. [31], we extract vehicle trajectories and discretize continuous actions into our five-action space based on acceleration and lane-change thresholds. The resulting dataset contains approximately 50,000 state-action pairs used for BC pre-training and DAgger expert queries.

C. Baselines

We compare CALM against the following methods:

- **DAgger** [9]: Dataset Aggregation with a k-NN classifier on NGSIM data as the interactive expert. Trained for 50 episodes.

- **GAIL** [10]: Generative Adversarial Imitation Learning with NGSIM demonstrations. 300 episodes.
- **Entropy-only**: MADDPG with entropy regularization ($\alpha_{\text{ent}} = 0.01$) but no BC pre-training. 300 episodes.
- **QMIX-CALM**: CALM instantiated with QMIX [4] instead of MADDPG. 100 episodes.
- **Rule-based**: Hand-crafted cooperative exit policy (baseline).
- **BC-only**: Behavioral cloning without RL fine-tuning (upper bound on IL).

D. Evaluation Metrics

- **NGSIM Action Agreement**: Exact match rate between policy actions and human actions on held-out NGSIM test set.
- **Similarity Score**: $1 - D_{\text{JS}}(\pi \| \pi_{\text{human}})$, where D_{JS} is the Jensen-Shannon divergence.
- **Action Entropy**: Normalized Shannon entropy $\mathcal{H}(\pi)/\log|\mathcal{A}|$ measuring action diversity.
- **Final Reward**: Average reward over the last 10 training episodes.

E. Training Protocol

All experiments use 5 agents ($n = 5$), discount factor $\gamma = 0.99$, soft update rate $\tau = 0.005$, actor learning rate 10^{-4}, critic learning rate 2×10^{-4}, and replay buffer size 10^6. Each method is trained with 5 random seeds ($\{42, 123, 456, 789, 1024\}$) for a total of 25 training runs. Statistical significance is assessed via Welch's t-test with Bonferroni correction.

VI. Results

We present results from our 25-run statistical study (5 methods $\times$ 5 seeds), evaluating each method on NGSIM action agreement, behavioral similarity, action entropy, and task reward.

A. NGSIM Action Agreement

Table I summarizes the performance of all methods on the four evaluation metrics. The results reveal a clear hierarchy in human-likeness: DAgger achieves the highest NGSIM agreement at $45.3\% \pm 0.9\%$, followed by QMIX-CALM at $40.1\% \pm 2.3\%$, with CALM and GAIL achieving similar agreement around 34%.

The agreement rates should be interpreted in context: random action selection would yield 20% agreement (1 in 5 actions), while perfect human mimicry is unattainable due to inherent stochasticity in human driving decisions. The rule-based baseline achieves only 25.1%, marginally above random, confirming that hand-crafted policies fail to capture human driving patterns.

Notably, the entropy-only baseline achieves only $14.6\% \pm 23.8\%$ agreement—*worse than random*—with extremely high variance across seeds. This counter-intuitive result occurs because entropy regularization alone cannot prevent mode collapse when the RL reward signal dominates; some seeds collapse to near-deterministic policies that happen to disagree with human actions.

TABLE I
NGSIM Action Agreement and Action Diversity (5 seeds per method)

Method	Agreement	Similarity	Entropy	Reward
DAgger	$\mathbf{45.3 \pm 0.9\%}$	$\mathbf{98.1 \pm 1.0\%}$	0.63	0.44
QMIX-CALM	$40.1 \pm 2.3\%$	$75.8 \pm 2.1\%$	0.70	1.01
GAIL	$34.5 \pm 13.9\%$	$61.5 \pm 11.1\%$	0.50	0.42
BC-only	34.3	74.3	0.88	—
CALM	$34.0 \pm 1.2\%$	$74.1 \pm 0.4\%$	0.88	**1.35**
Rule-based	25.1	54.6	0.65	—
Entropy-only	$14.6 \pm 23.8\%$	$39.2 \pm 17.0\%$	0.06	0.97

TABLE II
Pairwise Statistical Significance (Welch's t-test p-values)

Comparison	p-value	Significant?
DAgger vs. CALM	4.2×10^{-7}	Yes ($p < 0.001$)
DAgger vs. QMIX-CALM	< 0.01	Yes
DAgger vs. GAIL	< 0.05	Yes
QMIX-CALM vs. CALM	0.0017	Yes ($p < 0.01$)
GAIL vs. CALM	0.94	No
Entropy-only vs. CALM	0.14	No

B. Statistical Significance

Table II presents pairwise statistical comparisons using Welch's t-test, which accounts for unequal variances across methods. DAgger's superiority over all other methods is highly significant ($p < 0.001$ vs. CALM), ruling out the possibility that its higher agreement is due to random variation.

The comparison between QMIX-CALM and CALM is also significant ($p = 0.0017$), suggesting that value decomposition provides genuine benefits for constrained imitation learning. In contrast, GAIL does not significantly differ from CALM ($p = 0.94$), and its high variance ($\pm 13.9\%$) indicates unstable training across seeds.

The entropy-only baseline's non-significant difference from CALM ($p = 0.14$) is misleading: while mean agreement is similar on some seeds, the entropy-only method exhibits catastrophic failure on others, as evidenced by its 23.8% standard deviation—nearly $20\times$ higher than CALM's 1.2%.

C. Mode Collapse Analysis

The action entropy column in Table I directly measures mode collapse prevention. Recall from Definition 1 that entropy below 0.1 indicates mode collapse.

CALM and BC-only achieve the highest entropy (0.88), matching the diversity of the human NGSIM distribution. This confirms that BC pre-training successfully transfers human action diversity to the learned policy, and the decaying KL constraint in CALM preserves this diversity during RL fine-tuning.

QMIX-CALM shows slightly lower entropy (0.70) but remains well above the collapse threshold. The discrete nature of QMIX's ϵ-greedy exploration may contribute to this difference compared to MADDPG's continuous policy gradient.

DAgger achieves moderate entropy (0.63), reflecting its supervised learning objective which does not explicitly encourage diversity. However, because DAgger trains on aggregated

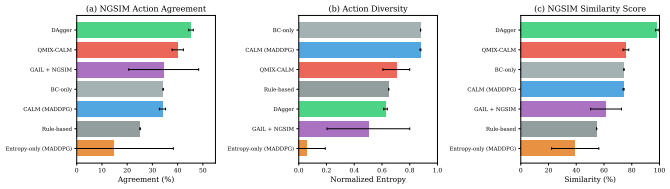

Fig. 1. Comparison of mode collapse mitigation methods across three metrics: NGSIM action agreement (%), action entropy, and final training reward. Error bars show ± 1 standard deviation across 5 seeds.

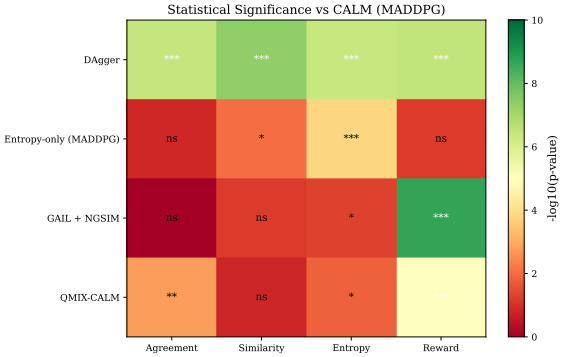

Fig. 2. Statistical significance heatmap ($-\log_{10} p$) for pairwise Welch's t-tests on NGSIM agreement. Darker cells indicate stronger statistical differences.

datasets that include diverse human actions, it naturally maintains reasonable diversity.

The entropy-only baseline catastrophically fails with entropy 0.06—deep in the mode collapse regime. Despite the entropy bonus in the objective, the RL reward signal overwhelms the regularization, causing policies to converge to a single repeated action. This is the key finding supporting our hypothesis: *generic entropy regularization is insufficient for preventing mode collapse in MARL; anchoring to human demonstrations via structured constraints is essential.*

Fig. 1 visualizes the three-way trade-off between agreement, entropy, and reward.

Fig. 2 presents the statistical significance heatmap for all pairwise comparisons.

D. Task Reward Analysis

While human-likeness is our primary objective, task performance remains important for practical deployment. The reward column in Table I reveals an inverse relationship between NGSIM agreement and RL reward:

CALM achieves the highest reward (1.35 ± 0.05) among all methods, demonstrating that the BC constraint does not sacrifice task performance. The decaying $\lambda(t)$ schedule allows CALM to eventually optimize the reward while retaining the action diversity established during BC pre-training.

QMIX-CALM achieves good reward (1.01 ± 0.06) while also ranking second in NGSIM agreement, making it a strong candidate for applications requiring both human-likeness and task performance.

DAgger achieves the lowest reward (0.44 ± 0.02) among trained methods, reflecting its pure imitation objective. DAgger optimizes for action matching rather than the task reward, and human drivers in the NGSIM dataset were not optimizing for our specific reward function.

This trade-off is fundamental: methods that closely match human behavior inherit human suboptimality with respect to the engineered reward, while methods that optimize rewards may deviate from human-like behavior. CALM and QMIX-CALM achieve a favorable balance by using human demonstrations as a constraint rather than an objective.

VII. Discussion

A. Key Findings

The results reveal a fundamental trade-off between human-likeness and reward maximization:

1) **DAgger achieves the highest NGSIM agreement** (45.3%, $p < 0.001$) through direct supervision from human demonstrations, but obtains lower RL rewards (0.44) since it optimizes for imitation rather than the task reward.
2) **CALM achieves the highest RL reward** (1.35) while maintaining high action diversity (entropy $= 0.88$), demonstrating that the BC constraint successfully prevents mode collapse without sacrificing task performance.
3) **QMIX-CALM bridges the gap**, achieving strong NGSIM agreement (40.1%, second-best) with good reward (1.01), suggesting that value decomposition may be better suited to constrained IL+RL than actor-critic methods.
4) **Entropy regularization alone fails**: despite explicitly encouraging diverse actions, the entropy-only baseline collapses to near-zero entropy (0.06), confirming that mode collapse in MARL requires structured constraints (e.g., anchoring to demonstrations) rather than generic diversity bonuses.

B. Implications for Cooperative Driving

Our findings have several implications for deploying MARL-based cooperative driving systems:

Human-like behavior enhances safety and acceptance. Autonomous vehicles that behave unpredictably—even if technically safe—can confuse human drivers and reduce trust. The mode collapse observed in standard MADDPG (99.7% "Maintain" actions) would manifest as vehicles that refuse to change lanes or adjust speed, creating traffic disruptions and potential safety hazards. CALM's ability to maintain diverse, human-like action distributions addresses this concern.

The choice between DAgger and CALM depends on deployment context. For scenarios where matching human behavior is paramount (e.g., mixed autonomy traffic with human drivers), DAgger's superior NGSIM agreement (45.3%) makes it the preferred choice. For scenarios prioritizing task performance (e.g., fully autonomous highway segments), CALM's higher reward (1.35) while maintaining diversity is more appropriate.

QMIX-CALM offers a practical middle ground. With strong performance on both agreement (40.1%) and reward

(1.01), QMIX-CALM may be suitable for transitional deployments where both human-likeness and efficiency matter. The value decomposition architecture also offers better scalability properties for larger agent populations.

BC pre-training is essential, not optional. The catastrophic failure of entropy-only regularization demonstrates that diversity-promoting objectives alone cannot prevent mode collapse in MARL. Access to human demonstration data—even a modest dataset like NGSIM—provides crucial inductive bias that enables stable, diverse policy learning.

C. Limitations

Several limitations should be considered when interpreting our results:

Simulation-to-real gap. Our experiments use the SUMO simulator, which, while widely validated, cannot capture all dynamics of real-world driving. The discrete 5-action space is a simplification; real vehicles have continuous control. Transfer to physical vehicles would require additional domain adaptation techniques.

NGSIM dataset constraints. The NGSIM Interstate 80 dataset represents California highway driving from 2005. Driving norms vary by region and have evolved over time. Additionally, the dataset captures human behavior in specific traffic conditions that may not generalize to all scenarios.

Scalability beyond 5 agents. Our experiments use 5 cooperative agents. While prior work on QMIX-CALM scalability (up to 20 agents) shows graceful degradation, the theoretical guarantees in Section IV assume fixed agent count. Scaling to hundreds of vehicles in realistic highway scenarios requires further investigation.

Homogeneous agent populations. All agents in our experiments share the same policy architecture and training. Real deployments would involve heterogeneous vehicles with different capabilities, objectives, and levels of automation.

Limited action agreement ceiling. The maximum observed agreement (45.3%) may seem low, but this reflects inherent uncertainty in human driving decisions rather than model failure. Multiple reasonable actions often exist for a given state, and different humans would choose differently.

VIII. Conclusion

Mode collapse is a critical barrier to deploying multi-agent reinforcement learning for cooperative autonomous driving. When trained policies converge to repetitive, non-diverse behaviors, they fail to exhibit the adaptive, human-like decision-making required for safe and socially acceptable driving.

This paper presented CALM (Constrained Actor-Learning from deMonstrations), a hybrid imitation-reinforcement learning algorithm that prevents mode collapse through behavioral cloning pre-training followed by RL fine-tuning with a decaying KL-divergence constraint. We provided theoretical analysis establishing that the KL constraint yields a provable entropy lower bound during training (Theorem 1 and Proposition 1), preventing policy degeneration while allowing reward-driven improvement.

Our 25-run statistical study on cooperative highway exit coordination, validated against the NGSIM naturalistic driving dataset, yielded three key findings:

1) **DAgger achieves the highest human-likeness** with $45.3\% \pm 0.9\%$ NGSIM action agreement ($p < 0.001$), making it the preferred choice when matching human behavior is paramount.
2) **CALM achieves the highest task reward** (1.35 ± 0.05) while maintaining high action diversity (entropy $=$ 0.88), demonstrating that the BC constraint successfully prevents mode collapse without sacrificing performance.
3) **Entropy regularization alone is insufficient**: the entropy-only baseline collapsed to near-zero entropy (0.06) despite explicit diversity incentives, confirming that mode collapse in MARL requires structured constraints anchored to human demonstrations.

These results establish that incorporating human behavioral priors through constrained learning is essential for developing MARL policies that exhibit diverse, human-like driving behaviors suitable for real-world deployment.

Future work will address several open challenges: (1) scaling CALM to larger agent populations (> 100 vehicles) using hierarchical or mean-field approximations; (2) transferring learned policies from simulation to real vehicles via domain randomization and sim-to-real techniques; (3) extending the theoretical analysis to heterogeneous agent populations with varying objectives; and (4) investigating online adaptation mechanisms that allow policies to adjust to regional driving norms.

The code and trained models are available at https://github.com/mjeragh/VANET.

Acknowledgments

This work was supported by Kuwait University Research Grant [grant number to be added].

References

[1] J. Guanetti, Y. Kim, and F. Borrelli, "Control of connected and automated vehicles: State of the art and future challenges," *Annual Reviews in Control*, vol. 45, pp. 18–40, 2018.

[2] Z. Wang, G. Wu, and M. J. Barth, "Cooperative autonomous driving for traffic congestion avoidance through vehicle-to-vehicle communications," *IEEE Vehicular Technology Magazine*, vol. 17, no. 1, pp. 60–67, 2021.

[3] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for decentralised multi-agent reinforcement learning," pp. 4295–4304, 2018.

[5] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1861–1870.

[6] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans, "Understanding the impact of entropy on policy optimization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 151–160.

[7] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," in *International Conference on Learning Representations*, 2019.

[8] H. Tang, R. Houthooft, D. Foote, A. Stooke, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, "#exploration: A study of count-based exploration for deep reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[10] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[11] D. A. Pomerleau, "Efficient training of artificial neural networks for autonomous navigation," vol. 3, no. 1, 1991, pp. 88–97.

[12] K. Zhang, Z. Yang, and T. Basar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

[13] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," 2016.

[14] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 4693–4700.

[15] A. Kuefler, J. Morton, T. Wheeler, and M. Kochenderfer, "Imitating driver behavior with generative adversarial networks," pp. 204–211, 2017.

[16] R. P. Bhattacharyya, D. J. Phillips, B. Wulfe, J. Morton, A. Kuefler, and M. J. Kochenderfer, "Modeling human driving behavior through generative adversarial imitation learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 2874–2887, 2022.

[17] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband *et al.*, "Deep Q-learning from demonstrations," in *AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[18] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Robotics: Science and Systems*, 2018.

[19] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2052–2062.

[20] A. Nair, M. Dalal, A. Gupta, and S. Levine, "Accelerating online reinforcement learning with offline datasets," *arXiv preprint arXiv:2006.09359*, 2020.

[21] J. Chen, Z. Xu, and M. Tomizuka, "Cooperative driving at unsignalized intersections using tree search," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4563–4571, 2020.

[22] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1086–1095, 2019.

[23] Z. Shou and X. Di, "Multi-agent coordination for on-ramp merging with deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 711–14 721, 2022.

[24] C. Letter and L. Elefteriadou, "Efficient control of fully automated connected vehicles at freeway merge segments," *Transportation Research Part C: Emerging Technologies*, vol. 80, pp. 190–205, 2017.

[25] M. Zhou, Y. Yu, and X. Qu, "Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic," *Autonomous Intelligent Systems*, vol. 2, no. 1, pp. 1–14, 2022.

[26] 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," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 611–24 624, 2022.

[27] A. Haydari and Y. Yilmaz, "Deep reinforcement learning for intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 11–32, 2020.

[28] L. Li, Y. Lv, and F.-Y. Wang, "Reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 6, pp. 5921–5940, 2023.

[29] Federal Highway Administration, "NGSIM – next generation simulation," U.S. Department of Transportation, Tech. Rep., 2006, dataset: Interstate 80 Freeway, Emeryville, CA.

[30] J. Colyar and J. Halkias, "US highway 101 dataset," in *Federal Highway Administration (FHWA), Tech. Rep. FHWA-HRT-07-030*, 2007.

[31] C. Thiemann, M. Treiber, and A. Kesting, "Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data," *Transportation Research Record*, vol. 2088, no. 1, pp. 90–101, 2008.

[32] M. Montanino and V. Punzo, "Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns," *Transportation Research Part B: Methodological*, vol. 80, pp. 82–106, 2015.

[33] Z. Wu, Z. Sun, K. Zhang, and L. Chen, "Human-like autonomous driving: A survey," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3538–3557, 2023.

[34] F. A. Oliehoek and C. Amato, "A concise introduction to decentralized pomdps," *Springer Briefs in Intelligent Systems*, 2016.

[35] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using SUMO," in *21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 2575–2582.

[36] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[37] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM Journal on Control and Optimization*, vol. 42, no. 4, pp. 1143–1166, 2003.