Robotics: Science and Systems 2024
Delft, Netherlands, July 15-July 19, 2024

Blending Data-Driven Priors in Dynamic Games

Justin Lidard[1,*], Haimin Hu[2,*], Asher Hancock[1,†], Zixu Zhang[2,†],
Albert Gimó Contreras[2], Vikash Modi[2], Jonathan A. DeCastro[3], Deepak Gopinath[3],
Guy Rosman[3], Naomi Ehrich Leonard[1], María Santos[1], and Jaime Fernández Fisac[2]

***Abstract*—As intelligent robots like autonomous vehicles become increasingly deployed in the presence of people, the extent to which these systems should leverage model-based game-theoretic planners versus data-driven policies for safe, interaction-aware motion planning remains an open question. Existing dynamic game formulations assume all agents are task-driven and behave optimally. However, in reality, humans tend to deviate from the decisions prescribed by these models, and their behavior is better approximated under a *noisy-rational* paradigm. In this work, we investigate a principled methodology to blend a data-driven *reference policy* with an optimization-based game-theoretic policy. We formulate *KLGame*, an algorithm for solving non-cooperative dynamic game with Kullback-Leibler (KL) regularization with respect to a general, stochastic, and possibly multi-modal reference policy. Our method incorporates, for each decision maker, a tunable parameter that permits *modulation* between task-driven and data-driven behaviors. We propose an efficient algorithm for computing multi-modal approximate feedback Nash equilibrium strategies of KLGame in real time. Through a series of simulated and real-world autonomous driving scenarios, we demonstrate that KLGame policies can more effectively incorporate guidance from the reference policy and account for noisily-rational human behaviors versus non-regularized baselines. Website with additional information, videos, and code: https://kl-games.github.io/.**

I. Introduction

Planning safe trajectories in the presence of multiple decision-making agents is a long-standing challenge in robotics. Desired performance and safety behaviors can often be encapsulated in mathematical models and algorithms that govern the operation of the agents, offering analytically tractable guarantees. However, first-principles planners may fall short in capturing the nuanced variety of scenarios that are subject to arise in interactive settings. In turn, data-driven priors can be extremely informative in these situations, with *policy alignment* strategies allowing agents to incorporate fine-tuned, scenario-specific heuristics as well as human-desired intent into their motion planning.

Two diverging approaches have recently emerged for planning in the presence of nearby agents. Dynamic game theory [1], which models interactions as a multi-agent optimal control problem, allows computing a task-optimal plan simultaneously for all agents in the scene. However, while these methods can effectively capture high-level rational behaviors, other agents do not always act rationally—i.e., they do not necessarily pick the game-theoretically optimal action. Moreover, gradient-based game-theoretic solvers may struggle to model *multi-modal*, *mixed-strategy* (stochastic) interactions, wherein distinct high-level behaviors may arise from the same objectives, leading to several local optima in the space of rewards. Conversely, behavior cloning methods provide a framework for end-to-end learning of multi-modal trajectory distributions from data, but may yield suboptimal or unsafe

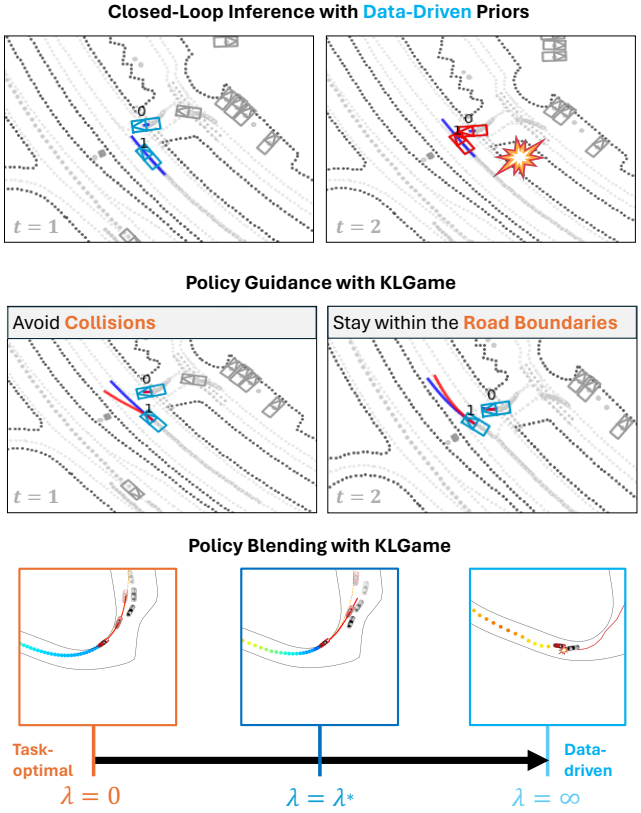

Fig. 1: Our proposed approach can seamlessly integrate data-driven policies with optimization-based dynamic game solutions. *Top:* Data-driven behavior predictors, such as transformer-based methods, provide *marginal* (i.e., single-agent) priors for human motion prediction, but may struggle to model *closed-loop*, multi-agent interactions. *Middle:* Our proposed method, KLGame, allows a robot to incorporate *guidance* from data-driven rollouts while performing online game-theoretic planning in closed-loop. *Bottom:* KLGame uses a *tunable* parameter λ that modulates behaviors on a spectrum: $\lambda = 0$ gives a deterministic dynamic game (*task-optimal*), and $\lambda \to \infty$ gives multi-modal behavior cloning. We call this tunability *policy blending*.

[1]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08540, USA
[2]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08540, USA
[3]Toyota Research Institute, Cambridge, MA 02139, USA
This research has been supported in part by an NSF Graduate Research Fellowship. This work is partially supported by Toyota Research Institute (TRI). It, however, reflects solely the opinions and conclusions of its authors and not TRI or any other Toyota entity.
*J. Lidard and H. Hu contributed equally.
†A. Hancock and Z. Zhang contributed equally.

predictions. Such data-driven methods exploit recent advances in autonomous vehicle perception systems by synthesizing high-dimensional sensor inputs, including a road map, traffic light states, and LIDAR data. While imitation learning can provide a flexible and efficient approximation for interactions modeled by dynamic game theory, predictions may be suboptimal or unsafe.

Integrated prediction and planning approaches [2, 3] are a popular middle-ground approach that combines first principles from game-theoretic planning with data-driven models derived from human motion. In this paradigm, a deep-learned motion forecasting system first produces a trajectory prediction before planning begins, reducing the computational load on the planner. However, substantially different behaviors may arise depending on whether agents adhere more strongly to the data-driven or optimal behavior. Fig. 1 provides an example of a multi-modal interaction where a human neglects an oncoming vehicle due to a visual occlusion. Since the vehicle's goal is near the human, the human's safety is directly in conflict with the target of the vehicle. Furthermore, the most likely motion forecast for both agents predicts that both will go straight, necessitating intervention to avoid a collision. Neither the game-theoretic nor imitation-learned plan provides a safe trajectory for both agents, calling for another method.

In this work, we introduce KLGame, a novel multi-modal game-theoretic planning framework that incorporates data-driven policy models (such as motion forecasting) into each agent's policy optimization through a regularization bonus in the planning objective. KLGame incentivizes the trajectories of planning agents to not only optimize hand-crafted cost heuristics, but also to adhere to a *reference policy*. In this work, we assume that the reference policy is distilled from data or expert knowledge, is stochastic, and may be multi-modal in general. In contrast to other integrated prediction and planning methods, KLGame (i) provides an analytically and computationally sound methodology for planning under strategy uncertainty, exactly solving the regularized stochastic optimization problem; and (ii) incorporates *tunable* multi-modal, data-driven motion predictions in the optimal policy through a scalar parameter, allowing the planner to modulate between purely data-driven and purely optimal behaviors.

Statement of contributions. We make two key contributions:

- We introduce *KLGame*, a novel stochastic dynamic game that blends *interaction-aware* task optimization with *closed-loop* policy guidance via Kullback-Leibler (KL) regularization. We provide an in-depth analysis in the linear-quadratic (LQ) setting with Gaussian reference policies and show KLGame permits an analytical *global feedback Nash equilibrium*, which naturally generalizes the solution of the maximum-entropy game [4].
- We propose an efficient and scalable trajectory optimization algorithm for computing approximate feedback Nash equilibria of KLGame with general nonlinear dynamics, costs, and *multi-modal* reference policies. Experimental results on Waymo's Open Motion Dataset demonstrate the efficacy of KLGame in leveraging data-driven priors compared to state-of-the-art methods [4]–[6].

II. Related Work

Our work relates to recent advances in game-theoretic motion planning, stochastic optimal control, and multi-agent trajectory prediction.

A. Game-Theoretic Motion Planning

Game-theoretic motion planning is a popular choice for modeling multi-agent non-cooperative interactions, such as autonomous driving [4, 5, 7]–[9], crowd navigation [10], distributed systems [11, 12], drone racing [13], and shared control [14]. A central focus of game-theoretic planning is local Nash equilibria (LNE) [1], in which no agent is unilaterally incentivized to deviate from their strategy. LNEs have been studied extensively in autonomous driving as a model for interaction in merging [15]–[17], highway overtaking [18]–[20], and racing [8, 21]. However, in many real-world settings, agents are not perfectly rational and may deviate significantly from actions they believe are optimal due to distractions [22], imperfect information [23, 24], or poor human-machine interfacing [25]. In this work, we adopt the principle of *bounded rationality* [26]–[29], in which agents act rationally with likelihood proportional to the distribution of utilities over actions.

While LNEs provide a mechanism for predicting and analyzing multi-agent interactions, multiple LNEs may arise depending on the initial condition of the joint (multi-agent) system [1]. Scenarios where multiple equilibria exist are called *multi-modal* due to the existence of distinct (and possibly stochastic) outcomes, the selection of which might depend on individual preferences. Multi-modality in planning has recently become a popular object of study for modeling high-level decision-making in games. Peters et al. [30] introduce an inference framework for maximum a posteriori equilibria aligned control by randomly selecting seed strategies and solving multiple seeds in parallel to an equilibrium, to then normalize their weights in a manner similar to a particle filter. So et al. [31, 32] show that equilibria can be efficiently clustered and inferred using a Bayesian update, which can then be used to hedge different strategies using a QMDP-style [33] cost estimate. Peters et al. [34] show that multiple game-theoretic contingencies can be solved, with an arbitrary branching time, using mixed-complementarity programming. Recent work [9] uses implicit dual control to tractably compute an active information gathering policy for a class of partially-observable stochastic games, while preserving the multi-modal information encoded in the robot's belief states with scenario optimization. Follow-up work [24] extends this idea to a high-dimensional adversarial setting using deep reinforcement learning to synthesize the robot's policy at scale.

Multi-modality in planning can arise from uncertain or unknown objectives. For example, in a merging setting, which player prefers to go first may directly alter the outcome of the interaction, and this *objective uncertainty* may lead to indecision. Cleac'h et al. [18] introduce a motion planning method for uncertain linear cost models by estimating cost parameters using an unscented Kalman filter. Chen et al. [35] introduce an MPC-based framework for interactive trajectory optimization using deep-learning predictions as input. Hu et al.

[36] study emergent coordination in multi-agent interactions when cost parameters follow an opinion-dynamic interaction over a graph. Recent work [37] uses differentiable optimization to learn open-loop mixed strategies for non-cooperative games based on trajectory data. Liu et al. [17] extend this idea to compute generalized equilibria, a setting where hard constraints are present, and use deep learning to bolster the computation performance of the game solver. Another follow-up work [38] further studies the setting of learning a feedback Nash equilibrium strategy. In the potential game setting, Diehl et al. [39] show that multiple open-loop Nash equilibria can be computed in parallel using deep-learned game parameters.

To the best of our knowledge, our approach is the first closed-loop, model-based game-theoretic planner that can incorporate multi-modal priors from data-driven trajectory prediction models for guided exploration and producing scene-consistent interactions. Our method makes only mild assumptions on the environment dynamics and planning costs (e.g. being differentiable). Moreover, our provision of a scalar regularization parameter permits *tunable* adherence to data-driven behavior at inference time, in contrast to traditional game-theoretic planning and end-to-end equilibrium learning, which are rigid in their respective assumptions about planning objectives (i.e. task-driven or data-driven).

B. Stochastic Optimal Control and Dynamic Games

In contrast to greedy-optimal, deterministic behavior predicted by some classical control-theoretic planners, stochastic behavior has been exhibited in many real-world settings, such as biological muscle control [40], animal foraging [41], and urban driving [42]. Stochastic optimal control (SOC) provides a mechanism for solving and explaining *stochastic* optimal policies through accommodating uncertainty in decision-making. Energy-based methods [43]–[46] are a well-studied technique for solving optimal policies by analyzing utility functions similar to the total energy in statistical physics. KL control [47]–[50] is a branch of energy-based methods that studies solutions for SOC problems where the control cost is augmented with a Kullback-Leibler divergence regularizer term that tries to keep the resultant policy close to some *reference policy*, which may have different uses depending on the task. For example, the reference policy could minimize control effort [47], guide exploration [51], or incorporate human feedback [52].

Due to the intractability of the expectation term for general stochastic optimal control and dynamic games, scenario optimization [53, 54] is often used as an approximate solution method. Schildbach et al. [55] uses scenario optimization for autonomous lane change assistance when the ego vehicle is interacting with other human-controlled vehicles, whose future motions are predicted with a pre-specified scenario generation model and incorporated in a model predictive control (MPC) problem. Chen et al. [35] develops a scenario-based MPC algorithm to handle multi-modal reactive behaviors of uncontrolled human agents. In [9, 56], a provably safe scenario-based MPC planner is proposed for interactive motion planning in uncertain, safety-critical environments, which improves task performance by preempting emergency safety maneuvers triggered by unlikely agent behaviors. Recent work [57] uses alternating direction method of multipliers (ADMM) based scenario optimization to tractably compute Nash equilibria for a class of constrained stochastic games subject to parametric uncertainty.

Maximum-entropy optimal control constitutes another branch of energy-based methods (essentially equivalent KL control with a uniform reference policy) that incentives entropy of the optimal policy, which can be used to incentivize policy robustness [58], learn from human data [59], and escape local minima in long-horizon tasks [60]. Mehr et al. [4] take a step forward into the multi-agent land and study a maximum-entropy dynamic game.

In this work, we provide a mechanism for incorporating policies guided by data-driven models from large aggregated datasets in model-based game-theoretic planning. We show that our formulation is a generalization of the maximum-entropy dynamic game when the reference policy becomes arbitrarily uninformative. We also leverage scenario optimization to tractably handle multi-modal reference policy at scale.

C. Multi-Modal Multi-Agent Motion Prediction

Multi-agent trajectory prediction seeks to model agent-to-agent interactions that greatly affect joint outcomes versus predicting independently for each agent [61]. Recently, multi-agent motion prediction [6, 62, 63] has enjoyed great success in predicting multi-modal human behavior in part due to advances in transformer-based architectures [64, 65], natural language processing [66], and diffusion models [67]. The output of a motion predictor is a weighted set of trajectory samples [68], typically in the form of a mixture model over discrete modes [69]. Finally, some approaches explicitly model discrete agent interactions in order to better account for them and improve accuracy [70]–[72].

In sample-based prediction, maintaining the diversity of samples to represent distinct outcomes is an active area of research. Recent works [73, 74] emphasize *multi-modality* by incorporating rollouts and fictitious play to increase diversity in the prediction framework. Farthest Point Sampling (FPS) [75, 76], Non-Maximum Suppression (NMS) [77], and/or neural adaptive sampling [78] provide methods for encouraging outcome diversity using pairwise distances between trajectories as a metric. Sample diversity makes prediction methods a popular data-driven choice for human behavior.

Prediction models have been incorporated in planning through an integrated prediction and planning approach, typically a model-predictive controller that seeks to avoid other agents [3, 79]. In [2], joint motion forecasts are incorporated as a constraint in the motion planning framework, allowing the planning agent to better understand high-dimensional scene data (such as the map) through low-dimensional action selection. In [80], a parameter-shared cost model is learned from data for all agents in the scene, allowing for scene-dependent cost formulation.

Generative modeling of traffic scenarios is also capable of compositing learned behavior behaviors with model-based

objectives. Recent works [81, 82] utilize signal temporal logic to guide the diffusion model generating desired behaviors. VBD [83] applies a game-theoretic gradient descent-ascent guidance strategy to generate interactive safety-critical scenarios with a diffusion model.

In this work, we provide a principled approach to integrate the *multi-modality* of motion predictors with closed-loop game-theoretic motion planning. Our framework complements end-to-end learned methods by allowing an agent to consider multiple motion modes and compute an optimal policy for each one, rather than needing to select a mode *a-priori*. Furthermore, by specifying additional incentives such as safety, KLGame allows a robot to reason *strategically* over time in a *closed-loop* manner, making it more robust to unmodelled scenarios through feedback when compared to open-loop end-to-end predictions.

III. Problem Formulation

Notation. We will take T as the planning horizon and define $[T]:=\{1,...,T\}$. Similarly, if there are N players we will use $[N]:=\{1,...,N\}$ to refer to the set containing all the players' indices. Throughout the paper we will use subscripts to refer to time and superscripts for player. For example, u_t^i will stand for the action of player i at time t. For a matrix M we use $M\succ 0$ (or $M\succeq 0$) to indicate M is positive (or semi-positive) definite. We use $\sigma(M)$ to indicate the set of singular values of M.

We consider a discrete-time dynamic game with N robots over a finite horizon T governed by the nonlinear dynamics

$$x_{t+1}=f_t(x_t,u_t), \tag{1}$$

where $x_t\in\mathcal{X}$ is the joint state of the dynamical system and $u_t:=(u_t^1,...,u_t^N)$ is the joint control input. Each player $i\in[N]$ applies an action according to a stochastic policy $\pi_t^i(u_t^i|x_t)$ at each time step t.

Each player i minimizes a cost J^i that quantifies the loss associated with the trajectory and control effort exerted by the player as well as the deviation with respect to a reference policy $\tilde{\pi}_t^i(u_t^i|x_t)$ over a time horizon T:

$$J^i(\pi^i)=\mathbb{E}^{\pi}\left[\sum_{t=0}^{T} l^i(x_t,u_t)+\lambda^i D_{KL}(\pi_t^i(\cdot|x_t)||\tilde{\pi}_t^i(\cdot|x_t))\right], \tag{2}$$

where $\pi:=(\pi^1,...,\pi^N)$ is the joint policy of all robots, λ is an N-vector of nonnegative scalars, and D_{KL} is the Kullback-Leibler (KL) divergence from the reference policy,

$$D_{KL}(\pi_t^i||\tilde{\pi}_t^i)=\int_{\mathcal{U}}\pi_t^i(u_t^i)\log\left(\frac{\pi_t^i(u_t^i)}{\tilde{\pi}_t^i(u_t^i)}\right)du_t^i. \tag{3}$$

For the KL divergence to be well-defined, we require that $\pi_t^i(\cdot|x_t)$ is absolutely continuous with respect to $\tilde{\pi}_t^i(\cdot|x_t)$, i.e., $\tilde{\pi}_t^i(\cdot|x_t)=0 \implies \pi_t^i(\cdot|x_t)=0$ [84]. We denote absolute continuity as $\pi_t^i << \tilde{\pi}_t^i$.

The scalar $\lambda^i\geq 0$ provides a natural "tuning knob" for the stochastic dynamic game, balancing the preference between optimizing the player's nominal performance objectives, encoded in l^i, and aligning its policy with the reference policy, $\tilde{\pi}^i$. We refer to this new class of dynamic game as *KLGame*, which can be compactly formulated as:

$$\begin{aligned}\min_{\pi^i}\quad & J^i(\pi^i)\\ \text{s.t.}\quad & x_{t+1}=f_t(x_t,u_t),\quad \forall t,\end{aligned} \tag{4}$$

for all $i\in[N]$. Herein, we use $\neg i$ to denote all agents except for i, and the superscript $(\cdot)^*$, as an indicator of optimality.

Recent work [38] revealed that feedback Nash equilibria (FBNE) strategies can generate more expressive and realistic interactions than their open-loop counterpart, which is an important property for interaction-rich and safety-critical motion planning tasks such as autonomous driving. For the same reason, we focus on seeking FBNE of game (4) in this paper.

Definition 1 (Feedback Nash Equilibrium [1])**.** *A policy π^{i*} defines an FBNE if no player has an incentive to unilaterally alter the strategy, i.e.,*

$$J(\pi^{i*},\pi^{\neg i*})\leq J(\pi^i,\pi^{\neg i*}),\quad \forall\pi_t^i(\cdot)\in\Pi^i,\ \forall t\in[T],\ \forall i\in[N].$$

We now introduce the main technical approach for seeking FBNE of KLGame.

IV. Kullback-Leibler Regularized Games

In this section, we propose an efficient algorithm for seeking FBNE of KLGame (4). Our roadmap towards such a game solver is as follows. First, we show that a KL-regularized Bellman equation adopts a closed-form solution. Then, we derive the analytical global Nash equilibrium strategy in the LQ setting with Gaussian reference policies. Finally, we propose an iterative linear-quadratic (ILQ) algorithm for efficiently approximating the game solution with general nonlinear dynamics, costs, and multi-modal reference policies based on the ILQ approximation principle. Fig. 2 presents the high-level architecture of our system.

A. Dynamic Programming Formulation

We propose to solve (4) with the dynamic programming principle, which gives the Bellman equation:

$$V^i(x)=\inf_{\pi^i}\left\{\mathbb{E}^{\pi}[\mathcal{Q}^i(x,u)]+\lambda^i D_{KL}[\pi^i(\cdot|x)||\tilde{\pi}^i(\cdot|x)]\right\}, \tag{5}$$

where $V^i(x)=\inf_{\pi^i}J^i(\pi^i)$ denotes the optimal value function and $\mathcal{Q}^i(x,u):=l^i(x,u)+V^{i\prime}(f(x,u))$ is the Q-value function. In the remainder of this section, we show that (5) admits an optimal solution of a particular structure.

We first state a helpful Lemma, which is a slight modification to the dual of the Donsker-Varadhan variational formula [85].

Lemma 1. *(Regularized variational formula [86, 87].) Let $\mathcal{Q}:\mathcal{X}\times\mathcal{U}\to\mathbb{R}$ be any bounded measurable function. Then, for any nonnegative real number λ, probability measures π, $\tilde{\pi}$ and states $x\in\mathcal{X}$:*

$$\begin{aligned}&-\lambda\log\int_{\mathcal{U}}e^{-\mathcal{Q}(x,u)}\tilde{\pi}(u|x)du\\ &\quad=\inf_{\pi}\left\{\mathbb{E}^{\pi}[\mathcal{Q}(x,u)]+\lambda D_{KL}(\pi(\cdot|x)||\tilde{\pi}(\cdot|x))\right\}.\end{aligned} \tag{6}$$

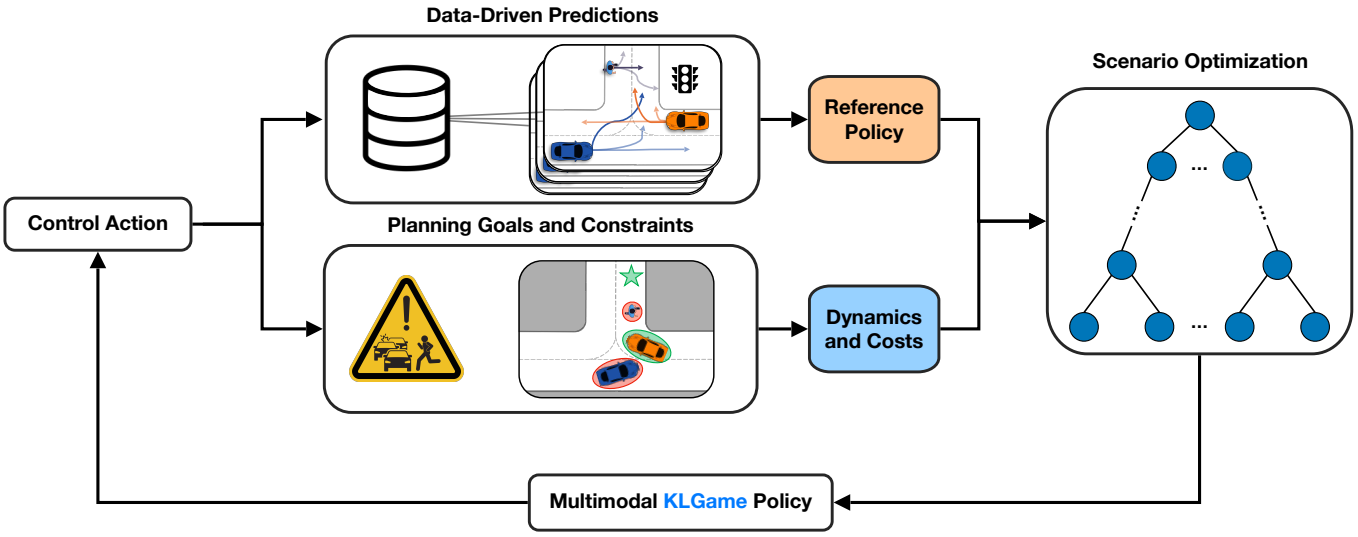

Fig. 2: Overview of the proposed algorithmic framework for computing an approximate feedback Nash equilibrium for KLGame. In essence, we use scenario optimization to compute a multi-modal mixed strategy via blending a multi-modal reference policy learned from human interaction data with optimization-based game-theoretic policy.

Moreover, the infimum in (6) is attained with minimizer π^* if π^* satisfies the expression:

$$\pi^*(u|x) = \frac{e^{-\mathcal{Q}(x,u)/\lambda}\tilde{\pi}(u|x)}{\int_{\mathcal{U}} e^{-\mathcal{Q}(x,u)/\lambda}\tilde{\pi}(u|x)du}. \tag{7}$$

Proof: See Appendix A. ■

Lemma 1 presents a convenient analytical form of the solution to Bellman equation (5): the optimal regularized policy is a (normalized) product of two density functions, one of which being precisely the reference policy. This result, as we will see momentarily, serves as a key step towards deriving the KLGame equilibrium strategy.

B. Global FBNE: The Linear-Quadratic Gaussian Case

Even without the KL regularizer in (2), seeking a local feedback Nash equilibrium of a non-LQ trajectory game is generally intractable [88]. However, if the system dynamics are linear, the cost functions are convex quadratic functions of the state (LQ), and the reference policy is Gaussian, the *global* Nash equilibrium can be computed efficiently. Specifically, the system dynamics are linear when

$$x_{t+1} = A_t x_t + \sum_{i\in[N]} B_t^i u_t^i + d_t, \tag{8}$$

where A_t, B_t^i are the state-transition matrices and $d_t \sim \mathcal{N}(0,\Sigma_d)$ Gaussian noise with zero mean and covariance Σ_d. The stage cost of each player is a quadratic function of state and control, i.e., $l^i(x_t,u_t) = x_t^\top Q_t^i x_t + \sum_{j\in[N]} {u_t^j}^\top R_t^{ij} u_t^j$, where Q^i and R^{ij} are positive semidefinite matrices of appropriate dimension penalizing state and control authority. The reference policies $\tilde{\pi}^i$ are Gaussian if

$$\tilde{\pi}_t^i = \mathcal{N}(\tilde{\mu}_t^i, \tilde{\Sigma}_t^i) \tag{9}$$

for all $t\in[T]$ and $i\in[N]$.

We call KLGame, in this special case, the KL-regularized Linear-Quadratic Gaussian (KL-LQG) game. Our goal is to find, for all agents i, the global FBNE strategy π^{i*} of a KL-LQG game. We show in the following theorem that π^{i*} is Gaussian with a *state-dependent* mean $\mu^{i*} = \mu^{i*}(x)$, which depends linearly on x, and a *state-independent* covariance Σ^{i*}. Furthermore, we show that μ^{i*} and Σ^{i*} are parameterized by the dynamics (A and B), task costs (Q and R), reference policy ($\tilde{\mu}$ and $\tilde{\Sigma}$), and value function parameters.

Theorem 1 (Global FBNE of KL-LQG Game)**.** *The N-player nonzero-sum KL-LQG dynamic game (4) admits a unique global feedback Nash equilibrium solution if,*

1) *The dynamics follow the functional form of (8), where $x_0 \sim \mathcal{N}(\mu_{x_0}, \Sigma_{x_0})$, $d_t \sim \mathcal{N}(0,\Sigma_d)$,*
2) *The costs, introduced in (2), have the functional form*

$$J^i(\pi^i) = \mathbb{E}^\pi \left[\sum_{t=0}^{T} \frac{1}{2} \left(x_t^\top Q_t^i x_t + \sum_{j\in[N]} {u_t^j}^\top R_t^{ij} u_t^j \right) \right] + \sum_{t=0}^{T} \lambda^i D_{KL}(\pi_t^i || \tilde{\pi}_t^i),$$

   where, for all $t\in[T]$, $Q_t^i \succeq 0$, $R_t^{ij} \succeq 0$, $\forall i,j\in[N], j\neq i$, $R_t^{ii} \succ 0$, $\forall i\in[N]$,
3) *The reference policies are Gaussian, as in (9).*

Moreover, the global (mixed-strategy) Nash equilibrium is a set of time-varying policies $\pi_t^{i*} = \mathcal{N}(\mu_t^{i*}, \Sigma_t^{i*})$, $\forall i\in[N]$ with mean and covariance given by

$$\begin{aligned} \mu_t^{i*} &= -K_t^i x_t - \kappa_t^i, \\ \Sigma_t^{i*} &= \left[\frac{1}{\lambda^i} \left(R_t^{ii} + {B_t^i}^T Z_{t+1}^i B_t^i \right) + \left(\tilde{\Sigma}_t^i \right)^{-1} \right]^{-1}, \end{aligned} \tag{10}$$

where (K_t^i, κ_t^i)*, the state-feedback matrix and constant, are given by solving the coupled KL-regularized Riccati equation:*

$$\begin{aligned}
&\left[\lambda^i(\tilde{\Sigma}_t^i)^{-1} + R_t^{ii} + B_t^{i\top} Z_{t+1}^i B_t^{i\top}\right] K_t^i + B_t^{i\top} Z_{t+1}^i \sum_{j\neq i} B_t^j K_t^j \\
&\qquad = B_t^{i\top} Z_{t+1}^i A_t, \\
&\left[\lambda^i(\tilde{\Sigma}_t^i)^{-1} + R_t^{ii} + B_t^{i\top} Z_{t+1}^i B_t^{i\top}\right] \kappa_t^i + B_t^{i\top} Z_{t+1}^i \sum_{j\neq i} B_t^j \kappa_t^j \\
&\qquad = B_t^{i\top} z_{t+1}^i - \lambda^i(\tilde{\Sigma}_t^i)^{-1} \tilde{\mu}_t^i.
\end{aligned} \tag{11}$$

Value function parameters (Z_t^i, z_t^i) *are computed recursively backward in time as*

$$\begin{aligned}
Z_t^i &= Q_t^i + \sum_{j\in[N]} (K_t^j)^T R_t^{ij} K_t^j + F_t^T Z_{t+1}^i F_t + \lambda^i K_t^{i\top} (\tilde{\Sigma}_t^i)^{-1} K_t^i, \\
z_t^i &= \sum_{j\in[N]} (K_t^j)^T R_t^{ij} \kappa_t^j + F_t^T \left(z_{t+1}^i + Z_{t+1}^i \beta_t\right) \\
&\quad + \lambda^i K_t^{i\top} (\tilde{\Sigma}_t^i)^{-1} (\kappa_t^i - \tilde{\mu}_t^i),
\end{aligned} \tag{12}$$

with terminal conditions $Z_{T+1}^i = 0$ *and* $z_{T+1}^i = 0$*. Here,* $F_t = A_t - \sum_{j\in[N]} B_t^j K_t^j$ *and* $\beta_t = -\sum_{j\in[N]} B_t^j \kappa_t^j$ *for all* $t \in [T]$*.*

Proof: See Appendix B. ∎

Theorem 1 reveals several important insights about KL-LQG games at equilibrium:

- The global FBNE must be a mixed strategy, taking the form of an unimodal Gaussian.
- The policy mean is a linear state-feedback control law, coinciding with the function form of FBNE of a deterministic LQ game [1]. Moreover, computing the global FBNE strategy enjoys the same order of complexity as the deterministic LQ game, since the coupled Riccati equation has the same dimension with additional red terms contributed by the reference policy,
- As the reference policy becomes arbitrarily uninformative, the FBNE of KL-LQG reduces to that of a maximum-entropy game [4],
- As $\lambda^i \to \infty$, the FBNE becomes the reference policy,
- As $\lambda^i \to 0$, the FBNE coincides with that of a deterministic LQ game.

Corollary 1 (Reduction to the MaxEnt Game)**.** *As the reference policies become arbitrarily uninformative, i.e.,* $\min \sigma(\tilde{\Sigma}_t^i) \to \infty, \forall i \in [N], \forall t \in [T]$*, a KL-LQG game reduces to the maximum-entropy dynamic game defined in [4] and its FBNE becomes an unguided, exploratory entropic cost equilibrium (ECE) strategy.*

Proof: Recall from [4] that an ECE in the LQ case is also an unimodal Gaussian. When $\min \sigma(\tilde{\Sigma}_t^i) \to \infty$, the contribution from the reference policy (red terms) in the covariance Σ_t^{i*} and Riccati equation vanishes, and the resulting covariance and Riccati equation match that of the maximum-entropy dynamic game. ∎

Corollary 2 (Reduction to the Reference Policy)**.** *As* $\lambda^i \to \infty$*, the FBNE of a KL-LQG game becomes the reference policy, i.e.,* $\lim_{\lambda^i\to\infty} \mu_t^{i*} = \tilde{\mu}_t^i$ *and* $\lim_{\lambda^i\to\infty} \Sigma_t^{i*} = \tilde{\Sigma}_t^i$*.*

Corollary 3 (Reduction to the Deterministic LQ Game)**.** *As* $\lambda^i \to 0$*, the FBNE of a KL-LQG game becomes a deterministic policy, i.e.,* $\lim_{\lambda^i\to 0} \mu_t^{i*} = -K_t^i x_t - \kappa_t^i$*, where* (K_t^i, κ_t^i) *is given by the coupled Riccati equation of the deterministic LQ game.*

While Theorem 1 lays the theoretical foundation of KLGame, the assumption that the reference policy $\tilde{\pi}_t^i \sim \mathcal{N}(\tilde{\mu}_t^i, \tilde{\Sigma}_t^i)$ implies that the guidance is ultimately *open-loop*: the moments of $\tilde{\pi}_t^i$ are independent of the state and thus unaware of players' interaction throughout the planning horizon. In the following, we extend Theorem 1 to the *closed-loop* setting by accounting for a *state-feedback* reference policy.

Proposition 1 (Closed-loop Guidance)**.** *Let all assumptions in Theorem 1 hold but assume reference policies are Gaussian conditioned on the time and state, i.e.,* $\tilde{\pi}_t^i(\cdot|x_t) \sim \mathcal{N}(\tilde{\mu}_t^i(x_t), \tilde{\Sigma}_t^i)$*, where* $\tilde{\mu}_t^i(x) = -\tilde{K}_t^i x - \tilde{\kappa}_t^i$*, for all* $t \in [T]$ *and* $i \in [N]$*. Under these assumptions, the global (mixed-strategy) Nash equilibrium is a set of time-varying policies* $\pi_t^{i*} = \mathcal{N}(\mu_t^{i*}, \Sigma_t^{i*})$*,* $\forall i \in [N]$ *with mean and covariance given by*

$$\begin{aligned}
\mu_t^{i*} &= -K_t^i x_t - \kappa_t^i, \\
\Sigma_t^{i*} &= \left[\frac{1}{\lambda^i}\left(R_t^{ii} + B_t^{i^T} Z_{t+1}^i B_t^i\right) + \left(\tilde{\Sigma}_t^i\right)^{-1}\right]^{-1},
\end{aligned} \tag{13}$$

where (K_t^i, κ_t^i) *is given by solving the coupled KL-regularized Riccati equation:*

$$\begin{aligned}
&\left[\lambda^i(\tilde{\Sigma}^i)^{-1} + R_t^{ii} + B^{i\top} Z_{t+1}^i B^{i\top}\right] K_t^i + B^{i\top} Z_{t+1}^i \sum_{j\neq i} B^j K_t^j \\
&\qquad = B^{i\top} Z_{t+1}^i A + \lambda^i(\tilde{\Sigma}_t^i)^{-1} \tilde{K}_t^i, \\
&\left[\lambda^i(\tilde{\Sigma}^i)^{-1} + R_t^{ii} + B^{i\top} Z_{t+1}^i B^{i\top}\right] \kappa_t^i + B^{i\top} Z_{t+1}^i \sum_{j\neq i} B^j \kappa_t^j \\
&\qquad = B^{i\top} z_{t+1}^i + \lambda^i(\tilde{\Sigma}_t^i)^{-1} \tilde{\kappa}_t^i.
\end{aligned} \tag{14}$$

Value function parameters (Z_t^i, z_t^i) *are computed recursively backward in time as*

$$\begin{aligned}
Z_t^i &= Q_t^i + \sum_{j\in[N]} (K_t^j)^T R_t^{ij} K_t^j + F_t^T Z_{t+1}^i F_t \\
&\quad + \lambda^i (K_t^i + \tilde{K}_t^i)^\top (\tilde{\Sigma}_t^i)^{-1} (K_t^i + \tilde{K}_t^i), \\
z_t^i &= \sum_{j\in[N]} (K_t^j)^T R_t^{ij} \kappa_t^j + F_t^T \left(z_{t+1}^i + Z_{t+1}^i \beta_t\right) \\
&\quad + \lambda^i (K_t^i + \tilde{K}_t^i)^\top (\tilde{\Sigma}_t^i)^{-1} (\kappa_t^i + \tilde{\kappa}_t^i),
\end{aligned} \tag{15}$$

with terminal conditions $Z_{T+1}^i = 0$ *and* $z_{T+1}^i = 0$*. Here,* $F_t = A_t - \sum_{j\in[N]} B_t^j K_t^j$ *and* $\beta_t = -\sum_{j\in[N]} B_t^j \kappa_t^j$ *for all* $t \in [T]$*.*

Proof: See Appendix C. ∎

We highlight in (14) and (15) terms that are different from Theorem 1 in orange, which stem from the state-feedback reference policy.

C. Solving General KLGame Planning Problems

Having gained insights into the theoretical properties of KLGame in the LQ setting with Theorem 1 and Proposition 1, we now return to the full KLGame setting with nonlinear dynamics (1), non-convex costs (2), and arbitrary (possibly

state-dependent) reference policies $\tilde{\pi}^i(u|x)$. The intractability of finding an exact FBNE in deterministic games [88] does not ease in our setting. Therefore, inspired by [5], we instead look for an approximate FBNE with the linear-quadratic-Laplace (LQL) approximation method, an extension of the ILQGame method in [5] to the KLGame setting, which we introduce in detail below.

1) The LQL approximation: We seek to find an approximate FBNE by iteratively solving a sequence of KL-LQG games until converging to a stationary point. This process is initialized with a *nominal* trajectory $\eta := \{\bar{x}_{[0:T]}, \bar{u}_{[0:T]}\}$, and is divided into two phases: backward pass and forward pass.

Backward Pass. During the backward pass, we linearize dynamics (1) and quadraticize cost (2) around η to obtain a linear dynamical system and quadratic cost functions that fit into the assumption of KL-LQG game (c.f. Theorem 1). Next, we use the Laplace approximation [89, Chapter 4] to obtain a Gaussian distribution that captures the reference policy locally around η. Specifically, the conditional probability distribution of a reference policy $\tilde{\pi}_t^i(u_t^i|x_t)$ is approximated as:

$$\tilde{\pi}_t^i(u_t^i|x_t) \approx \mathcal{N}\Big(\tilde{\mu}_t^i(x_t;\bar{\eta}_t), \tilde{\Sigma}_t^i(\bar{\eta}_t)\Big), \tag{16}$$

where $\bar{\eta}_t := \{x_t, u_t\}$, the mean function is

$$\tilde{\mu}_t^i(x_t;\bar{\eta}_t) := \operatorname{argmax}_{\tilde{u}^i(x_t)} \tilde{\pi}_t^i(\tilde{u}^i|\bar{x}_t), \tag{17}$$

and the covariance matrix is

$$\tilde{\Sigma}_t^i(\bar{\eta}_t) := -\Big[\nabla^2_{u^i} \ln \tilde{\pi}_t^i(u^i|x_t)\big|_{u^i=\tilde{\mu}_t^i(x_t;\bar{\eta}_t)}\Big]^{-1}. \tag{18}$$

The Laplace-approximated distribution in (16) is a Gaussian whose mean is centered at a mode $\tilde{\mu}_t^i(\cdot)$ of the reference policy $\tilde{\pi}_t^i(\cdot)$. As discussed, it is desirable to use closed-loop guidance with a state-dependent reference policy (c.f. Proposition 1). We note that, in this case, *functional optimization* is required in order to find the state-dependent mean function $\tilde{\mu}_t^i(x_t;\bar{\eta}_t)$ in (17). While an exact optimization is likely intractable, we provide a practical implementation to find such a function. First, we sample for $t \in [T]$ the original reference policy $\tilde{\pi}_t^i(u_t^i|\bar{x}_t)$ conditioned on the current nominal trajectory η and compute the sample means (control actions). Then, we solve an ILQR problem, which tracks those controls, for a time-varying linear state-feedback control law. This control law may be used as the mean function in (17), and the covariance matrix in (18) can be computed by setting u^i to those sample means.

With the linearized dynamics, quadraticize costs, and Laplace-approximated reference policy obtained above, we can now formulate a KL-LQG game, whose global FBNE can be found by leveraging Theorem 1 (or Proposition 1).

Forward Pass. The forward pass aims to update the nominal trajectory η with the KL-LQG strategies. This can be done by forward-simulating the system with players' control deviation at each time set to the KL-LQG strategy mean $\mu_t^{i,*}(x_t;\bar{\eta}_t)$, i.e., $u_t^i = \bar{u}_t^i + \delta u_t^i$, $\delta u_t^i = \mu_t^{i,*}(x_t;\bar{\eta}_t)$ for all $i \in [N]$ and $t \in [T]$.

Line Search. A common issue in ILQ-based trajectory optimization is that directly applying the mean of the KL-LQG strategy in the forward pass may cause the system to deviate too much from the nominal trajectory η, resulting in divergence of the algorithm. A possible remedy is line search [90, Chapter 3], which controls the trajectory update by scaling down the KL-LQG strategy parameters. Specifically, for a line search parameter $\varepsilon \in (0,1]$ and KL-LQG strategy $\mathcal{N}(-K_t^i x_t - \kappa_t^i, \Sigma_t^i)$, the adjusted strategy is $\mathcal{N}(-\varepsilon K_t^i x_t - \varepsilon\kappa_t^i, \varepsilon\Sigma_t^i)$, which is then used in the forward pass to obtain a new nominal trajectory η. The procedure starts with $\varepsilon \leftarrow 1$ and repeats by decreasing ε by half until a convergence criterion is met. Two commonly used criteria are decreasing of the social cost (i.e., the sum of all players' costs) and that the updated nominal trajectory is sufficiently close to the previous one.

2) Accounting for multi-modality with scenario optimization: The LQL approximation may be extended to account for both *multi-modal* approximate FBNE KLGame policy *and* reference policy with M discrete modes:

$$\tilde{\pi}_t^i(u_t^i|x_t) = \sum_{m=1}^{M} w_t^{i,m} \tilde{\pi}_t^{i,m}(u_t^i|x_t), \tag{19}$$

where *mixture weights* $w_t^{i,m} \geq 0$, $\sum_{m=1}^M w_t^{i,m} = 1$, and *mixture components* $\tilde{\pi}_t^{i,m}(u_t^i|x_t)$ are arbitrary probability density functions. Due to the need for capturing multi-modality in the KLGame and reference policy, it no longer suffices to perform forward and backward passes along a single nominal trajectory η. In order to tractably compute a multi-modal approximate FBNE of KLGame in this setting, we propose to use scenario optimization (SO) [53, 54] as the underlying solution framework. SO performs trajectory optimization over a scenario tree, whose root node is the initial state x_0, which then descends into $\underline{M}$ nodes ($\underline{M} \leq M$), each corresponding to a different mode, either specified by the user or determined at random. Each branch of the tree is a *scenario* governed by a sequence of reference policy modes. While most LQL approximation concepts carry over, we discuss necessary modifications to the backward and forward pass procedures under the SO framework below.

Modified Backward Pass. First, we need to change the Laplace approximation to account for the multi-modality of the reference policy $\tilde{\pi}_t^i(u_t^i|x_t)$. We propose to approximate the reference policy as a *Gaussian mixture model (GMM)*, i.e., $\tilde{\pi}_t^i(u_t^i|x_t) \approx \sum_{m=1}^M w_t^{i,m} \mathcal{N}\Big(\tilde{\mu}_t^{i,m}(x_t;\bar{\eta}_t), \tilde{\Sigma}_t^{i,m}(\bar{\eta}_t)\Big)$. This is obtained by inheriting the mixture coefficients $w_t^{i,m}$ from the original reference policy (19) and applying Laplace approximation to each mixture component $\tilde{\pi}_t^{i,m}(\cdot)$.

Next, we modify the backward-time computation of the KL-LQG strategies so that it is performed over the scenario tree, thus preserving the time causality of the resulting strategies [91]. Specifically, given a scenario tree with node set $\mathcal{NS}$, for each node $n \in \mathcal{NS}$ with time step t, the value function parameters of the succeeding time step in Theorem 1 (or Prop. 1) are obtained as a weighted sum of children node solutions, i.e., $Z_{t+1}^{i,n} = \sum_{\tilde{n}\in\Phi(n)} w_{t+1}^{i,\tilde{n}} Z_{t+1}^{i,\tilde{n}}$ and $z_{t+1}^{i,n} = \sum_{\tilde{n}\in\Phi(n)} w_{t+1}^{i,\tilde{n}} z_{t+1}^{i,\tilde{n}}$, where $\Phi(n)$ is the set of children nodes of n.

Finally, we propose to obtain a *GMM* approximate FBNE policy of the KLGame for each non-leaf node n in the

scenario tree:

$$\pi_t^{i,n,*}(u_t^i|x_t) = \sum_{\tilde{n}\in\Phi(n)} w_t^{i,\tilde{n}} \mathcal{N}\Big(\tilde{\mu}_t^{i,\tilde{n}}(x_t;\bar{\eta}_t), \tilde{\Sigma}_t^{i,\tilde{n}}(\bar{\eta}_t)\Big).$$

In this GMM policy, the mixture coefficients $w_t^{i,\tilde{n}}$ are set to the values used for tree construction, and the mean $\tilde{\mu}_t^{i,\tilde{n}}(x_t;\bar{\eta}_t)$ and the covariance $\tilde{\Sigma}_t^{i,\tilde{n}}(\bar{\eta}_t)$ are computed with (10) (or (13)) by setting the succeeding value function parameters to that of the children node, i.e, $Z_{t+1}^i = Z_{t+1}^{i,\tilde{n},*}$ and $z_{t+1}^i = z_{t+1}^{i,\tilde{n},*}$.

Modified Forward Pass. The forward pass computation is now performed over the scenario tree: at a node n with nominal state $\bar{x}_t^n$, nominal control $\bar{u}_t^{i,n}$, and KLGame policy $\pi_t^{i,n,*}(u_t^i|x_t)$, for each mode m of $\pi_t^{i,n,*}(\cdot)$, we compute a control action $u_t^{i,m} = \bar{u}_t^{i,n} + \tilde{\mu}_t^{i,m}(x_t;\bar{\eta}_t)$. Those control actions bring the system to the $\underline{M}$ children nodes of n, and the procedure continues until the nominal state and control are updated for all nodes in the scenario tree.

The Algorithm. We summarize the procedure of computing a GMM approximate FBNE of multi-modal KLGame in Algorithm 1. We omit the algorithm for the basic KLGame introduced in the previous section since it is a special case of the multi-modal one. In practice, we apply the KLGame policy in receding horizon fashion: each player i executes action $u^i = \bar{u}^{i,r} + \delta u^{i,r}$, where $\bar{u}^{i,r}$ is the nominal control of the root node r and $\delta u^{i,r}$ is sampled from the GMM policy $\pi^{i,r,*}(u^i|x)$ of the root node, time horizon is shifted, and KLGame is solved repeatedly once a new state measurement is received.

Computation Complexity. The computation complexity of multi-modal KLGame is affected by the state dimension n, each player's control dimension m_i, the number of players N, the number of scenario tree branches H, and the number of LQL iterations (c.f. Section IV-C1). Assuming $n > m_i$ for each player i, in each iteration, along a single tree branch, the complexity of the LQ approximation and solving the coupled Riccati equation of the corresponding KL-LQG game (c.f. [1, Corollary 6.1]) are $\mathcal{O}(Nn^2)$ and $\mathcal{O}(N^3n^3)$, respectively, coinciding with that of ILQGame [5]. The Laplace approximation has complexity $\mathcal{O}(Nn^3)$. The computation bottleneck of Algorithm 1 is the matrix inversion operation when solving the coupled Riccati equation, which has complexity $\mathcal{O}(N^3n^3)$. Therefore, the overall per-iteration complexity of multi-modal KLGame is $\mathcal{O}(HN^3n^3)$ (for ILQGame, the complexity is $\mathcal{O}(N^3n^3)$). Our implementation leverages the automatic vectorization of JAX [92] to efficiently batch-compute gradients and Hessians across different players and time steps. Throughout our experiments, we observe that Algorithm 1, when warmstarted and solved in a receding horizon fashion, typically converges with a small number of LQL iterations. As shown in Fig. 12, our solver achieves a 10Hz planning rate for a reasonable number of agents (around four) in complex traffic interaction scenarios where agents are in close proximity.

V. Experimental Results

In this section, we study the role of the reference policy in helping KLGame find diverse game solutions more amenable to execution than existing methods. We compare our method against a mixture of deterministic, stochastic, and data-driven baselines on three simulated interaction scenarios with nonconvex costs and mode uncertainty. In all simulations, we use the 4D kinematic bicycle model [93] to describe the motion of the vehicles. We implement a receding horizon version of multi-modal KLGame using JAX [92], which runs at 10Hz on a desktop with an AMD Ryzen 9 7950X CPU.

Algorithm 1 Multi-modal KLGame

Input: Initial state x_0, M-mode reference policy $\tilde{\pi}_t^i(u_t^i|x_t)$, branching number $\underline{M} \le M$

Initialization: Construct a scenario tree with node set $\mathcal{NS}$ and nominal trajectories $\eta_{\mathcal{NS}} := \{x^n, u^{i,n}\}_{n\in\mathcal{NS}}$

while `Not converged` **do**

$\{\pi_t^{i,n}(\cdot)\}_{n\in\mathcal{NS}} \leftarrow$ BackwardPass(x_0, $\eta_{\mathcal{NS}}$, $\tilde{\pi}_t^i(\cdot)$)

$\{\pi_t^{i,n}(\cdot)\}_{n\in\mathcal{NS}}, \eta_{\mathcal{NS}} \leftarrow$ ForwardPass&LineSearch(x_0, $\eta_{\mathcal{NS}}$, $\{\pi_t^{i,n}(\cdot)\}_{n\in\mathcal{NS}}$)

end while

Output: GMM approximate FBNE strategies $\{\pi_t^{i,n,*}(\cdot)\}_{n\in\mathcal{NS}}$ and optimized nominal trajectories $\{x^{n,*}, u^{i,n,*}\}_{n\in\mathcal{NS}}$

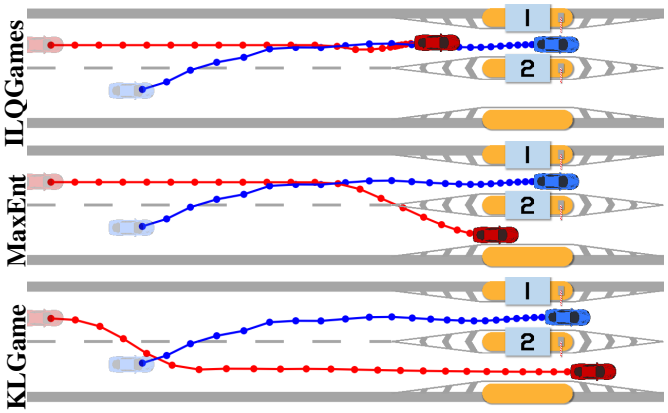

Fig. 3: Two-player tollbooth interaction. Player 1 (red) and Player 2 (blue) are approaching a tollbooth on a two-lane road (upper: 1, lower: 2). Both players receive a coordination bonus for taking opposite lanes (e.g. for reducing traffic), but Player 2 has a strong preference to merge into Lane 1. Since players incur a large penalty for deviating from the lane centerline, Player 1 may get caught in a local cost minimum (Lane 1) without external guidance to switch to Lane 2.

A. Two-Player Tollbooth Selection

In this experiment, we use the highway tollbooth coordination example from [36] to showcase the KLGame algorithm's ability to incorporate data-driven exploration in order to escape local minima despite nonconvex costs. We introduce two baseline methods.

ILQGames [5]. ILQGames is a popular algorithm for solving near-LNE policies via a set of coupled algebraic Riccati equations. While optimizing for socially optimal costs allows players to find mutually beneficial trajectories, nonconvexity in costs coupled with a lack of exploration in the state space may incentivize the joint system to remain stuck at a local minimum.

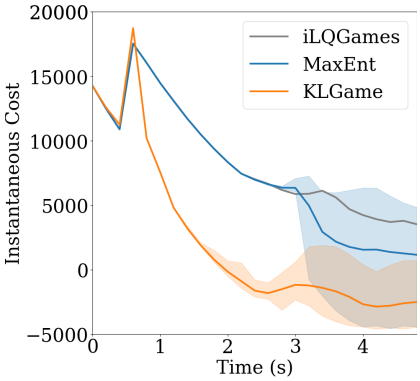

Fig. 4: Instantaneous task costs for ILQGames, MaxEnt, and KLGame methods over 100 trials. The initial spike in cost is due to Player 2 merging into Player 1's lane, eliminating the coordination bonus. MaxEnt incorporates stochastic exploration to find lower-cost trajectories, but incurs high variance. KLGame uses guidance from the reference policy to find a coordinating trajectory at an earlier iteration than MaxEnt, recovering the coordination bonus sooner and giving a lower overall cost.

TABLE I: Effect of Regularization in Tollbooth Interaction

Method	CR ↑	SR↑	Prog. (m) ↑	Cost (10^3)↓
ILQGames	0.00±0.00	**1.00**±0.00	45.0±0.00	8.45±0.00
MaxEnt	0.28±0.43	0.68±0.48	46.8±3.03	6.91±1.64
KLGame (Ours)	**1.00**±0.00	**1.00**±0.00	**51.8**±5.83	**4.22**±1.12

Maximum Entropy Game [31, 32]. Inspired by the success of the maximum-entropy framework in games, we explore how random exploration can be used to escape local minima when costs are nonconvex. In maximum-entropy games, the optimal exploration rate is proportional to the Hessian of the control cost [31]. In contrast, the optimal KLGame policy incorporates the mean and covariance of a reference policy, allowing better exploration of focal regions of the state space, such as right or left turns.

Fig. 3 compares qualitative behavior from KLGame against two methods that incorporate different regularization (no regularization for ILQGames and entropy regularization for MaxEnt). Two cars are rapidly approaching two toll stations and wish to get through the stations as quickly as possible and without slowing down. To capture the representative for each method, we show the *mean clustered equilibrium* for the three different methods. The ILQGames Player 1 is fully deterministic, and Player 1 must break and wait for Player 2 instead of changing lanes. The MaxEnt game Player 1 is able to explore and find the coordinating plan, but such exploration results in a delayed switch, and moreover, a penalty for crossing the lane boundary. The KLGame Player 1 incorporates a reference policy to find the coordinating plan. Using only a constant turn rate reference as a proxy for interactive toll-booth data, the KLGame Player 1 merges early and incurs the largest coordination bonus. Fig. 4 depicts instantaneous costs for the three methods, with KLGame incurring the lowest instantaneous and cumulative cost.

Table I presents performance comparisons for the three

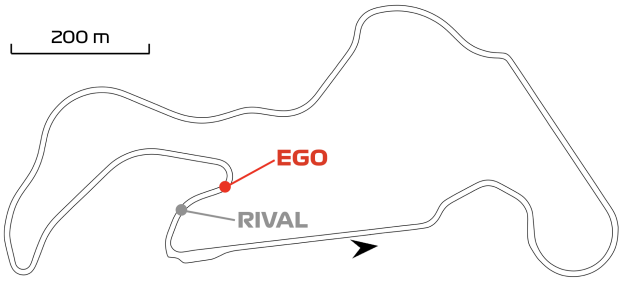

Fig. 5: Initial positions of the ego and rival vehicles on the Thunderhill Raceway. The track direction is indicated by the black arrow.

Fig. 6: Driving simulator used for collecting expert racing data.

methods over 100 trials. We report the following methods: (i) coordination rate (CR): the proportion of trials that end with the vehicles in opposite lanes, (ii) safety rate (SR): the proportion of trials where the vehicles remain in-bounds and do not collide, (iii) road progress, the distance that Player 1 travels from their initial position (as a measure of efficiency), (iv) the minimum distance to Player 2, and (v) the time-averaged cost incurred over the full rollout, averaged over all of the trials. Using a simple reference policy, KLGame is able to coordinate at a much higher rate than other methods while maintaining safety. **Key Takeaway.** Mixing the game's payoff with the expert's reference policy allows the game solver to *break out of a suboptimal equilibrium.*

B. Autonomous Car Racing

In this second experiment, we leverage KLGame within the more challenging context of autonomous car racing, delving into the advantages offered by the multi-modal KLGame policy framework.

Setup. We construct a simulated racing environment mirroring the Thunderhill Raceway in Willows, CA, USA, scaled accurately to 1:1 (see Fig. 5). In the race, the ego vehicle (red) is approaching a leading, rival vehicle (grey), and both vehicles are constrained to remain within the track boundaries. To make the scenario more challenging, we assume that only the ego vehicle has the responsibility to avoid a collision.

Reference policy and rival behaviors. To obtain a reference policy, we use a driving simulator (Fig. 6) to collect driving data (states, actions, track information, etc.) from races performed by two expert human racers in Carla [94]. With this

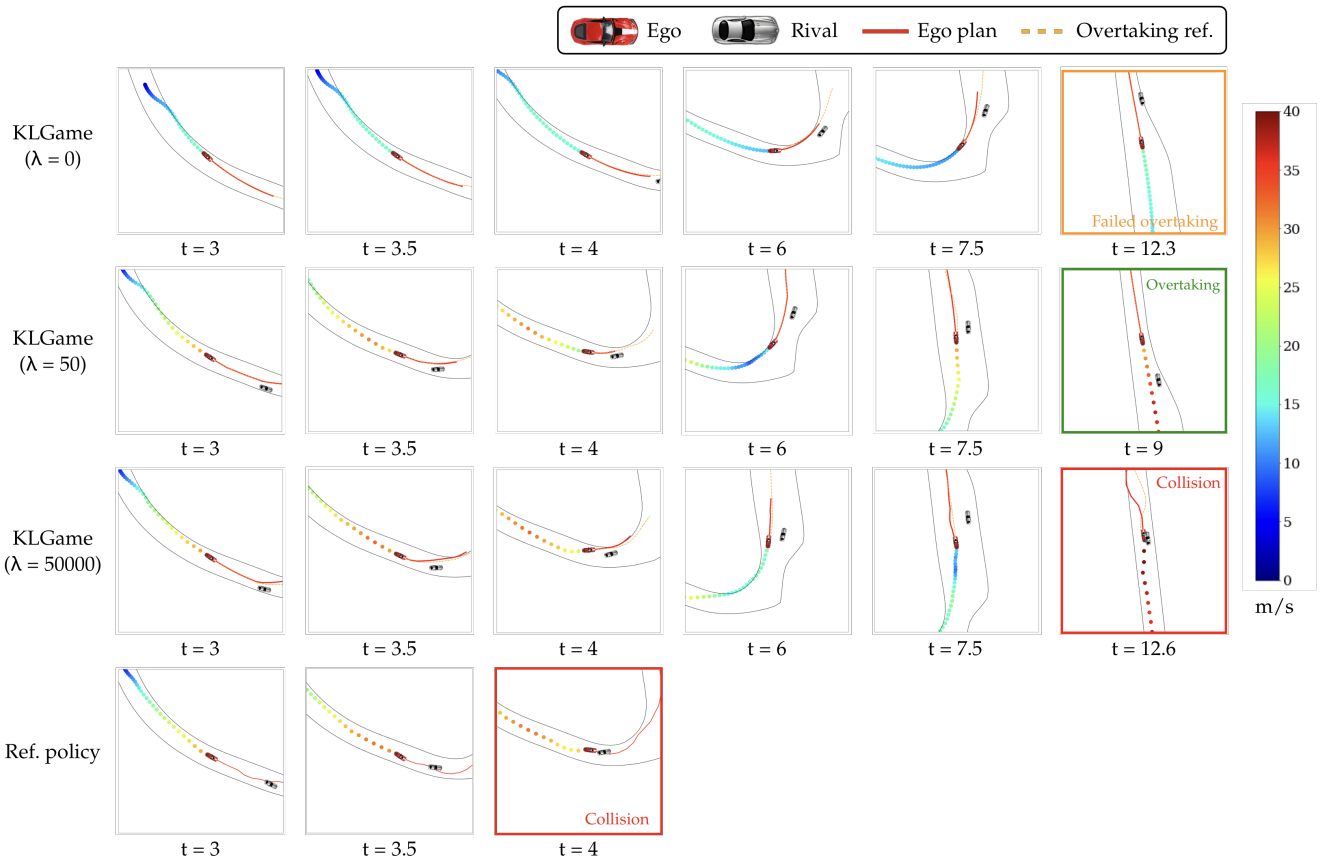

Fig. 7: Autonomous racing with the ego car using the basic (unimodal) KLGame guided by a single overtaking reference policy learned from human demonstration. *First row:* With un-regularized KL-Game ($\lambda = 0$), the policy has no incentive to overtake and as a result, the ego cannot pass the rival. *Second row:* For $\lambda = 50$, a successful overtaking is observed. *Third row:* For $\lambda = 5 \times 10^4$, the ego vehicle's action, heavily influenced by the reference policy, leads to a collision after a failed overtake attempt. *Fourth row:* The reference policy leads to a collision when the ego attempts to overtake the rival aggressively.

data, we employ an inverse dynamic game training procedure inspired by works such as [17, 38]. This approach involves defining a set of basis functions for the game's cost—each representing a distinct racing specification like optimizing lap times, collision avoidance, and track boundary adherence—and learning a context-dependent cost weight model. The model, which takes as inputs driving data and outputs the cost weights, is learned as a deep neural network leveraging automatic differentiation in JAX [92]. Given the limited volume of human-generated data, we use a dataset aggregation (DAgger) strategy [95] to enrich the training dataset for the reference policy. To obtain a stochastic reference policy, we use a game-theoretic variant of the Boltzmann distribution [9, Section 3.2], in which the Q-value function is obtained by solving the game with the inverse-learned cost weight model, which captures the noisily-rational aspect of human decision making. Finally, we produce the rival's behavior with a defensive ILQGame policy [5], whose parameters (cost terms and weights) are not accessible to either the reference or KLGame policies.

Basic KLGame. We first evaluate a basic unimodal KLGame policy, wherein the reference policy is simply the inverse-learned game policy that aggressively attempts to overtake the rival. Snapshots from the simulation are plotted in Fig. 7. For $\lambda = 0$, the ego simply follows the racing line (the time-optimal path of a race track) and fails to overtake the rival due to a lack of such incentive. As λ is increased to 50, the ego car overtakes the rival safely. When $\lambda = 5 \times 10^4$, the KLGame policy behaves closely to the aggressive reference policy, which leads to a collision. Even though the ego's nominal game cost contains a collision avoidance term, it is ultimately overpowered by the regularization term weighted by the large value of λ and is no longer capable of producing safe behaviors. Finally, as a sanity check, we directly apply the reference policy to the ego vehicle and witness a collision, as expected.

Multi-modal KLGame. Motivated by the result shown above for *Basic KLGame*, where blending only an aggressive reference policy can lead to unsafe outcomes, this experiment demonstrates how a *multi-modal KLGame (MM-KLGame)* (c.f. Sec. IV-C2) with an additional safety-enhancing mode performs in the racing scenario. The second mode we use is a *following* policy that controls the ego car to (i) follow the rival while keeping a constant distance when the rival is leading or (ii) defend against the rival after overtaking it. During each planning cycle, instead of picking the mode via sampling, we use rollout-based safety filters [96, Section 3.3] and perform a collision check for the KLGame-optimized trajectories asso-

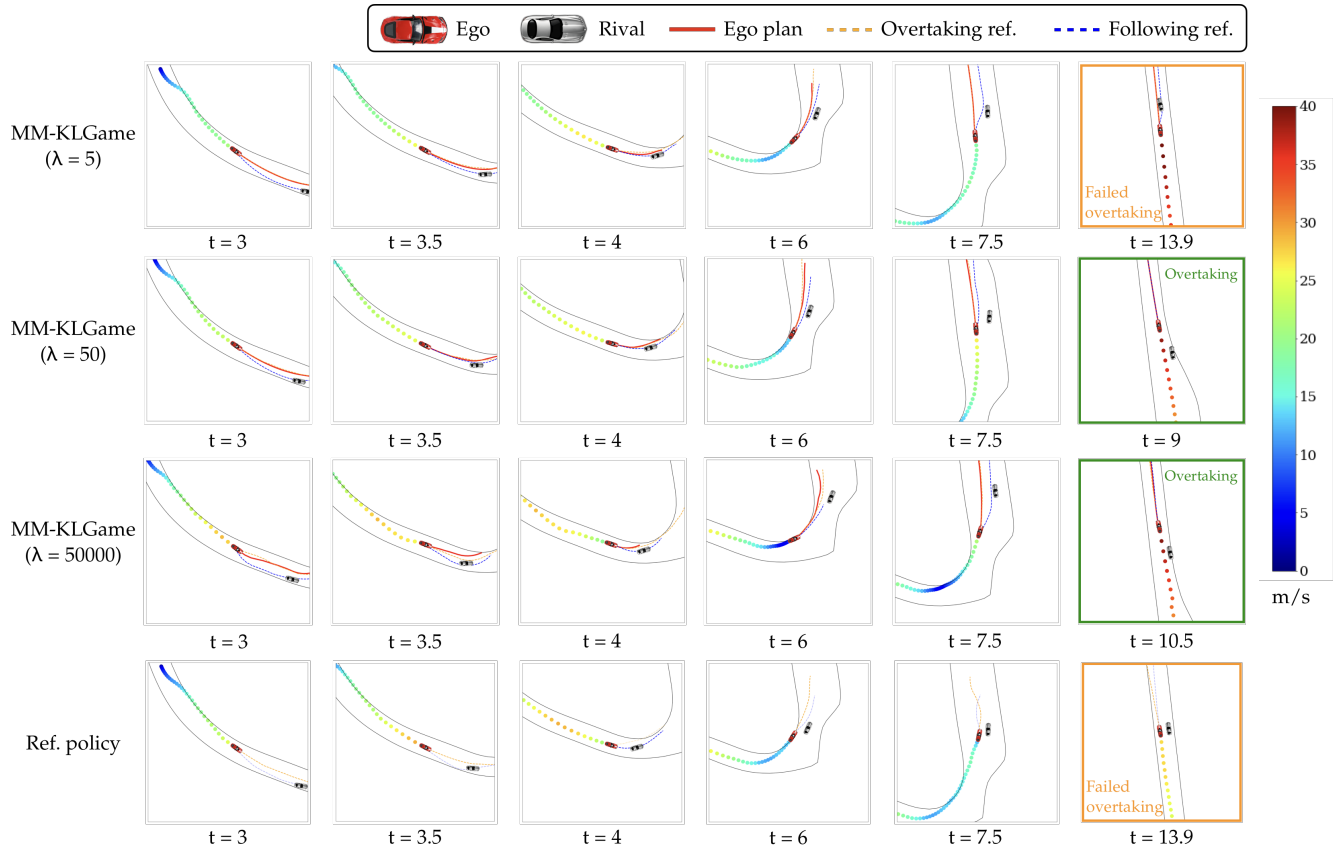

Fig. 8: Autonomous racing with the ego car using a multi-modal KLGame policy mixing an overtaking and a following mode. *First row:* With little regularization ($\lambda = 5$), the ego is too conservative to overtake the rival. *Second row:* For $\lambda = 50$, the ego successfully overtakes the rival without incurring a collision. *Third row:* For $\lambda = 5 \times 10^4$, the MM-KLGame policy results in a safe overtake. *Fourth row:* The multi-modal reference policy fails to overtake but averts collision.

ciated with the overtaking mode: if they are collision-free, we choose the overtaking mode; otherwise, we pick the following mode. Again, we vary the value of λ and inspect the qualitative difference in the resulting ego behaviors. The simulation snapshots are shown in Fig. 8. When λ takes a relatively small value, $\lambda = 5$, the MM-KLGame is not sufficiently guided by the overtaking policy and fails to overtake the rival. When $\lambda = 50$, the ego using MM-KLGame behaves similarly to the unimodal case and safely overtakes the rival. However, in the case when $\lambda = 5 \times 10^4$, instead of incurring a collision as in the unimodal KLGame case, the MM-KLGame leads to a safe overtake thanks to the additional following mode in the multi-modal reference policy. We also apply the reference policy to the ego using the same mode selection rule, resulting in a *shielded* inverse dynamic game policy. Interestingly, the ego car fails to overtake the rival due to a more frequent application of the following policy triggered by the unsafe overtaking policy, which ultimately slows down the ego vehicle. This example also demonstrates the potential of KLGame using a multi-modal reference policy that mixes a data-driven policy and an optimization-based policy.

Racing Statistics. Finally, in order to quantitatively assess the performance of our proposed KLGame and MM-KLGame policies, we compare them against two baselines: (1). inverse dynamic game with a neural cost model (the reference policy), and (2). the shielded inverse dynamic game policy. We simulate 100 *randomized* racing scenarios, each with different initial positions on the Thunderhill track (for both the ego and rival vehicles), as well as varying levels of aggressiveness in the rival's defending behaviors. Each scenario is evaluated for each policy using the same random seed. We use the same blending factor λ for both KLGame and MM-KLGame policies across all simulation instances. The safe rate and overtaking rate statistics obtained from these trials are shown in Table II. Since the reference policy is learned from a dataset containing aggressive overtaking demonstrations, it achieves the highest overtaking rate but falls short of the safe rate. The (uni-modal) KLGame policy significantly improves the safe rate by 12% compared to the reference policy, albeit at the cost of a 4% reduction in the overtaking rate. The multi-modal KLGame policy, due to the additional car following mode, achieves the highest safety rate (95%) among the four policies, while maintaining a

TABLE II: Racing Statistics

Method	Safe Rate ↑	Overtaking Rate↑
Inverse Dynamic Game [17, 38]	0.73	**1.00**
Shielded Inverse Dynamic Game	0.90	0.86
KLGame (Ours)	0.85	0.96
Multi-modal KLGame (Ours)	**0.95**	0.92

decent overtaking rate (92%). Finally, the shielded inverse game policy, despite reaching a high safe rate, is ultimately the most conservative policy due to the planner-agnostic safety filter.

Key Takeaway. Our blended multi-modal KLGame policy can balance between competitiveness and cautiousness in *fast and rivalrous* interactions such as car racing.

C. Scene-Consistent Simulated Agents in Waymax

Modern trajectory forecasting models [6, 62, 97] learn natural driving behaviors from data and generate accurate forecasts across different scenarios without the need for manual feature selection. However, these models may encounter difficulties when they are directly used to simulate interactive multi-agent scenarios, as the predicted marginal trajectories may not be compatible with others [98]. In our experiment, we demonstrate the ability of the proposed KLGame to capitalize on data-driven priors for generating natural and scene-consistent reference policies for interactive and safety-critical driving scenarios involving multiple agents.

MTR Reference Policy. In this experiment, we utilize a pre-trained MTR [62] motion prediction model, which outputs trajectories represented by a Gaussian mixture model (GMM) as the reference policy for KLGame. We then use the MTR-guided reference policy to emulate agent behaviors on Waymax [99] with the interactive validation dataset of Waymo Open Motion Dataset [42]. Because the MTR model only outputs a GMM of future trajectories for each agent, we estimate the control sequences leading to the mean trajectory of the corresponding modes via optimization. The reference policy can be constructed from a mixture of control sequences with a fixed covariance and mode probabilities from MTR. This step can be skipped if the model can directly generate a policy given the state.

KLGame on Waymax. For our implementation of KLGame on Waymax, we use *only two* generic optimization targets as cost functions: collision distance and control magnitude regularization, with the rest of the agent's optimization objective coming from the MTR reference policy guidance. Alongside the control regularization, each agent is also tasked with maximizing its signed distance to the closest players and off-road areas. These objectives prompt the agents to prevent collisions effectively and adhere to the road. The signed distance between agents is calculated as the Minkowski difference between two footprints, while the distance to the off-road areas is obtained from road boundaries. In our experiments, all agents are replanned every second in a receding horizon fashion with a 2-second planning horizon and 0.1s discrete timestep.

The first scenario (Fig. 9) consists of four agents in the game, where one of the agents (vehicle 1) merges into the oncoming traffic from behind a parked vehicle (vehicle 28). In doing so, vehicle 1 obstructs vehicle 0's path. The socially optimal trajectories given by KLGame at $t=0$ are such that vehicle 0's modifies its trajectory to slightly veer left and decelerates agent 1. At the replanning stage $t=1$, despite MTR's reference policy leading vehicle 9 towards the edge of the road at the end of the planning horizon, KLGame manages to optimize the trajectory, ensuring it adheres to the lane center, without requiring explicit information about the lane and reference path. The simulated scenario using the proposed KLGame maintains the scene collision-free throughout the episode, while marginal prediction from MTR leads to a collision between vehicles 0 and 1 at $t=1.8s$.

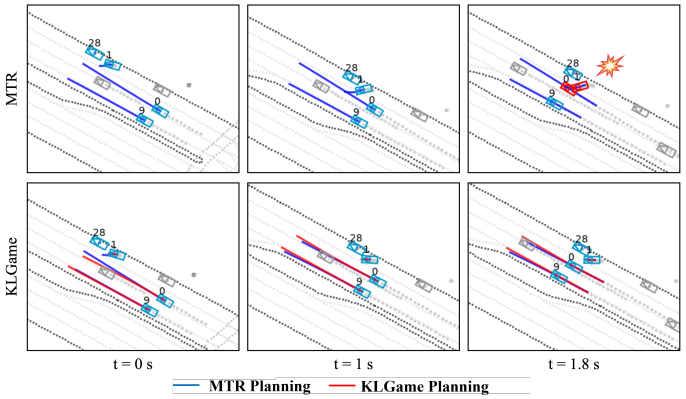

Fig. 9: As car 1 pulls out from the parked vehicle, KLGame optimizes the joint trajectories, guiding both car 0 and car 1 to exhibit yielding behaviors to prevent a collision. However, the reference MTR policy, in contrast, results in a collision.

In the second scenario (Fig. 10), vehicle 0 merges onto the main road, interacting with vehicle 1. The standard MTR policy attempts to slow down vehicle 0, but results in a collision. In contrast, KLGame refines the joint trajectories, planning for vehicle 0 to change lanes, thereby avoiding a collision. While the aggressive evasive maneuver for vehicle 1 at $t=3$ initially risks taking the vehicle off the road, the reference policy regularization combined with the optimization goal assist vehicle 1 in maintaining the proper course on subsequent replanning stages.

Finally, Fig. 11 shows that KLGame can yield diverse, yet scene-consistent, traffic plans, particularly when paired with a multi-modal reference policy. In this intersection scenario, agent 44 approaches the intersection from the wrong side of the road. This unlikely event leads to the marginal predictions from MTR having a hard time keeping the scenario collision-free. In the first rollout, as vehicle 44 opts for a mode to proceed straight, KLGame plans emergency maneuvers for both vehicles to avoid collision. In the subsequent rollout, vehicle 44 samples the mode of turning right. Despite a proximate interaction, KLGame computes the optimal joint trajectory, thereby successfully preventing a collision.

Waymax Statistics. We compare our proposed KLGame method against four baselines: (1) ILGames [5], (2) MaxEnt dynamic game [4], (3) the Intelligent Driver Model (IDM) [100], and (4) MTR [62]. To measure the robustness of each planner to modeling errors, we evaluated each method against an IDM agent in 500 scenarios from the Waymo Open Motion Dataset [101] *interactive validation* split, which is composed of especially challenging interactive scenarios selected by [102]. The safe rate, collision rate, and ADE along each trajectory are reported in Table III. Since the ILQGames and MaxEnt dynamic game baselines do not use a multi-modal

Method	Collision Rate ↓	ADE [3s] (m) ↓	ADE [5s] (m) ↓	ADE [8s] (m) ↓	ADE [avg] (m)
ILQGames [5]	0.13	0.23	1.02	3.34	1.53
MaxEnt [4]	0.13	0.23	1.02	3.34	1.53
IDM [100]	0.10	0.81	3.03	8.04	3.96
MTR [62]	0.12	0.15	0.70	2.19	1.01
KLGame (Ours)	0.09	0.18	0.80	2.45	1.14

TABLE III: Results for a KLGame ego-agent planner against several other ego-agent baselines on Waymax [99] when the other agent follows the IDM [100] model.

reference policy, we instead incorporate reference tracking via the most likely MTR mode. The multimodal KLGame significantly improves on the unimodal ILQGames and MaxEnt dynamic game. Furthermore, KLGame achieves the lowest collision rate among all methods, including data-driven MTR [62] while maintaining the second-lowest ADE scores by a large margin. With only a 12% increase in ADE ($2.45\ m$ with KLGame versus $2.19\ m$ with MTR), KLGame reduces the number of collisions by 25% (45 with KLGame versus 60 with MTR). Overall, these results suggest that the multi-modal policy solved by KLGame encourages safety even when the other agents in the game are not modeled perfectly.

Key Takeaway. The KLGame planner allows integrating the state-of-the-art large-scale, data-driven behavior model with an optimization-based game policy to improve safety for complex tasks such as urban autonomous driving *at scale*.

VI. Limitations and Future Work

In this paper, we empirically demonstrate that, by tuning the KL-regularization weight λ, we can trade off task performance with data-driven (human) behaviors. However, we have not yet delved into what the "optimal" value of λ should be and how to find it. Such questions become more relevant in human-AI shared autonomy, which can naturally benefit from applying KLGame to ease the issue of *automation surprise* [103, 104] via blending the AI policy with the data-driven human policy. We believe KLGame can be extended to account for human expectations by identifying the human-preferred value of λ via, e.g., differentiable game-theoretic planning [17, 37, 38] or inference-based planning techniques [9, 30, 105, 106]. In addition, motivated by the growing need for *interactive* driving datasets and benchmarks, we see an open opportunity to use KLGame for generating scene-consistent and interaction-rich traffic predictions for autonomous driving.

VII. Conclusion

In this work, we propose KLGame, an interactive planning and prediction framework that blends data-driven priors with task-optimal behaviors in a game-theoretic setting. By incorporating information about the uncertainty of other agents' intents—represented by a multi-modal probability distribution—KLGame enables a robot to plan for multiple future scenarios while also providing a mechanism for plan generation through random sampling. Through detailed simulations and real-world autonomous driving data, we demonstrate KLGame's ability to incorporate both optimization-based and data-driven priors into robot motion planning. In essence, KLGame represents a significant stride in integrating data-driven priors with game-theoretic planning, showcasing its prowess in the realm of autonomous driving. Beyond the scope of self-driving cars, the general philosophy of KLGame opens doors to broader applications in human-centered robotics and artificial intelligence.

References

[1] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1998.

[2] J. L. V. Espinoza, A. Liniger, W. Schwarting, D. Rus, and L. Van Gool, "Deep interactive motion prediction and planning: Playing games with motion prediction models," in *L4DC*. PMLR, 2022, pp. 1006–1019.

[3] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas, "Safe planning in dynamic environments using conformal prediction," *IEEE Robotics and Automation Letters*, 2023.

[4] N. Mehr, M. Wang, M. Bhatt, and M. Schwager, "Maximum-entropy multi-agent dynamic games: Forward and inverse solutions," *IEEE Transactions on Robotics*, 2023.

[5] D. Fridovich-Keil, E. Ratner, L. Peters, A. D. Dragan, and C. J. Tomlin, "Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum differential games," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 1475–1481.

[6] S. Shi, L. Jiang, D. Dai, and B. Schiele, "MTR++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying," *arXiv preprint arXiv:2306.17770*, 2023.

[7] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus, "Social behavior for autonomous vehicles," *PNAS*, vol. 116, no. 50, pp. 24 972–24 978, 2019.

[8] M. Wang, Z. Wang, J. Talbot, J. C. Gerdes, and M. Schwager, "Game-theoretic planning for self-driving cars in multivehicle competitive scenarios," *IEEE T-RO*, vol. 37, no. 4, pp. 1313–1325, 2021.

[9] H. Hu, D. Isele, S. Bae, and J. F. Fisac, "Active uncertainty reduction for safe and efficient interaction planning: A shielding-aware dual control approach," *The International Journal of Robotics Research*, 2023.

[10] M. Sun, F. Baldini, P. Trautman, and T. Murphey, "Move beyond trajectories: Distribution space coupling for crowd navigation," ser. Robotics: Science and Systems, 2021.

[11] Z. Williams, J. Chen, and N. Mehr, "Distributed potential ilqr: Scalable game-theoretic trajectory planning for multi-agent interactions," *arXiv preprint arXiv:2303.04842*, 2023.

[12] H. Hu, K. Gatsis, M. Morari, and G. J. Pappas, "Non-cooperative distributed MPC with iterative learning," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 5225–5232, 2020.

[13] R. Spica, E. Cristofalo, Z. Wang, E. Montijano, and M. Schwager, "A real-time game theoretic planner for autonomous two-player drone racing," *IEEE Transactions on Robotics*, vol. 36, no. 5, pp. 1389–1403, 2020.

[14] S. Musić and S. Hirche, "Haptic shared control for human-robot collaboration: a game-theoretical approach," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 10 216–10 222, 2020.

[15] P. Geiger and C.-N. Straehle, "Learning game-theoretic models of multiagent trajectories using implicit layers," in *AAAI*, vol. 35, no. 6, 2021, pp. 4950–4958.

[16] S. Le Cleac'h, M. Schwager, Z. Manchester, *et al.*, "ALGAMES: A fast solver for constrained dynamic games," in *RSS*, 2020.

[17] X. Liu, L. Peters, and J. Alonso-Mora, "Learning to play trajectory games against opponents with unknown objectives," *IEEE Robotics and Automation Letters*, 2023.

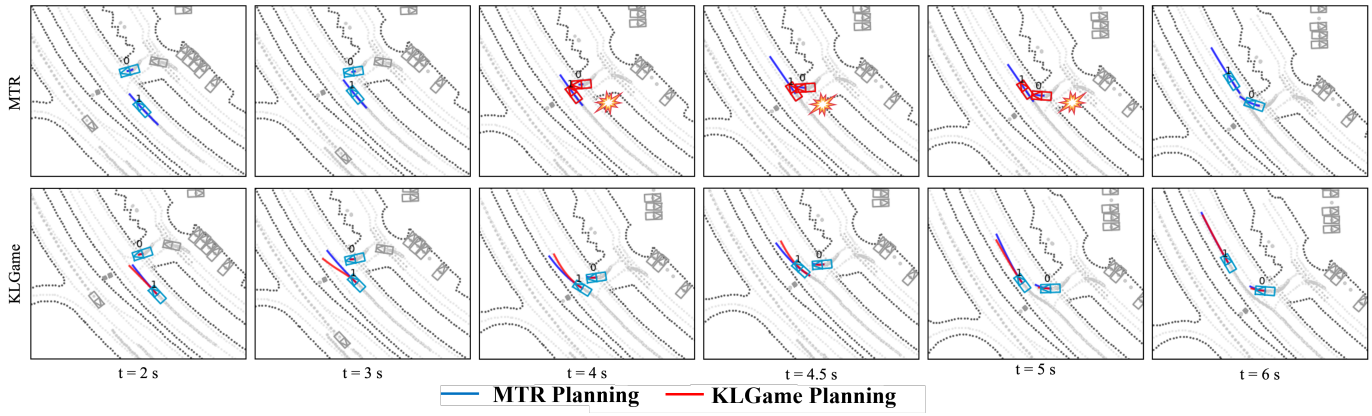

Fig. 10: As car 1 merges into the main road, the MTR-derived policy slows down car 0 but ultimately results in a collision. In comparison, KLGame optimizes the joint trajectories by planning for car 0 to switch lanes and yield, thus avoiding a collision.

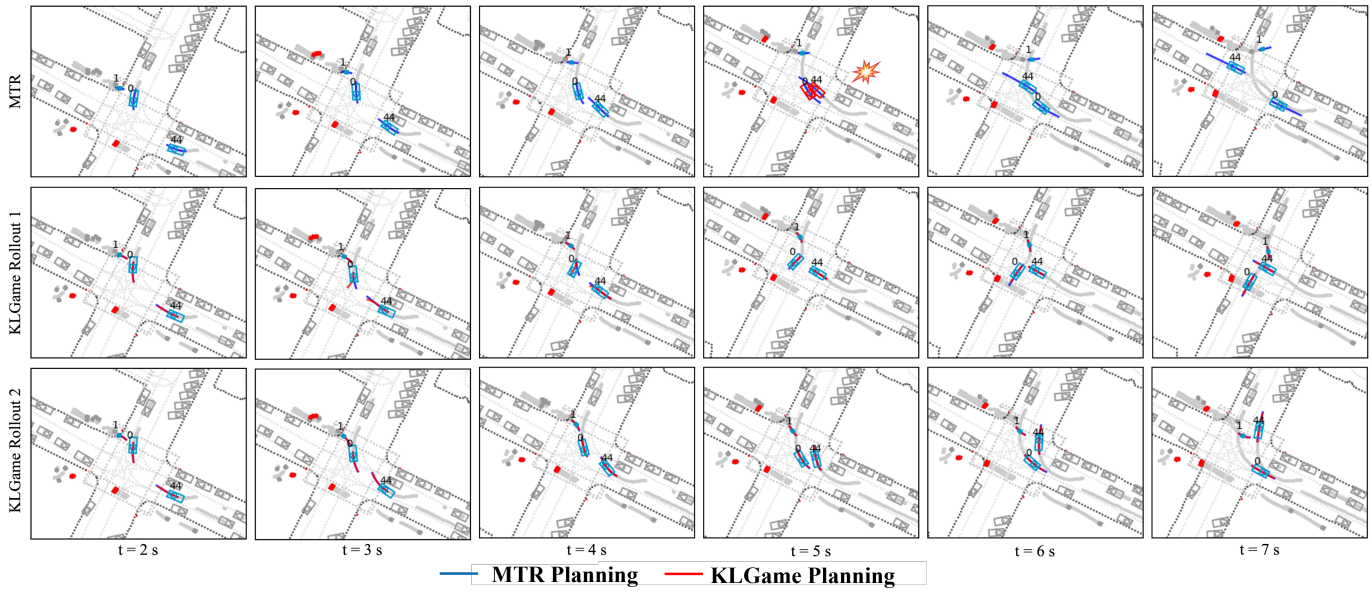

Fig. 11: Car 44 approaches the intersection from the incorrect side of the road. In the first rollout, car 44 opts for a mode to proceed straight, causing KLGame to plan an emergency maneuver to circumvent collision. In the subsequent rollout, car 44 samples the mode of turning right. Despite a proximate interaction, KLGame computes the optimal joint trajectory, thereby successfully averting a collision. This demonstrates KLGame's ability to generate diverse trajectories with a multi-modal reference policy.

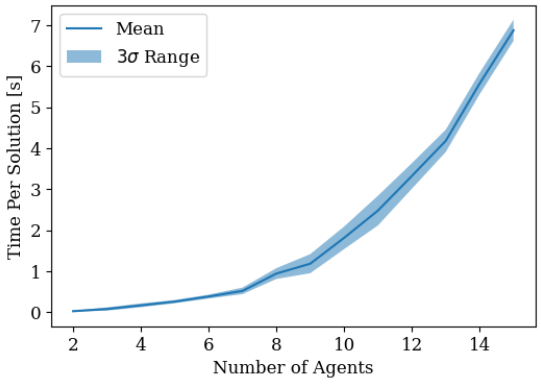

Fig. 12: We examined the worst case runtime and its 3-σ limits for KLGame with different number agents. In each trial, We repeated each trails for 80 times and ensure that the KLGame solver iterates 15 times with 15 line searches.

[18] S. Le Cleac'h, M. Schwager, and Z. Manchester, "LUCIDgames: Online unscented inverse dynamic games for adaptive trajectory prediction and planning," *IEEE RA-L*, vol. 6, no. 3, pp. 5485–5492, 2021.

[19] J. F. Fisac, E. Bronstein, E. Stefansson, D. Sadigh, S. S. Sastry, and A. D. Dragan, "Hierarchical Game-Theoretic planning for autonomous vehicles," in *ICRA*, May 2019, pp. 9590–9596.

[20] H. Hu and J. F. Fisac, "Active uncertainty reduction for human-robot interaction: An implicit dual control approach," in *Algorithmic Foundations of Robotics XV*, 2022, pp. 385–401.

[21] W. Schwarting, A. Pierson, S. Karaman, and D. Rus, "Stochastic dynamic games in belief space," *IEEE T-RO*, vol. 37, no. 6, pp. 2157–2172, 2021.

[22] F. Steinberger, R. Schroeter, and C. N. Watling, "From road distraction to safe driving: Evaluating the effects of boredom and gamification on driving behaviour, physiological arousal, and subjective experience," *Computers in Human Behavior*, vol. 75, pp. 714–726, 2017.

[23] A. Talebpour, H. S. Mahmassani, and S. H. Hamdar, "Modeling

lane-changing behavior in a connected environment: A game theory approach," *Transportation Research Procedia*, vol. 7, pp. 420–440, 2015.
[24] H. Hu, Z. Zhang, K. Nakamura, A. Bajcsy, and J. F. Fisac, "Deception game: Closing the safety-learning loop in interactive robot autonomy," in *Conference on Robot Learning*. PMLR, 2023, pp. 3830–3850.
[25] K. L. Young, S. Koppel, and J. L. Charlton, "Toward best practice in human machine interface design for older drivers: A review of current design guidelines," *Accident Analysis & Prevention*, vol. 106, pp. 460–467, 2017.
[26] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, *et al.*, "Maximum entropy inverse reinforcement learning." in *AAAI*, vol. 8, 2008, pp. 1433–1438.
[27] M. Wulfmeier, D. Rao, D. Z. Wang, P. Ondruska, and I. Posner, "Large-scale cost function learning for path planning using deep inverse reinforcement learning," *The International Journal of Robotics Research*, vol. 36, no. 10, pp. 1073–1087, 2017.
[28] B. Evens, M. Schuurmans, and P. Patrinos, "Learning mpc for interaction-aware autonomous driving: A game-theoretic approach," in *2022 European Control Conference (ECC)*, 2022, pp. 34–39.
[29] T. Phan-Minh, F. Howington, T.-S. Chu, S. U. Lee, M. S. Tomov, N. Li, C. Dicle, S. Findler, F. Suarez-Ruiz, R. Beaudoin, *et al.*, "Driving in real life with inverse reinforcement learning," *arXiv preprint arXiv:2206.03004*, 2022.
[30] L. Peters, D. Fridovich-Keil, C. J. Tomlin, and Z. N. Sunberg, "Inference-based strategy alignment for general-sum differential games," *arXiv preprint arXiv:2002.04354*, 2020.
[31] O. So, Z. Wang, and E. A. Theodorou, "Maximum entropy differential dynamic programming," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3422–3428.
[32] O. So, P. Drews, T. Balch, V. Dimitrov, G. Rosman, and E. A. Theodorou, "MPOGames: Efficient multimodal partially observable dynamic games," in *ICRA*. IEEE, 2023, pp. 3189–3196.
[33] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, "Learning policies for partially observable environments: Scaling up," in *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 362–370.
[34] L. Peters, A. Bajcsy, C.-Y. Chiu, D. Fridovich-Keil, F. Laine, L. Ferranti, and J. Alonso-Mora, "Contingency games for multi-agent interaction," *arXiv preprint arXiv:2304.05483*, 2023.
[35] Y. Chen, U. Rosolia, W. Ubellacker, N. Csomay-Shanklin, and A. D. Ames, "Interactive multi-modal motion planning with branch model predictive control," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5365–5372, 2022.
[36] H. Hu, K. Nakamura, K.-C. Hsu, N. E. Leonard, and J. F. Fisac, "Emergent coordination through game-induced nonlinear opinion dynamics," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 8122–8129.
[37] L. Peters, D. Fridovich-Keil, L. Ferranti, C. Stachniss, J. Alonso-Mora, and F. Laine, "Learning mixed strategies in trajectory games," in *Proc. of Robotics: Science and Systems (RSS)*, 2022.
[38] J. Li, C.-Y. Chiu, L. Peters, S. Sojoudi, C. Tomlin, and D. Fridovich-Keil, "Cost inference for feedback dynamic games from noisy partial state observations and incomplete trajectories," *arXiv preprint arXiv:2301.01398*, 2023.
[39] C. Diehl, T. Klosek, M. Krueger, N. Murzyn, T. Osterburg, and T. Bertram, "Energy-based potential games for joint motion forecasting and control," in *7th Annual Conference on Robot Learning*, 2023.
[40] E. Todorov, "Optimality principles in sensorimotor control," *Nature neuroscience*, vol. 7, no. 9, pp. 907–915, 2004.
[41] F. Bartumeus and S. A. Levin, "Fractal reorientation clocks: Linking animal behavior to statistical patterns of search," *Proceedings of the National Academy of Sciences*, vol. 105, no. 49, pp. 19 072–19 077, 2008.
[42] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *CVPR*, 2021, pp. 9710–9719.
[43] J. Kim and I. Yang, "Hamilton-jacobi-bellman equations for maximum entropy optimal control," *arXiv preprint arXiv:2009.13097*, 2020.
[44] E. A. Theodorou and E. Todorov, "Relative entropy and free energy dualities: Connections to path integral and kl control," in *2012 ieee 51st ieee conference on decision and control (cdc)*. IEEE, 2012, pp. 1466–1473.
[45] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *International conference on machine learning*. PMLR, 2017, pp. 1352–1361.
[46] D. Garg, S. Chakraborty, C. Cundy, J. Song, and S. Ermon, "Iq-learn: Inverse soft-q learning for imitation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4028–4039, 2021.
[47] E. Todorov, "Linearly-solvable markov decision problems," *Advances in neural information processing systems*, vol. 19, 2006.
[48] ——, "Compositionality of optimal control laws," *Advances in neural information processing systems*, vol. 22, 2009.
[49] P. Guan, M. Raginsky, and R. M. Willett, "Online markov decision processes with kullback–leibler control cost," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1423–1438, 2014.
[50] K. Ito and K. Kashima, "Kullback–leibler control for discrete-time nonlinear systems on continuous spaces," *SICE Journal of Control, Measurement, and System Integration*, vol. 15, no. 2, pp. 119–129, 2022.
[51] J. Ok, A. Proutiere, and D. Tranos, "Exploration in structured reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[52] R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, A. Michi, *et al.*, "Nash learning from human feedback," *arXiv preprint arXiv:2312.00886*, 2023.
[53] D. Bernardini and A. Bemporad, "Stabilizing model predictive control of stochastic constrained linear systems," *IEEE Trans. Autom. Control*, vol. 57, no. 6, pp. 1468–1480, 2011.
[54] M. C. Campi, S. Garatti, and F. A. Ramponi, "A general scenario theory for nonconvex optimization and decision making," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4067–4078, 2018.
[55] G. Schildbach and F. Borrelli, "Scenario model predictive control for lane change assistance on highways," in *IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 611–616.
[56] H. Hu, K. Nakamura, and J. F. Fisac, "SHARP: Shielding-aware robust planning for safe and efficient human-robot interaction," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5591–5598, 2022.
[57] J. Li, C.-Y. Chiu, L. Peters, F. Palafox, M. Karabag, J. Alonso-Mora, S. Sojoudi, C. Tomlin, and D. Fridovich-Keil, "Scenario-game admm: A parallelized scenario-based solver for stochastic noncooperative games," *arXiv preprint arXiv:2304.01945*, 2023.
[58] B. Eysenbach and S. Levine, "Maximum entropy RL (provably) solves some robust RL problems," *arXiv preprint arXiv:2103.06257*, 2021.
[59] Z. Wu, L. Sun, W. Zhan, C. Yang, and M. Tomizuka, "Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5355–5362, 2020.
[60] S. Pitis, H. Chan, S. Zhao, B. Stadie, and J. Ba, "Maximum entropy gain exploration for long horizon multi-goal reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7750–7761.
[61] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.
[62] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," *arXiv preprint arXiv:2209.13508*, 2022.
[63] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *ECCV*. Springer, 2020, pp. 683–700.
[64] Z. Zhou, L. Ye, J. Wang, K. Wu, and K. Lu, "Hivt: Hierarchical vector transformer for multi-agent motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8823–8833.
[65] X. Jia, P. Wu, L. Chen, Y. Liu, H. Li, and J. Yan, "Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding," *IEEE transactions on pattern analysis and machine intelligence*, 2023.
[66] A. Seff, B. Cera, D. Chen, M. Ng, A. Zhou, N. Nayakanti, K. S. Refaat, R. Al-Rfou, and B. Sapp, "Motionlm: Multi-agent motion forecasting as language modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8579–8590.
[67] C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, D. Anguelov, *et al.*, "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9644–9653.
[68] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, D. Weiss, B. Sapp, Z. Chen, and J. Shlens, "Scene transformer: A unified architecture for predicting multiple agent trajectories," in *ICLR*, June 2021.

[69] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov, and B. Sapp, "MultiPath++: Efficient information fusion and trajectory aggregation for behavior prediction," Nov. 2021.

[70] S. Kumar, Y. Gu, J. Hoang, G. C. Haynes, and M. Marchetti-Bowick, "Interaction-based trajectory prediction over a hybrid traffic graph," in *IROS*. IEEE, 2021, pp. 5530–5535.

[71] Q. Sun, X. Huang, J. Gu, B. C. Williams, and H. Zhao, "M2I: From factored marginal trajectory prediction to interactive prediction," in *CVPR*, 2022, pp. 6543–6552.

[72] Y. Ban, X. Li, G. Rosman, I. Gilitschenski, O. Meireles, S. Karaman, and D. Rus, "A deep concept graph network for interaction-aware trajectory prediction," in *ICRA*. IEEE, 2022, pp. 8992–8998.

[73] J. Lidard, O. So, Y. Zhang, J. DeCastro, X. Cui, X. Huang, Y.-L. Kuo, J. Leonard, A. Balachandran, N. Leonard, and G. Rosman, "Nashformer: Leveraging local nash equilibria for semantically diverse trajectory prediction," 2023.

[74] Z. Huang, H. Liu, and C. Lv, "Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving," *arXiv preprint arXiv:2303.05760*, 2023.

[75] X. Huang, S. G. McGill, J. A. DeCastro, L. Fletcher, J. J. Leonard, B. C. Williams, and G. Rosman, "DiversityGAN: Diversity-aware vehicle motion prediction via latent semantic sampling," *IEEE RA-L*, vol. 5, no. 4, pp. 5089–5096, 2020.

[76] S. Shiroshita, S. Maruyama, D. Nishiyama, M. Y. Castro, K. Hamzaoui, G. Rosman, J. DeCastro, K.-H. Lee, and A. Gaidon, "Behaviorally diverse traffic simulation via reinforcement learning," in *IROS*. IEEE, 2020, pp. 2103–2110.

[77] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, *et al.*, "TNT: Target-driven trajectory prediction," in *Conference on Robot Learning*. PMLR, 2021, pp. 895–904.

[78] X. Huang, G. Rosman, I. Gilitschenski, A. Jasour, S. G. McGill, J. J. Leonard, and B. C. Williams, "HYPER: Learned hybrid trajectory prediction via factored inference and adaptive sampling," in *ICRA*. IEEE, 2022, pp. 2906–2912.

[79] A. Dixit, L. Lindemann, S. X. Wei, M. Cleaveland, G. J. Pappas, and J. W. Burdick, "Adaptive conformal prediction for motion planning among dynamic agents," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 300–314.

[80] Z. Huang, H. Liu, J. Wu, and C. Lv, "Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving," *IEEE transactions on neural networks and learning systems*, 2023.

[81] Z. Zhong, D. Rempe, D. Xu, Y. Chen, S. Veer, T. Che, B. Ray, and M. Pavone, "Guided conditional diffusion for controllable traffic simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3560–3566.

[82] Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu, M. Pavone, and B. Ray, "Language-guided traffic simulation via scene-level diffusion," in *Conference on Robot Learning*. PMLR, 2023, pp. 144–177.

[83] Z. Huang, Z. Zhang, A. Vaidya, Y. Chen, C. Lv, and J. F. Fisac, "Versatile scene-consistent traffic scenario generation as optimization with diffusion," *arXiv preprint arXiv:2404.02524*, 2024.

[84] E. Çinlar, *Probability and stochastics*. Springer, 2011.

[85] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. iv," *Communications on pure and applied mathematics*, vol. 36, no. 2, pp. 183–212, 1983.

[86] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*. John Wiley & Sons, 2011.

[87] P. Dupuis, "Representations and weak convergence methods for the analysis and approximation of rare events," *Padova notes*, 2019.

[88] F. Laine, D. Fridovich-Keil, C.-Y. Chiu, and C. Tomlin, "The computation of approximate generalized feedback nash equilibria," *SIAM Journal on Optimization*, vol. 33, no. 1, pp. 294–318, 2023.

[89] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[90] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[91] A. Mesbah, "Stochastic model predictive control: An overview and perspectives for future research," *IEEE Control Systems Magazine*, vol. 36, no. 6, pp. 30–44, 2016.

[92] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs," 2018. [Online]. Available: http://github.com/google/jax

[93] X. Zhang, A. Liniger, and F. Borrelli, "Optimization-based collision avoidance," *IEEE Transactions on Control Systems Technology*, vol. 29, no. 3, pp. 972–983, 2020.

[94] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[95] 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.

[96] K.-C. Hsu, H. Hu, and J. F. Fisac, "The safety filter: A unified view of safety-critical control in autonomous systems," *Annual Review of Control, Robotics, and Autonomous Systems*, 2023.

[97] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," July 2022.

[98] Y. Chen, B. Ivanovic, and M. Pavone, "Scept: Scene-consistent, policy-based trajectory predictions for planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 103–17 112.

[99] C. Gulino, J. Fu, W. Luo, G. Tucker, E. Bronstein, Y. Lu, J. Harb, X. Pan, Y. Wang, X. Chen, *et al.*, "Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research," *arXiv preprint arXiv:2310.08710*, 2023.

[100] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[101] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[102] L. Zhang, Z. Peng, Q. Li, and B. Zhou, "Cat: Closed-loop adversarial training for safe end-to-end driving," in *Conference on Robot Learning*. PMLR, 2023, pp. 2357–2372.

[103] N. B. Sarter and D. D. Woods, "Team play with a powerful and independent agent: Operational experiences and automation surprises on the Airbus A-320," vol. 39, no. 4, pp. 553–569.

[104] G. A. Jamieson, G. Skraaning, and J. Joe, "The B737 MAX 8 accidents as operational experiences with automation transparency," vol. 52, no. 4.

[105] J. F. Fisac, A. Bajcsy, S. L. Herbert, D. Fridovich-Keil, S. Wang, C. J. Tomlin, and A. D. Dragan, "Probabilistically safe robot planning with confidence-based human predictions," in *Robotics: Science and Systems*, 2018.

[106] R. Tian, L. Sun, A. Bajcsy, M. Tomizuka, and A. D. Dragan, "Safety assurances for human-robot interaction via confidence-aware game-theoretic human models," in *ICRA*. IEEE, 2022, pp. 11 229–11 235.

[107] S. Aggarwal, M. Bastopcu, T. Başar, *et al.*, "Policy Optimization finds Nash Equilibrium in Regularized General-Sum LQ Games," *arXiv preprint arXiv:2404.00045*, 2024.

APPENDIX

A. Proof of Lemma 1

By a Lagrange multiplier argument, we guess the minimizer of (6) is given by

$$\frac{d\pi^*}{d\tilde{\pi}} = \frac{e^{-\mathcal{Q}(x,u)/\lambda}}{\int_{\mathcal{U}} e^{-\mathcal{Q}(x,u)/\lambda} d\tilde{\pi}}. \tag{20}$$

Next, we rearrange the term inside the infimum, which we denote by I^π.

$$\begin{aligned}
I^\pi &= \mathbb{E}^\pi[\mathcal{Q}(x,u)] + \lambda D_{KL}(\pi||\tilde{\pi}) \\
&= \int_{\mathcal{U}} \mathcal{Q}(x,u) d\pi + \lambda D_{KL}(\pi||\tilde{\pi}) \\
&= \int_{\mathcal{U}} \mathcal{Q}(x,u) d\pi + \lambda \int_{\mathcal{U}} \log\left(\frac{d\pi}{d\tilde{\pi}}\right) d\pi \\
&= \int_{\mathcal{U}} \mathcal{Q}(x,u) d\pi + \lambda \int_{\mathcal{U}} \log\left(\frac{d\pi}{d\pi^*}\frac{d\pi^*}{d\tilde{\pi}}\right) d\pi \\
&= \lambda \int_{\mathcal{U}} \log(e^{(\mathcal{Q}(x,u)/\lambda)}) d\pi + \lambda \int_{\mathcal{U}} \log\left(\frac{d\pi^*}{d\tilde{\pi}}\right) d\pi + \lambda D_{KL}(\pi||\pi^*) \\
&= \lambda \int_{\mathcal{U}} \log\left(e^{(\mathcal{Q}(x,u)/\lambda} \times \frac{e^{-\mathcal{Q}(x,u)/\lambda}}{\int_{\mathcal{U}} e^{-\mathcal{Q}(x,u)/\lambda} d\tilde{\pi}}\right) d\pi + \lambda D_{KL}(\pi||\pi^*) \\
&= -\lambda \int_{\mathcal{U}} \log\left(\int_{\mathcal{U}} e^{-\mathcal{Q}(x,u)/\lambda} d\tilde{\pi}\right) d\pi + \lambda D_{KL}(\pi||\pi^*) \\
&= -\lambda \log\left(\int_{\mathcal{U}} e^{-\mathcal{Q}(x,u)/\lambda} d\tilde{\pi}\right) \underbrace{\int_{\mathcal{U}} d\pi}_{=1} + \lambda D_{KL}(\pi||\pi^*) \\
&\geq -\lambda \log\left(\int_{\mathcal{U}} e^{-\mathcal{Q}(x,u)/\lambda} d\tilde{\pi}\right),
\end{aligned} \tag{21}$$

where we substitute the definition (20) in the penultimate equality. Note that if $\pi << \tilde{\pi}$ and $\tilde{\pi} << \pi^*$, then $\pi << \pi^*$. Recalling that $\lambda D_{KL}(\pi||\pi^*) \geq 0$, we attain the infimum $\inf_\pi I^\pi$ exactly when $\pi = \pi^*$.

B. Proof of Theorem 1

We restate the theorem for convenience.

The N-player nonzero-sum KL-LQG dynamic game (4) admits a unique global feedback Nash equilibrium solution if,

1) The dynamics follow the functional form

$$x_{t+1} = A_t x_t + \sum_{i\in[N]} B_t^i u_t^i + d_t,$$

where $x_0 \sim \mathcal{N}(\mu_{x_0}, \Sigma_{x_0})$, $d_t \sim \mathcal{N}(0, \Sigma_d)$,

2) The costs have the functional form

$$J^i(\pi^i) = \mathbb{E}^\pi\left[\sum_{t=0}^{T} \frac{1}{2}\left(x_t^\top Q_t^i x_t + \sum_{j\in[N]} u_t^{j\top} R_t^{ij} u_t^j\right) + \sum_{t=0}^{T} \lambda^i D_{KL}(\pi_t^i||\tilde{\pi}_t^i)\right],$$

where, for all $t \in [T]$, $Q_t^i \succeq 0$, $R_t^{ij} \succeq 0$, $\forall i,j \in [N], j \neq i$, $R_t^{ii} \succ 0$, $\forall i \in [N]$,

3) The reference policies are Gaussian, i.e., $\tilde{\pi}_t^i \sim \mathcal{N}(\tilde{\mu}_t^i, \tilde{\Sigma}_t^i)$ for all $t \in [T]$ and $i \in [N]$.

Moreover, the global (mixed-strategy) Nash equilibrium is a set of time-varying policies $\pi_t^{i*} = \mathcal{N}(\mu_t^{i*}, \Sigma_t^{i*})$, $\forall i \in [N]$ with mean and covariance given by

$$\begin{aligned}
\mu_t^{i*} &= -K_t^i x_t - \kappa_t^i, \\
\Sigma_t^{i*} &= \left[\frac{1}{\lambda^i}\left(R_t^{ii} + B_t^{i^T} Z_{t+1}^i B_t^i\right) + \left(\tilde{\Sigma}_t^i\right)^{-1}\right]^{-1},
\end{aligned} \tag{22}$$

where (K_t^i, κ_t^i) is given by solving the coupled KL-regularized Riccati equation:

$$\begin{aligned}
\left[\lambda^i(\tilde{\Sigma}_t^i)^{-1} + R_t^{ii} + B_t^{i\top} Z_{t+1}^i B_t^{i\top}\right] K_t^i + B_t^{i\top} Z_{t+1}^i \sum_{j\neq i} B_t^j K_t^j &= B_t^{i\top} Z_{t+1}^i A_t, \\
\left[\lambda^i(\tilde{\Sigma}_t^i)^{-1} + R_t^{ii} + B_t^{i\top} Z_{t+1}^i B_t^{i\top}\right] \kappa_t^i + B_t^{i\top} Z_{t+1}^i \sum_{j\neq i} B_t^j \kappa_t^j &= B_t^{i\top} z_{t+1}^i - \lambda^i(\tilde{\Sigma}_t^i)^{-1}\tilde{\mu}_t^i.
\end{aligned} \tag{23}$$

Value function parameters (Z_t^i, z_t^i) are computed recursively backward in time as

$$\begin{aligned}
Z_t^i &= Q_t^i + \sum_{j\in[N]} (K_t^j)^T R_t^{ij} K_t^j + F_t^T Z_{t+1}^i F_t + \lambda^i K_t^{i\top}(\tilde{\Sigma}_t^i)^{-1} K_t^i, \\
z_t^i &= \sum_{j\in[N]} (K_t^j)^T R_t^{ij} \kappa_t^j + F_t^T \left(z_{t+1}^i + Z_{t+1}^i \beta_t\right) + \lambda^i K_t^{i\top}(\tilde{\Sigma}_t^i)^{-1}(\kappa_t^i - \tilde{\mu}_t^i),
\end{aligned} \tag{24}$$

with terminal conditions $Z_{T+1}^i = 0$ and $z_{T+1}^i = 0$. Here, $F_t = A_t - \sum_{j\in[N]} B_t^j K_t^j$ and $\beta_t = -\sum_{j\in[N]} B_t^j \kappa_t^j$ for all $t \in [T]$.

Proof: We proceed via induction. The base case $(t = T)$ follows directly from the hypothesis of the cost functional form in the theorem statement. For the induction step, we assume that the state-value function is a quadratic form at time $t+1$,

$$V_{t+1}^{i*} = \frac{1}{2} x_{t+1}^\top Z_{t+1}^i x_{t+1} + z_{t+1}^{i\top} x_t + \eta_{t+1}^i, \tag{25}$$

and we show this holds at time t. Denote $\pi_t = (\pi_t^i, \pi_t^{\neg i})$. Therefore, we may write

$$\begin{aligned}
V_t^{i*}(x_t) &= \min_{\pi_t^i} \mathbb{E}^{\pi_t}\left[\frac{1}{2}\left(x_t^\top Q_t^i x_t + \sum_{j\in[N]} u_t^{j\top} R_t^{ij} u_t^j\right) + \lambda^i D_{KL}\left(\pi_t^i || \tilde{\pi}_t^i\right) + \mathbb{E}^{x_{t+1}}\left[V_{t+1}^{i*}(x_{t+1})\right]\right] \\
&= \min_{\pi_t^i} \mathbb{E}^{\pi_t}\Bigg[\underbrace{\frac{1}{2}\left(x_t^\top Q_t^i x_t + \sum_{j\in[N]} u_t^{j\top} R_t^{ij} u_t^j\right)}_{(1)} + \underbrace{\lambda^i D_{KL}\left(\pi_t^i || \tilde{\pi}_t^i\right)}_{(2)} + \underbrace{\mathbb{E}^{x_{t+1}}\left[V_{t+1}^{i*}\left(A_t x_t + \sum_{j\in[N]} B_t^j u_t^j + d_t\right)\right]}_{(3)}\Bigg].
\end{aligned} \tag{26}$$

Let us now consider each term in the brackets:

$$(1) = \mathbb{E}^{\pi_t}\left[\frac{1}{2}\left(x_t^\top Q_t^i x_t + \sum_{j\in[N]} u_t^{j\top} R_t^{ij} u_t^j\right)\right] = \frac{1}{2} x_t^\top Q_t^i x_t + \frac{1}{2}\left(\sum_{j\in[N]} \mu_t^{j\top} R_t^{ij} \mu_t^j + \mathrm{tr}\left(R_t^{ij}\Sigma_t^j\right)\right), \tag{27}$$

$$(2) = \mathbb{E}^{\pi_t}\left[\lambda^i D_{KL}\left(\pi_t^i || \tilde{\pi}_t^i\right)\right] = \frac{\lambda^i}{2}\left(n_{u^i} - \mathrm{logdet}(\Sigma_t^i) + \mathrm{logdet}(\tilde{\Sigma}_t^i) + \mathrm{tr}\left(\left(\tilde{\Sigma}_t^i\right)^{-1}\Sigma_t^i\right) + \left(\mu_t^i - \tilde{\mu}_t^i\right)^\top \left(\tilde{\Sigma}_t^i\right)^{-1}\left(\mu_t^i - \tilde{\mu}_t^i\right)\right), \tag{28}$$

$$(3) = \mathbb{E}^{\pi_t}\left[\mathbb{E}^{\pi_t, d}\left[\frac{1}{2}\left(A_t x_t + \sum_{j\in[N]} B_t^j u_t^j + d_t\right)^\top Z_{t+1}^i \left(A_t x_t + \sum_{j\in[N]} B_t^j u_t^j + d_t\right) + z_{t+1}^{i\top}\left(A_t x_t + \sum_{j\in[N]} B_t^j u_t^j + d_t\right) + \eta_{t+1}^i\right]\right]. \tag{29}$$

Expanding the terms inside the brackets of Equation (29), taking expectations, and dropping terms not affecting the minimum, (29) yields

$$\frac{1}{2}\left(\sum_{j\in[N]} B_t^j \mu_t^j\right)^\top Z_{t+1}^i \left(\sum_{j\in[N]} B_t^j \mu_t^j\right) + \sum_{j\in[N]} (A_t x_t)^\top Z_{t+1}^i B_t^j \mu_t^j + \frac{1}{2}\mathrm{tr}\left(Z_{t+1}^i\left(\Sigma_d + \sum_{j\in[N]} B_t^{j\top}\Sigma_t^j B_t^j\right)\right) + \sum_{j\in[N]} z_{t+1}^{i\top} B_t^j \mu_t^j. \tag{30}$$

Similarly, combining Equations (27), (28), (30) and dropping terms not affecting the minimum, we have

$$\begin{aligned}
V_t^{i*}(x_t) = &\frac{1}{2}\left(\sum_{j\in[N]} \mu_t^{j\top} R_t^{ij} \mu_t^j + \mathrm{tr}\left(R_t^{ij}\Sigma_t^j\right)\right) \\
&+\frac{\lambda^i}{2}\left(-\mathrm{logdet}(\Sigma_t^i) + \mathrm{logdet}(\tilde{\Sigma}_t^i) + \mathrm{tr}\left(\left(\tilde{\Sigma}_t^i\right)^{-1}\Sigma_t^i\right) + \left(\mu_t^i - \tilde{\mu}_t^i\right)^\top \left(\tilde{\Sigma}_t^i\right)^{-1}\left(\mu_t^i - \tilde{\mu}_t^i\right)\right) \\
&+\frac{1}{2}\left(\sum_{j\in[N]} B_t^j \mu_t^j\right)^\top Z_{t+1}^i \left(\sum_{j\in[N]} B_t^j \mu_t^j\right) + \sum_{j\in[N]} \left(A_t x_t\right)^\top Z_{t+1}^i B_t^j \mu_t^j \\
&+\frac{1}{2}\mathrm{tr}\left(Z_{t+1}^i\left(\Sigma_d + \sum_{j\in[N]} B_t^{j\top}\Sigma_t^j B_t^j\right)\right) + \sum_{j\in[N]} z_{t+1}^{i\top} B_t^j \mu_t^j,
\end{aligned} \tag{31}$$

which is convex quadratic in μ_t^i and convex in Σ_t^i. Therefore, by the first-order optimality conditions, we may take the gradient of Equation (31) and equate it to zero to find the optimal mean controls and the corresponding covariances. The optimal control law is linear state-feedback in the form of

$$\mu_t^{i*} = -K_t^i x_t - \kappa_t^i. \tag{32}$$

Plugging (32) into the gradient, with respect to mean controls, of (31) and equating to zero yields coupled Riccati equations

$$\begin{aligned}
&\left(R_t^{ii} + \lambda^i\left(\tilde{\Sigma}_t^i\right)^{-1} + B_t^{i\top} Z_{t+1} B_t^i\right) K_t^i + B_t^{i\top} Z_{t+1}^i \sum_{i\neq j\in[N]} B_t^j K_t^j = B_t^{i\top} Z_{t+1}^i A_t, \\
&\left(R_t^{ii} + \lambda^i\left(\tilde{\Sigma}_t^i\right)^{-1} + B_t^{i\top} Z_{t+1} B_t^i\right) \kappa_t^i + B_t^{i\top} Z_{t+1}^i \sum_{i\neq j\in[N]} B_t^j \kappa_t^j = z_{t+1}^{i\top} B_t^i - \lambda^i\left(\tilde{\Sigma}_t^i\right)^{-1}\tilde{\mu}_t^i.
\end{aligned} \tag{33}$$

Repeating the same steps above, but this time imposing first-order optimality conditions with respect to the covariance Σ_t^i yields

$$R_t^{ii} + \lambda^i(\tilde{\Sigma}_t^i)^{-1} - \lambda^i(\Sigma_t^i)^{-1} + B_t^{i\top} Z_{t+1}^i B_t^i = 0, \tag{34}$$

thereby giving the optimal covariance as

$$\Sigma_t^{i*} = \left[\frac{1}{\lambda^i}\left(R_t^{ii} + B_t^{i\top} Z_{t+1}^i B_t^i + \lambda^i\left(\tilde{\Sigma}_t^i\right)^{-1}\right)\right]^{-1}. \tag{35}$$

To solve the Riccati equations, it is necessary to derive expressions for Z_t^i and z_t^i, which can be accomplished by taking expectations in (26) and rearranging with the optimal control law of (32). Note that solving for η_t^i is not necessary, as it will not affect the optimal policy. Following the lines of [5], we remove the $\min_{\pi_t^i}$ since we already imposed first-order optimality via our control scheme:

$$\begin{aligned}
V_t^{i*}(x_t) = &\mathbb{E}^{\pi_t^*}\frac{1}{2}\left(x_t^\top Q_t^i x_t + \sum_{j\in[N]}\left(-K_t^j x_t - \kappa_t^j\right)^\top R_t^{ij}\left(-K_t^j x_t - \kappa_t^j\right)\right) \\
&+\lambda^i D_{KL}\left(\pi_t^{i*} || \tilde{\pi}_t^{i*}\right) \\
&+\mathbb{E}^{\pi_t^*}\mathbb{E}^{x_{t+1}}\frac{1}{2}\left(A_t x_t + \sum_{j\in[N]} B_t^j\left(-K_t^j x_t - \kappa_t^j\right) + d_t\right)^\top Z_{t+1}^i\left(A_t x_t + \sum_{j\in[N]} B_t^j\left(-K_t^j x_t - \kappa_t^j\right) + d_t\right) \\
&+\mathbb{E}^{\pi_t^*}\mathbb{E}^{x_{t+1}} z_{t+1}^{i\top}\left(A_t x_t + \sum_{j\in[N]} B_t^j\left(-K_t^j x_t - \kappa_t^j\right) + d_t\right) + \eta_{t+1}^i.
\end{aligned} \tag{36}$$

Taking expectations and rearranging yields the recursion parameters for the quadratic value function, which we now show. Firstly, let

$$\begin{aligned}
F_t &= A_t - \sum_{j\in[N]} B_t^j K_t^j, \\
\beta_t &= -\sum_{j\in[N]} B_t^j \kappa_t^j.
\end{aligned} \tag{37}$$

Then we have

$$
\begin{aligned}
V_t^{i*}(x_t) = \frac{1}{2} x_t^\top \left[Q_t^i + \sum_{j\in[N]} K_t^{j\top} R_t^{ij} K_t^j + F_t^\top Z_{t+1}^i F_t + \lambda^i K_t^{i\top} \left(\tilde{\Sigma}_t^i\right)^{-1} K_t^i \right] x_t \\
+ \left[\sum_{j\in[N]} K_t^{j\top} R_t^{ij} \kappa_t^j + F_t^\top \left(z_{t+1}^i + Z_{t+1}^i \beta_t\right)^\top + \lambda^i K_t^{i\top} \left(\tilde{\Sigma}_t^i\right)^{-1} \left(\kappa_t^i - \tilde{\mu}_t^i\right) \right] x_t \\
+ \cdots,
\end{aligned} \tag{38}
$$

where $\cdots$ denotes the constant values pertaining relevant for calculating the η_t^i terms, but are unnecessary for deriving the optimal policy. The terms in the brackets are precisely the recursions needed:

$$
\begin{aligned}
Z_t^i &= Q_t^i + \sum_{j\in[N]} K_t^{j\top} R_t^{ij} K_t^j + F_t^\top Z_{t+1}^i F_t + \lambda^i K_t^{i\top} \left(\tilde{\Sigma}_t^i\right)^{-1} K_t^i, && Z_{T+1}^i = 0, \\
z_t^i &= \sum_{j\in[N]} K_t^{j\top} R_t^{ij} \kappa_t^j + F_t^\top \left(z_{t+1}^i + Z_{t+1}^i \beta_t\right)^\top + \lambda^i K_t^{i\top} \left(\tilde{\Sigma}_t^i\right)^{-1} \left(\kappa_t^i - \tilde{\mu}_t^i\right), && z_{T+1}^i = 0,
\end{aligned} \tag{39}
$$

thereby completing our proof by construction. Finally, we note that, in the above proof, we have assumed that the coupled Riccati equations are uniquely solvable to ensure a unique FBNE. A more detailed discussion on the solution uniqueness of coupled Riccati equations can be found in concurrent work [107]. ■

C. Proof of Proposition 1

The proof follows the same structure as in Theorem 1. We develop the expressions of (26) with the expressions for (27) and (29) remaining the same. The expression in (28) needs to be modified to account for the time dependence in the reference policy. We have:

$$
\begin{aligned}
(2)\ \ \lambda^i D_{KL}(\pi_t^i \| \tilde{\pi}_t^i) &= \frac{\lambda^i}{2} \left[n_{u^i} - \operatorname{logdet}\Sigma_t^i + \operatorname{logdet}\tilde{\Sigma}_t^i + \operatorname{tr}((\tilde{\Sigma}_t^i)^{-1}\Sigma_t^i) + (\mu^{i*} - \tilde{\mu}^{i*})^\top (\tilde{\Sigma}_t^i)^{-1} (\mu^{i*} - \tilde{\mu}^{i*}) \right] \\
&= \frac{\lambda^i}{2} \left[n_{u^i} - \operatorname{logdet}\Sigma_t^i + \operatorname{logdet}\tilde{\Sigma}_t^i + \operatorname{tr}((\tilde{\Sigma}_t^i)^{-1}\Sigma_t^i) + (\mu^{i*} + \tilde{K}_t^i x + \tilde{\kappa}_t^i)^\top (\tilde{\Sigma}_t^i)^{-1} (\mu^{i*} + \tilde{K}_t^i x + \tilde{\kappa}_t^i) \right] \\
&= \frac{\lambda^i}{2} \Big[n_{u^i} - \operatorname{logdet}\Sigma_t^i + \operatorname{logdet}\tilde{\Sigma}_t^i + \operatorname{tr}((\tilde{\Sigma}_t^i)^{-1}\Sigma_t^i) + \mu^{i*\top} (\tilde{\Sigma}_t^i)^{-1} \mu^{i*} + (\tilde{K}_t^i x_t + \tilde{\kappa}_t^i)^\top (\tilde{\Sigma}_t^i)^{-1} (\tilde{K}_t^i x_t + \tilde{\kappa}_t^i) \\
&\quad + \mu^{i*\top} (\tilde{\Sigma}_t^i)^{-1} (\tilde{K}_t^i x_t + \tilde{\kappa}_t^i) + (\tilde{K}_t^i x_t + \tilde{\kappa}_t^i)^\top (\tilde{\Sigma}_t^i)^{-1} \mu^{i*} \Big].
\end{aligned} \tag{40}
$$

Following the same arguments as in Appendix B, we take derivatives and enforce them to be 0 to obtain the following set of coupled Riccati equations:

$$
\begin{aligned}
\left[\lambda^i (\tilde{\Sigma}^i)^{-1} + R_t^{ii} + B^{i\top} Z_{t+1}^i B^{i\top}\right] K_t^i + B^{i\top} Z_{t+1}^i \sum_{j\neq i} B^j K_t^j &= B^{i\top} Z_{t+1}^i A + \lambda^i (\tilde{\Sigma}_t^i)^{-1} \tilde{K}_t^i, \\
\left[\lambda^i (\tilde{\Sigma}^i)^{-1} + R_t^{ii} + B^{i\top} Z_{t+1}^i B^{i\top}\right] \kappa_t^i + B^{i\top} Z_{t+1}^i \sum_{j\neq i} B^j \kappa_t^j &= B^{i\top} z_{t+1}^i + \lambda^i (\tilde{\Sigma}_t^i)^{-1} \tilde{\kappa}_t^i.
\end{aligned} \tag{41}
$$

Substituting the formulas for μ_t^i back into the expression for $V_t^i(x)$ we obtain the formulas in (15).