Permissive Safety Through Trusted Inference: Verifiable Belief-Space Neural Safety Filters for Assured Interactive Robotics

Haimin Hu[0000−0002−4217−4776]

Department of Computer Science, Johns Hopkins University, USA
haimin@cs.jhu.edu

Abstract. Autonomous robots that interact with people must make safe and efficient decisions under human-induced uncertainty, such as their preferences, goals, competency, and willingness to cooperate. Safety filters are a popular approach for ensuring safety in interactive robotics, since their modular design separates safety from performance, allowing robots to operate safely around people with minimal impact on task efficiency. While traditional safety filters typically operate only in the physical space, neglecting the robot's ability to learn and adapt online, the recently proposed belief-space safety filter (BELIEFSF) reasons about robot safety in closed-loop with runtime inference that actively reduces the robot's uncertainty online, thereby reducing conservativeness in filtering. However, providing formal safety guarantees for robots deploying BELIEFSF remains a significant challenge due to errors in runtime inference and neural approximation of safety filters required to handle the high dimensionality of belief spaces. In this paper, we propose an algorithmic approach to certify high-probability safety of BELIEFSF using conformal prediction, while explicitly accounting for the reliability of the robot's runtime inference module. Our method leverages the structure of belief-space safety filtering by focusing verification on a region where inference is expected to be reliable. It preserves the simplicity and sample complexity of standard conformal prediction, yet can certify a substantially less conservative safety filter. Through a simulated human–vehicle interaction benchmark, we show that our approach verifies a significantly more permissive belief-space safety filter than a standard conformal prediction baseline.

Keywords: Neural Safety Filters · Safety Verification · Conformal Prediction · Assured Interaction · Belief-Space Motion Planning

1 Introduction

Autonomous robots are increasingly deployed in safety-critical and human-populated environments, from social navigation to mixed-autonomy traffic and human–robot teaming in warehouses, where human intentions are unobserved and their behaviors may be strongly influenced by the robot's actions. In such settings, safety involves more than physical interactions (e.g., collision avoidance); it also depends on *what the robot believes* about other agents and how it updates those beliefs online as new information is gathered during interaction. This coupling between inference and decision-making

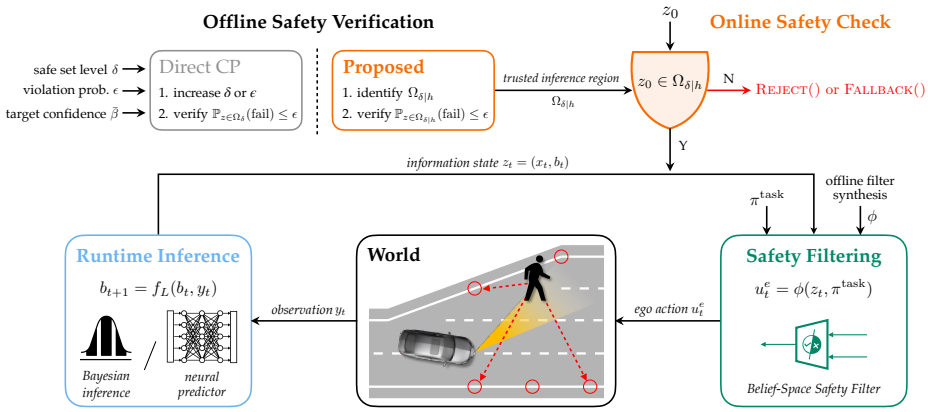

Fig. 1: Offline safety verification and online deployment of belief-space safety filters. Compared to direct conformal prediction, our proposed verification scheme jointly reasons about safety and inference, yielding a larger verified safe set and, consequently, a more permissive safety filter.

is central to interactive robotics, yet it poses a fundamental challenge for safety: even when a control policy is nominally safe under perfect information, incorrect or delayed inference can lead to catastrophic failures. On the other hand, if the robot's safety mechanism ignores its ability to learn from future observations and reduce uncertainty over the course of an interaction, it can become overly conservative and unnecessarily restrict performance [2,21,13].

The recently proposed belief-space safety filters (BELIEFSF) [22] offer a principled way to account for the evolving uncertainty by closing the loop between the robot's safety analysis and runtime inference that updates its belief about other agents. In particular, BELIEFSF assesses safety in an augmented state space that describes both the agents' physical behaviors and the ego robot's internal belief about other agents, thereby explicitly capturing the robot's ability to handle safety-critical events as a function of how accurately it can infer other agents' behavior through observations made during interactions (Figure 1). While BELIEFSF can guarantee safety under assumptions on accurate inference and safety-filter synthesis, how can we certify closed-loop safety in realistic, high-dimensional interaction settings where neither inference nor filter validity is perfect? Recent work [27] uses conformal prediction (CP) to verify general (non-belief) safety filters learned as neural networks and provide probabilistic safety guarantees based on trajectory rollouts. However, directly applying CP to verify belief-space safety filters can conflate safety failures caused by inaccurate inference with those caused by the safety filter itself (e.g., neural approximation errors), needlessly hurting the filter's *permissiveness*. For example, when the robot's runtime inference is generally accurate but occasionally fails in specific regions of the state space, standard CP can certify a desired safety level only by shrinking the safe set *uniformly*, thus restricting the robot's operation even in regions where safety could in fact be ensured. This observation suggests that safety verification for belief-space safety filters should explicitly account for the quality of the robot's runtime inference.

Contributions. We propose a novel inference-aware verification framework for providing high-probability safety guarantees in interactive robot autonomy. Our key insight is that, rather than directly verifying the filtered system end-to-end, *focusing the filter's verification and deployment on regions where the inference is expected to be reliable can substantially reduce its conservativeness.* To the best of our knowledge, this paper is the first to establish closed-loop safety guarantees for interactive motion planning in belief spaces without assumptions on perfect inference or filter validity. We show that our approach can yield a significantly larger certified safe set and more permissive filter than a standard CP baseline through extensive simulation studies of an 18-dimensional interactive driving scenario with a strategically deceptive opponent agent.

2 Related Work

Safety Filters. Safety filters monitor a robot's proposed task actions and, when necessary, modify them to avoid upcoming safety failures. Representative classes of safety filters include, for example, model predictive safety filtering, which reasons about safety by rolling out or optimizing system trajectories [41,8,40,33], control barrier functions, which smoothly modifies the proposed action via optimization to slow down the system when it approaches the boundary of the safe set [3,43,37,34], and reachability-based safety filters, which uses a controlled-invariant set [11] and the associated safe control policy obtained via computing or approximating the solution to a Hamilton–Jacobi reachability problem [31,6,16,19,42,26]. Beyond motion planning, recent work [9,35] shows that safety filters can help reinforcement learning (RL) algorithms converge to an optimal constrained policy that is both safe and performant. While effective at enforcing safety, conventional safety filtering methods tend to be overly conservative in interactive scenarios, since they reason only in the physical space and do not account for the robot's ability to reduce uncertainty about other agents. In particular, safety filtering in this case is *open-loop* in the sense that the robot treats its uncertainty over other agents as static and assumes that it will not acquire new information to better understand other agents' behavior. Alternatively, by proactively reasoning about the *future information* they might receive, robots can make more informed decisions and become less conservative. Motivated by this idea, Bajcsy et al. [5] formulate a reachability problem in the joint space of physical states and estimated parameters of a human model to analyze how quickly the robot can learn such parameters in order to make safe, timely decisions. Belief-space safety filters (BELIEFSF) [22] extends this idea to fully coupled multi-agent interaction scenarios where the robot is not merely an observer, but has the ability to actively engage with other agents to gain new information, leading to a *closed-loop-information* safety analysis. In order to leverage richer and more complex observations, e.g., RGB images, recent works [32,38,29] develop latent-space safety filters that reason about safety with a learned world model or reduced order model.

We develop, for the first time, a safety verification framework for neural approximations of belief-space safety filters. Our analysis also suggests a path toward provable guarantees for emerging safety filters that operate in learned latent spaces.

Safety Verification and Conformal Prediction. While it is generally challenging to synthesize learning-based control policies that are provably safe by design, safety ver-

ification has emerged as a practical way to certify the safety of neural controllers prior to their deployment. Existing safety verification frameworks for autonomous systems predominantly focus on settings with known dynamic models using various control-theoretic tools, e.g., reachable tubes [20,15], barrier certificates [36], hybrid system analysis [24], and mixed-integer optimization [1]. Conformal prediction (CP) [4] has become a popular tool for providing safety guarantees for complex, learning-enabled dynamical systems since it is distribution-free and does not require knowledge of the underlying system dynamics. Lindemann et al. [28] use CP to calibrate prediction regions of other agents' future trajectories, which are incorporated in a downstream model predictive controller for safe planning. Dixit et al. [14] extends this approach to account for distribution shift using adaptive CP. More recently, advanced CP variants [30,10] have been proposed to explicitly handle interactions in multi-agent scenarios. Most closely related to our work are CP algorithms for certifying neural HJ-based safe sets and controllers. Most notably, Lin and Bansal [27] leverage split CP to provide high-probability safety guarantees for systems using neural HJ safe controllers. They also show that such a CP formulation reduces to a scenario optimization problem.

While Lin and Bansal's algorithm focuses on single-system safety in the physical space, our work builds on it to certify neural safety filters that operate in the joint physical–belief space for interactive robotics under uncertainty. We show that focusing verification on regions where inference is expected to be reliable can yield a substantially more permissive belief-space safety filter than an inference-agnostic CP baseline.

3 Problem Setup and Background

Problem Setup. Consider an "ego" autonomous robot (e) interacting with one or multiple "opponent" agents (o), described by a dynamical system

$$x_{t+1} = f(x_t, u_t^e, u_t^o), \tag{1}$$

where $x_t = (x_t^e, x_t^o) \in \mathcal{X} \subseteq \mathbb{R}^{n_x}$ is the joint state vector of the ego and the opponent, $u_t^e \in \mathcal{U}^e \subset \mathbb{R}^{m_e}$ and $u_t^o \in \mathcal{U}^o \subset \mathbb{R}^{m_o}$ are the control vectors of the robot and the opponent, respectively, and $f : \mathcal{X} \times \mathcal{U}^e \times \mathcal{U}^o \to \mathcal{X}$ describes the physical interactions. We capture catastrophic safety events (e.g., collisions) with the *failure set*:

$$\mathcal{F} \coloneqq \{x \in \mathcal{X} \mid g(x) < 0\}, \tag{2}$$

where $g : \mathcal{X} \to \mathbb{R}$ is a safety *margin* function quantifying how close state x is to violating safety.

In this paper, we are interested in verifying that a learned *safe set* Ω (assuming $\Omega \cap \mathcal{F} = \emptyset$) and a corresponding control policy can guarantee safe interactions between the ego robot and the opponent with high probability within a finite horizon T, if the system starts within Ω:

$$\mathbb{P}_{x_0 \in \Omega}(x_t \in \mathcal{F}) \leq \epsilon, \ \forall t = 0, 1, \ldots, T. \tag{3}$$

Remark 1. Alternatively, one may also verify a stronger argument than (3) that safe set Ω is controlled-invariant (i.e., the state x does not leave Ω under a specific control policy): $\mathbb{P}_{x_0 \in \Omega}(x_t \in \Omega^c) \leq \epsilon, \ \forall t = 0, 1, \ldots, T$.

Safety Filters, Maximal Safe Set, and Hamilton–Jacobi Reachability. In order to achieve safety objective (3) while minimally interfering with the ego robot's task performance, we consider *safety filter*, an automatic process that monitors the operation of the ego robot at runtime and intervenes, when deemed necessary, by modifying its task policy $\pi^{\text{task}}(x_t)$ with a filtered action $u_t = \phi(x_t, \pi^{\text{task}})$ to prevent a potential safety failure. Here, the function form of ϕ depends on the intervention type of a safety filter, which includes, e.g., switching, transition, and optimization; our method is designed to work for generic value-based safety filters [17, Sec. 3.2] and therefore agnostic to the specific type of ϕ as long as a safety value is available for the filter's monitoring.

Ideally, we want the most *permissive* safety filter that works within the *maximal* safe set Ω^*, which contains all states in $\mathcal{X}$ where recursive safety is guaranteed despite the worst-case opponent action. In theory, such Ω^* and the corresponding safe controller can be computed with game-theoretic Hamilton–Jacobi (HJ) reachability analysis [31,18,42]. HJ analysis captures the opponent's worst-case behavior with an infinite-horizon zero-sum dynamic game. In discrete time, the game's solution can be obtained via the fixed-point Isaacs equation [23] (the game-theoretic counterpart to the Bellman equation):

$$V(x) = \min \left\{ g(x), \max_{u^e \in \mathcal{U}^e} \min_{u^o \in \mathcal{U}^o} V\big(f(x, u^e, u^o)\big) \right\}. \tag{4}$$

The *safety value function* $V(\cdot)$ encodes the maximal safe set with its zero superlevel set, i.e., $\Omega^* := \{x \mid V(x) \geq 0\}$. Given a value function $V(\cdot)$, the optimal ego's defense and opponent's attack policies $(\pi^{\bullet}, \pi^{o,*})$ are obtained by taking argmax/argmin controls in (4). Existing work has proposed several ways of using HJ policies in safety filter ϕ, including switching [31], rollouts [33], and smooth blending [12,34], each leading to a different intervention type and practical tradeoff between safe set volume Ω and the robot's task performance.

Despite its theoretical soundness, traditional HJ analysis relies on grid-based dynamic programming, making it intractable to scale to high-dimensional state spaces. Recent work has used data-driven methods to scale up the computation of HJ-based safety filter synthesis by approximating value function V or safe control policy $\pi^{\bullet}$ with neural networks learned through, for example, deep reinforcement learning [16,18,42,26] and self-supervised learning [7]. Although such *neural safety filters* have shown practical scalability and robustness in safe control for complex robotic systems, they do not come with any safety guarantees on the closed-loop system. Recently, Lin and Bansal [27] have developed a verification algorithm to certify high-probability safety guarantees (3) for neural HJ-based safety filters using conformal prediction [4], which we briefly introduce in the next section.

Verifying Neural Safety Filters with Conformal Prediction. For a *single-agent* system e (i.e., in the absence of opponent o) equipped with a neural safety value function V and control policy $\pi^{\bullet}$, Lin and Bansal [27] seek to verify that recursive safety holds with high probability for $\Omega_\delta := \{x \mid V(x) \geq \delta\}$, the δ-superlevel set of V, within a time horizon $[0, T]$. Given a robot trajectory $\mathbf{x}_{x_0} = (x_0, x_1, \ldots, x_T)$ resulted from applying policy π, define its *empirical safety score* as:

$$J_\pi(\mathbf{x}_{x_0}) = \min_{t \in [0,T]} g(x_t), \tag{5}$$

which is non-negative if and only if the ego robot always stays outside failure set $\mathcal{F}$ for all $t \in [0, T]$. The verification proceeds as follows. First, N independent and identically distributed (i.i.d.) initial states $x_0^{1:N}$ are sampled within Ω_δ according to some probability distribution $\mathbb{P}$ over Ω_δ. Since Ω_δ typically does not have a closed form, rejection sampling is commonly used in this step. Next, safety scores $J_{\tilde{\pi}^{\boldsymbol{\Phi}}}(\mathbf{x}_{x_i})$ are computed for $i = 1, 2, \ldots, N$ by forward simulating the system from x_0^i under $\tilde{\pi}^{\boldsymbol{\Phi}}$. Finally, we count k empirically unsafe trajectories, where $J_{\tilde{\pi}^{\boldsymbol{\Phi}}} < 0$. The following proposition provides a probabilistic safety guarantee of the closed-loop system operating within Ω_δ.

Proposition 1 (Lemma 5 [27]). *If a safety coverage (violation) parameter $\epsilon \in (0, 1)$ and a confidence parameter $\beta \in (0, 1)$ satisfy $\sum_{i=0}^{k} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta$, then, with probability at least $1 - \beta$, $\mathbb{P}_{x \in \Omega_\delta}(J_{\tilde{\pi}^{\boldsymbol{\Phi}}}(\mathbf{x}_x) \leq 0) \leq \epsilon$.*

The proof of Proposition 1 is based on standard split CP [4] and can be found in [27]. In this paper, we extend and apply Lin and Bansal's method to the verification of a recently proposed belief-space safety filter [22] that safeguards against a potentially adversarial opponent while adapting to uncertainty during the interaction. We introduce this safety filter in the next section.

BELIEFSF: Reducing Conservativeness by Filtering in Information Space. Standard neural approximations to the two-player safety value function in (4), e.g., the ISAACS framework [18], assume that the opponent may take any action from its control set $\mathcal{U}^o$ at all times. This assumption can lead to overly conservative solutions, such as excessively defensive robot behaviors that prevent efficient task execution, or a learned safe set Ω whose volume is significantly smaller compared to that of Ω^*.

Belief-Space Safety Filter (BELIEFSF) [22] proposes to make the filtering *more permissive* by conditioning the opponent's control bound on the robot's inference (runtime learning) algorithm, which reduces the robot's uncertainty about the opponent at runtime. For instance, in the vehicle–pedestrian interaction scenario described in the running example and visualized in Figure 2, imagine that when the human begins to step onto the highway by crossing the road boundary, the robot will have made the corresponding observation, which shifts its belief towards the hypothesis that the human's true intent is to cross the road. Assuming that the robot's belief about human intent is indeed accurate (we will revisit this assumption shortly), the safety filter can temporarily discard other low-probability hypotheses about the human's intent and subsequently "shrink" the opponent's control bound during the safety analysis, thus reducing the conservativeness of the filter.

Concretely, in BELIEFSF, we model the opponent's *hidden type* [25] with a discrete variable $\theta \in \Theta$, which may encode its individual characteristics that are unknown to the ego, e.g., its intended destination, semantic class, and willingness to cooperate, etc. The ego robot then represents its uncertainty over the unknown type θ through a *belief* state $b(\theta) \in \Delta$. Using an inference (runtime learning) algorithm, e.g., Bayesian inference and neural motion predictors, the robot can update (and hopefully refine) its belief over time as new observations are made. We can combine physical dynamics (1) and this runtime inference process to define a joint dynamical system:

$$z_{t+1} = F(z_t, u_t^e, u_t^o) := \begin{bmatrix} f(x_t, u_t^e, u_t^o) \\ f_L(b_t, y_t) \end{bmatrix}, \tag{6}$$

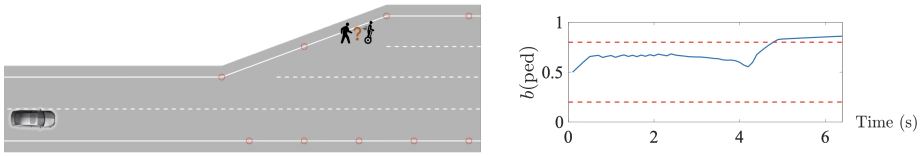

Fig. 2: BELIEFSF applied to vehicle–pedestrian interaction (running example). *Left:* The interaction scene, where an ego autonomous vehicle is uncertain about the semantic class (pedestrian or Segway rider) and intended destination (red dots) of an opponent human crossing the road. *Right:* A representative closed-loop belief trajectory of $b(\theta_{\text{class}} = \text{ped})$, where the dashed red lines represent threshold $\epsilon_{\theta_{\text{class}}} = 0.2$. The adversarial opponent learned to exploit the ego robot by generating motions that keep the robot's belief close to uniform, thereby confusing its inference module and making it challenging to timely identify the opponent's hidden type θ and the corresponding control bound $\mathcal{U}^o_\theta(x_t)$.

where $z_t := (x_t, b_t) \in \mathcal{Z}$ is an *information* (i.e., joint physical–belief) state vector, $y_t := \sigma(x_t, u^o_t) \in \mathcal{Y}$ is the robot's current observation with an observation map σ modeling partial observability, and $f_L : \Delta \times \mathcal{Y} \to \Delta$ is the dynamic model of the robot's runtime inference process. Joint dynamics (6) open up the opportunities of safety analysis and filtering in *information space*. We can now extend the standard Isaacs equation (4) to the information state z:

$$V(z) = \min \left\{ g(x), \max_{u^e \in \mathcal{U}^e} \min_{u^o \in \widehat{\mathcal{U}}^o(z)} V\left(F(z, u^e, u^o)\right) \right\}. \tag{7}$$

Here, $\widehat{\mathcal{U}}^o(z)$ is the ego's *inferred control bound* of the opponent—a key identity that bridges runtime inference and closed-loop safety—defined as

$$\widehat{\mathcal{U}}^o(z_t) := \bigcup_{\theta \in \Theta} \widehat{\mathcal{U}}^o_\theta(z_t), \quad \widehat{\mathcal{U}}^o_\theta(z_t) := \begin{cases} \mathcal{U}^o_\theta(x_t), & \text{if } b_t(\theta) \geq \epsilon_\theta \\ \emptyset, & \text{if } b_t(\theta) < \epsilon_\theta \end{cases} \tag{8}$$

where $\mathcal{U}^o_\theta(x_t)$ is a known (possibly state-dependent) set of *admissible* opponent actions associated with hypothesized type θ, and $\epsilon_\theta \geq 0$ is a designer-specified threshold. Below, we use a running example to illustrate that BELIEFSF improves practical safety and performance compared to neural safety filters synthesized only in physical space.

Running Example (Vehicle–Pedestrian Interaction).

System. We consider an 18D system from [22] describing an uncertain vehicle–pedestrian interaction scenario, as depicted in Figure 2. The autonomous vehicle (ego e) and crossing human (opponent o) are modeled as a 5D kinematic bicycle model and a 2D particle with velocity control, respectively. The information state vector is structured as $z := (x^e, x^o, b(\theta_{\text{class}}), b(\theta_{\text{goal}})) \in \mathbb{R}^{18}$, where $\theta_{\text{class}} \in \{\text{ped}, \text{seg}\}$ models the ego's hypotheses about the opponent's semantic class (pedestrian or Segway rider), with associated control sets $\mathcal{U}^o_{\text{ped}} = [-0.75, 0.75] \times [-2, 2]$ m/s and $\mathcal{U}^o_{\text{seg}} = [-0.75, 0.75] \times [-8, 0]$ m/s, and $\theta_{\text{goal}} \in \{g_i\}_{i=1}^9$ models the hypotheses about the human's goal position,

scattered along the road boundary (red dots in Figure 2), with associated control set $\mathcal{U}^o_{g_i}$ defined as a box centered around a goal-reaching nominal control action for g_i. The robot uses Bayesian inference as the learning dynamics f_L: $b_{t+1}(\theta) \propto \mathbb{P}(u^o_t \mid x_t; \theta) b_t(\theta)$. The likelihood function is a Gaussian $\mathbb{P}(u^o_t \mid x_t; \theta) = \mathcal{N}(\mu_\theta, \Sigma_\theta)$, where μ_θ is the average control of the bound $\mathcal{U}^o_\theta(x_t)$ and Σ_θ is a predefined covariance matrix.

Safety Specifications. A trajectory is considered safe if (1) no collision occurs between the ego vehicle and the human, and (2) the ego vehicle remains within the road boundary. We define the failure set $\mathcal{F}$ using a safety margin function $g(x)$ that encodes both collision avoidance and road-boundary constraints, where $g(x) < 0$ indicates that at least one safety constraint is violated.

BELIEFSF Synthesis and Deployment. We learned a BELIEFSF ϕ using a game-theoretic adversarial RL approach [18,42] that approximately solves a time-discounted variant of Isaacs equation (7). The result includes a neural value function V and safety fallback policy $\pi^{\blacktriangledown}$. The training details can be found in [22]. At runtime, we deploy ϕ as a value-based, switch-type [17, Sec. 3.1] safety filter:

$$u^e = \phi\left(z, \pi^{\text{task}}(z)\right) := \begin{cases} \pi^{\text{task}}(z), & V\Big(F\left(z, \pi^{\text{task}}(z), u^o\right)\Big) \geq \delta, \\ \pi^{\blacktriangledown}(z), & \text{otherwise}, \end{cases}$$

where the safety fallback policy only overrides the task policy to maintain non-negativity (hence safety) of $V(\cdot)$ when state z approaches the boundary of the learned safe set Ω_δ.

Results. We ran 1000 randomized trials with the opponent applying an adversarial policy, also synthesized with game-theoretic RL. The BELIEFSF with $\delta = 0$ achieved an empirical safe rate of 94.3 % and a task completion rate of 99.8 % (a trial is considered timeout if $t > 10$ s). In contrast, the baseline safety filter that reasons only in the physical state space, synthesized with the same game-theoretic RL algorithm, yielded an empirical safe rate of 77.7 % and a task completion rate of only 37.6 %. *These results demonstrate that* BELIEFSF *is significantly safer and more performant* in practice than standard neural HJ-based safety filters operating purely in physical spaces. It is also worth noting that, beyond physical interactions, the opponent in this example learns to exploit the ego agent in the belief space through *deceptive behaviors* that hinder the ego's inference by keeping its belief close to uniform, as evidenced by the belief trajectory in Figure 2. This observation underscores the importance of providing formal safety guarantees for BELIEFSF in scenarios involving strategic, potentially adversarial opponents that may deliberately exploit the ego robot's runtime inference.

Now, one may ask under what conditions a BELIEFSF ϕ can *guarantee* safety for system (1). It has been shown in [22] that closed-loop safety can be ensured under the *Inference Hypothesis* (Definition 1), that is, given an optimal value function V according to (7), if the inferred opponent control bound $\widehat{\mathcal{U}}^o(z_t)$ predicted by f_L can always include the opponent's executed control action u^o_t, then recursive safety is guaranteed when the ego applies BELIEFSF against *any* opponent policy. However, such assumption may be too strong to hold in practice for two reasons: (1) due to the high dimensionality of information spaces[1], the value function V and associated safe policy $\pi^{\blacktriangledown}$ need

[1] The joint system F in the running example is 18D; grid-based HJ analysis cannot practically scale beyond 7D.

to be approximated using, e.g., neural networks, which inevitably introduces error, and (2) the opponent action is likely outside the inferred bound $\widehat{\mathcal{U}}^o(z_t)$ due to, e.g., delayed inference or poorly designed parameters ($\mathcal{U}^o_\theta(x_t)$ and ϵ_θ in (8)), thus invalidating the Inference Hypothesis. In this paper, we develop a practical algorithm to guarantee probabilistic safety for BELIEFSF without relying on the Inference Hypothesis.

Definition 1 (Inference Hypothesis [22]). *At each time t, the ego's belief b_t must have assigned no less than ϵ_θ probability to at least one θ consistent with opponent action u^o_t. This is equivalent to requiring $u^o_t \in \widehat{\mathcal{U}}^o(z_t),\ \forall t \geq 0$, with $\widehat{\mathcal{U}}^o(z_t)$ defined in (8).*

4 Permissive and Assured Belief-Space Safety Filtering

Problem Statement: Permissive Safety Verification of BELIEFSF. In this paper, our main objective is to develop a verification algorithm for BELIEFSF, to certify that its safety guarantees hold with high probability within a learned belief-space safe set. We further use insights gained from verification to modify the deployment of belief-space safety filters. In particular, we show that verifying and deploying BELIEFSF in the *trusted inference region*—a subset of the learned safe set where the robot's inference algorithm is expected to perform reliably—is key to reducing the practical conservativeness of belief-space safety filtering.

4.1 Conformal Probabilistic Safety Verification for BELIEFSF

We start by considering a more direct extension of Lin and Bansal's verification algorithm [27] to the regime of belief-space safety. Define $\mathcal{D}$ as an unknown distribution over the trajectories of information states $\mathbf{z} = (z_0, z_1, \ldots, z_T)$. Next, we make a standard assumption in conformal prediction for safe multi-agent interaction, e.g., MPC-based social navigation [28].

Assumption 1 *The control inputs of the ego robot and opponent (u^e_t, u^o_t) do not change the distribution $\mathcal{D}$ for all $t \in [0, T]$.*

Remark 2. Assumption 1 is a standard assumption for ensuring exchangeability commonly used in conformal prediction for interactive systems [28]. Intuitively, it assumes that applying the safety filter does not alter the underlying distribution of interaction trajectories. When the robot's filtered actions substantially influence the opponent's behavior and alter the interaction distribution, this assumption may no longer hold. In such cases, recent CP methods that explicitly account for interaction-induced distribution shift [30,10] provide a promising direction for extending our framework.

Given an ego robot's task policy $\pi^{\text{task},e}$, opponent's policy π^o, and a BELIEFSF ϕ defined by neural value function V and safe control policy $\pi^{\blacktriangledown}$, our goal is to verify that within the δ-superlevel set $\Omega_\delta := \{z \mid V(z) \geq \delta\}$, system (6) in closed-loop with ϕ is recursively safe, i.e., $x_t \notin \mathcal{F},\ \forall t \in [0, T]$, with high probability. The approach, which we refer to as DIRECTCP, resembles that of Lin and Bansal [27], with the main difference that the sampling is now performed in the information space that

encompasses both the ego and opponent's physical states as well as belief states. We first sample N i.i.d. initial states $z_0^{1:N} \in \Omega_\delta$ and obtain trajectories $\mathbf{z}_{z^{1:N}}$ by forward simulating joint system F with the ego applying $\phi(\cdot, \pi^{\text{task},e})$ and the opponent applying policy π^o. Then, we compute empirical safety scores $J_\phi(\mathbf{z}_{z_i}) = \min_{t\in[0,T]} g(x_t)$ for $i = 1, 2, \ldots, N$ and count k empirically unsafe trajectories where $J_\phi(\cdot) < 0$.

Running Example: We applied DIRECTCP to verify BELIEFSF in the running example, where the opponent applies a fixed adversarial policy synthesized with game-theoretic RL. The verification horizon was $T = 10$ s. We used a confidence parameter $\beta = 10^{-5}$. With a safe set level $\delta = 0.1$, we obtained $k = 438$ failure cases from $N = 20000$ samples, leading to a certified safety level (coverage) $1-\epsilon = 97.34\%$. Among the 20000 trials, inference failed in 475 cases, 338 of which also resulted in unsafe outcomes. The safety level is increased to 98.7% with $\delta = 0.7$. Additional results with varying safe set levels δ can be found in Figure 4.

As reported in the above running example, the empirical safe rate when using BELIEFSF is significantly lower when the robot's inference algorithm has failed, i.e., when it does not account for the opponent's executed action u_t^o in the inferred bound $\widehat{\mathcal{U}}^o(z_t)$. This phenomenon is expected since the synthesis of BELIEFSF via Isaacs equation (7) assumes that the robot's inference is accurate and that the opponent's control always lies within the predicted control bound. Motivated by this observation, in the next section, we propose an improved CP procedure by explicitly accounting for the robot's inference performance in belief-space safety filtering, so that the verification algorithm can return a tighter (i.e., smaller) safety coverage parameter ϵ without needing to shrink the safe set volume with an excessively large δ, thus reducing the conservativeness of safety filtering.

Remark 3. Although the opponent policy used in verification is an approximation of a worst-case adversarial policy, the safety guarantees provided by conformal prediction are *distributional*: they hold with respect to the trajectory distribution induced by this fixed opponent policy, rather than a worst-case guarantee over all admissible opponents. Nevertheless, verifying against a fixed, highly exploiting opponent remains useful in practice. Empirically, we often observe that such opponents tend to induce safety-critical interactions that stress both the safety filter and the runtime inference module [22,42,33]. When the ego policy is paired with a less aggressive opponent, the resulting trajectory is typically no more safety-critical than those seen during verification, so the empirical failure rate under the same ego policy is expected to be no higher (and often lower) than what is observed during verification.

4.2 Verifying and Deploying BELIEFSF in Trusted Inference Regions

Our key insight is that, when the robot's inference algorithm is sufficiently reliable, i.e., it can successfully contain u_t^o within $\widehat{\mathcal{U}}^o(z_t)$ for the majority of state space with high probability, we may choose to identify and "carve out" the small subset in state space where the inference algorithm can fail. This way, we can still expect to recover a large portion of the safe set. To formalize this idea, we start by introducing the *inference score*

$J_L = h_L(z)$, where $h_L : \mathcal{Z} \to \mathbb{R}$ is a user-specified function mapping an information state z to a score J_L that accesses the *quality* of the robot's inference algorithm; the inference algorithm is *believed* to be accurate for physical state x in z if $h_L(z) \geq 0$. In the context of BELIEFSF, a reasonable design of the inference score function can be as simple as a classifier that outputs $h_L(z) = 1$ if it *predicts* for the information state trajectory with $z_0 = z$ that $u_t^o \in \widehat{\mathcal{U}}^o(z_t),\ \forall t \geq 0$, and $h_L(z) = -1$ otherwise; alternatively, h_L can be a function of the predicted probability or confidence score[2], typically available (or easy to calculate) from the output of modern multi-modal trajectory prediction models [39]. Note that the inference score function h_L is purely an evaluation on the quality of the inference algorithm f_L, and should therefore be chosen *independently* of the safety filter ϕ. We may now define the notion of *trusted inference region*, a subset of the learned safe set where we also expect the robot's inference algorithm to be accurate.

Definition 2 (Trusted Inference Region). *Given a score function h_L and learned information-space safe set Ω_δ, the trusted inference region is defined as $\Omega_{\delta|h} := \{z \mid V(z) \geq \delta, h_L(z) \geq 0\}$.*

Running Example: We trained a classifier $h_L(z)$ that outputs 1 if it predicts that the information state trajectory satisfies $u_t^o \in \widehat{\mathcal{U}}^o(z_t),\ \forall t \in [0, T]$, and -1 otherwise. The classifier was trained on a dataset of 10^6 randomly sampled initial information states z_0 with labels obtained by rolling out joint dynamics F. On a test dataset of size 2×10^5, $h_L(z)$ leads to a false inference rate (i.e., rejection rate) of 4.67% with a misclassification rate of 0.92%.

Now, we are ready to introduce Joint Inference–Safety Test (JIST), an inference-aware CP-based safety verification algorithm tailored for BELIEFSF. Our goal is to verify that system trajectories starting within $\Omega_{\delta|h}$ under BELIEFSF ϕ is safe with high probability. Similar to DIRECTCP, we begin by sampling N i.i.d. initial states $z_0^{1:N}$ within the trusted inference region $\Omega_{\delta|h}$. We then forward-simulate F under ϕ and π^o to obtain trajectories $\mathbf{z}_{z^{1:N}}$. Since, ultimately, we care about safety assurances under ϕ, we only need to compute the empirical safety score $J_\phi(\mathbf{z}_{z_i}) = \min_{t\in[0,T]} g(x_t)$ for $i = 1, 2, \ldots, N$ and count k failure cases, the same as in DIRECTCP. The above JIST verification procedure is summarized in Algorithm 1. Typically, we want the verification result to hold with high confidence defined by a target value $\bar{\beta}$ that is sufficiently small. In the event that $\beta > \bar{\beta}$, one may rerun Algorithm 1 by increasing ϵ, δ, or N, thereby achieving the desired confidence at the expense of increased safety conservativeness or sample complexity. Compared to DIRECTCP, JIST does not introduce any additional sample or computation complexity, since it only differs by focusing sampling in the trusted inference region $\Omega_{\delta|h}$. In Section 5, we will see that this minimal change can lead to a significantly reduced conservativeness for belief-space safety filtering in practice. The following theorem provides a probabilistic safety guarantee of the closed-loop system F operating within the trusted inference region $\Omega_{\delta|h}$.

[2] For example, if the model's predicted probability or confidence of belief trajectory $\mathbf{b} = (b_0, b_1, \ldots)$ is too low, h_L can return a negative number to predict poor inference performance.

Theorem 1 (Safety Verification of BELIEFSF with JIST). *Given a trusted inference region $\Omega_{\delta|h}$ defined in Definition 2, if a safety coverage parameter $\epsilon \in (0,1)$ and a confidence parameter $\beta \in (0,1)$ satisfy $\sum_{i=0}^{k} \binom{N}{i} \epsilon^{i}(1-\epsilon)^{N-i} \leq \beta$, then $\mathbb{P}_{z\in\Omega_{\delta|h}}(J_\phi(\mathbf{z}_z) \leq 0) \leq \epsilon$ with probability at least $1-\beta$.*

Proof. The proof follows directly from standard split CP results and resembles that of [27, Theorem 3]. We provide a brief sketch here for completeness. Define the nonconformity score as the negative safety score $-J_\phi(\mathbf{z}_z)$ and set $\alpha = \frac{k+1}{N+1}$. By [4, Theorem 1], we have $\mathbb{P}_{z\in\Omega_{\delta|h}}(J_\phi(\mathbf{z}_z) \geq -\hat{q}) \geq 1-\alpha$, where quantile $\hat{q}$ is the $(N-k)$-th safety score, i.e., the largest negative score of $J_\phi(\cdot)$. Therefore, we have that $\mathbb{P}_{z\in\Omega_{\delta|h}}(J_\phi(\mathbf{z}_z) > 0) \geq \frac{N-k}{N+1}$. In addition, by [4, Sec. 3.2], we can further show that the distribution of safety coverage admits a closed-form Beta distribution, i.e., $\mathbb{P}_{z\in\Omega_{\delta|h}}(J_\phi(\mathbf{z}_z) > 0) \sim \mathrm{Beta}(N-k, k+1)$. Equivalently, we have $\mathbb{P}(p \leq \epsilon) = \mathbb{P}(1-p \geq 1-\epsilon) = 1 - F_{\mathrm{Beta}(N-k,k+1)}(1-\epsilon) = 1 - \sum_{i=0}^{k} \binom{N}{i} \epsilon^{i}(1-\epsilon)^{N-i}$, where $p := \mathbb{P}_{z\in\Omega_{\delta|h}}(J_\phi(\mathbf{z}_z) \leq 0)$ and $F_{\mathrm{Beta}(\cdot,\cdot)}$ denotes the Beta CDF. Hence, if (ϵ, β) satisfy $\sum_{i=0}^{k} \binom{N}{i} \epsilon^{i}(1-\epsilon)^{N-i} \leq \beta$, then $\mathbb{P}(p \leq \epsilon) \geq 1-\beta$, i.e., with probability at least $1-\beta$ we have $\mathbb{P}_{z\in\Omega_{\delta|h}}(J_\phi(\mathbf{z}_z) \leq 0) \leq \epsilon$. □

Once a trusted inference region $\Omega_{\delta|h}$ and BELIEFSF ϕ are verified with Algorithm 1, it is straightforward to deploy them online for interactive motion planning with safety guarantees, as shown in Figure 1. In essence, $\Omega_{\delta|h}$ is used as an additional safety check to safety filter ϕ's own monitoring scheme. It is only at the beginning of the interaction that we check if the initial information state z_0 is within $\Omega_{\delta|h}$. If so, we may apply belief-space safety filtering in the standard way as in [22], and the resulting trajectory is provably safe according to the verification results. Otherwise, the system raises a flag and executes either REJECT() or FALLBACK(). The REJECT() option prevents the system from applying policies filtered under ϕ, since they may not be safe; for example, in assisted driving, this amounts to handing control back to the human driver. The FALLBACK() option attempts to replace ϕ with a backup filter, if available, that is more conservative but certified safe at the given state z_0. The procedure of deploying a verified BELIEFSF is summarized in Algorithm 2. Note that Algorithm 2 may also be used in a receding horizon fashion by recursively checking if $z_t \in \Omega_{\delta|h}$, which is useful when the actual control horizon is longer than the verification horizon T.

Remark 4. Since JIST couples safety verification with the quality of inference, it is expected to outperform DIRECTCP in settings where the cause of safety violations is dominated by incorrect inference, that is, the robot's runtime inference and neural safety filter are sufficiently accurate so that the "correct-inference" regime yields a substantially higher empirical safe rate. As both runtime inference (e.g., trajectory prediction models) and neural safety filter synthesis methods continue to improve in accuracy, JIST can directly benefit from these advances by expanding the certified safe set without requiring additional verification data.

Remark 5. Independent of the safety filter, we can optionally run a conformal prediction procedure to quantify the reliability of the inference module. This evaluation can help inform the choice between JIST and DIRECTCP by checking whether inference errors are sufficiently rare to warrant coupling safety with inference during verification.

Algorithm 1: Offline CP-based Safety Verification with JIST

Input: Violation probability $\epsilon > 0$, safe set level $\delta \geq 0$, inference score function h_L, learned belief-space value function V and safety filter ϕ, ego's task policy $\pi^{\text{task},e}$, opponent policy π^o, sample size $N > 0$, horizon $T > 0$

Output: Verified trusted inference region $\Omega_{\delta|h}$ with confidence parameter β

$k \leftarrow 0$
for $i = 0, 1, \ldots, N-1$ **do**
  Sample information state z_0^i from $\Omega_{\delta|h}$
  Obtain trajectory $\mathbf{z}_{z^i} = (z_0^i, z_1^i, \ldots, z_T^i)$ by forward simulating joint system F with ϕ, $\pi^{\text{task},e}$, and π^o
  Compute empirical safety score $J_\phi(z_i) = \min_{t\in[0,T]} g(x_t)$
  if $J_\phi(z_i) < 0$ **then**
    $k \leftarrow k + 1$
Compute a confidence parameter β such that $\sum_{i=0}^{k} \binom{N}{i} \epsilon^{\,i}(1-\epsilon)^{\,N-i} \leq \beta$

Algorithm 2: Safe Interactive Planning with Verified BELIEFSF

Input: Trusted inference region $\Omega_{\delta|h}$, initial information state z_0, learned belief-space safety filter ϕ, ego's task policy $\pi^{\text{task},e}$, horizon $T > 0$

if $z_0 \in \Omega_{\delta|h}$ **then**
  for $t = 0, 1, \ldots, T$ **do**
    Execute filtered robot's control $u_t^e = \phi(z_t, \pi^{\text{task},e})$
    Observe new state x_{t+1}
    Update belief state: $b_{t+1} \leftarrow f_L(b_t, y_t)$
else
  REJECT() or FALLBACK()

5 Case Study

In this section, we report verification and empirical results in the running example when the ego robot applies JIST. For brevity, we refer readers to previous running example sections for details on the system setup, BELIEFSF synthesis and deployment, opponent modeling, and training of the inference score function $h_L(\cdot)$. Our goal is to evaluate whether coupling conformal verification with inference quality (JIST) can certify a more permissive belief-space safety filter than DIRECTCP while maintaining the same probabilistic safety guarantees. Concretely, our evaluation focuses on two questions:

- for a fixed safe set level δ and calibration sample size N, whether JIST yields a tighter safety coverage parameter ϵ;
- for a fixed safety coverage ϵ and calibration sample size N, whether JIST certifies a more permissive filter, measured by rejection rate during online deployment.

For both methods, conformal prediction was performed using the same number of rollout samples ($N = 20000$) and confidence parameter ($\beta = 10^{-5}$). The inference score function $h_L(\cdot)$ used by JIST was trained separately from, and independently of,

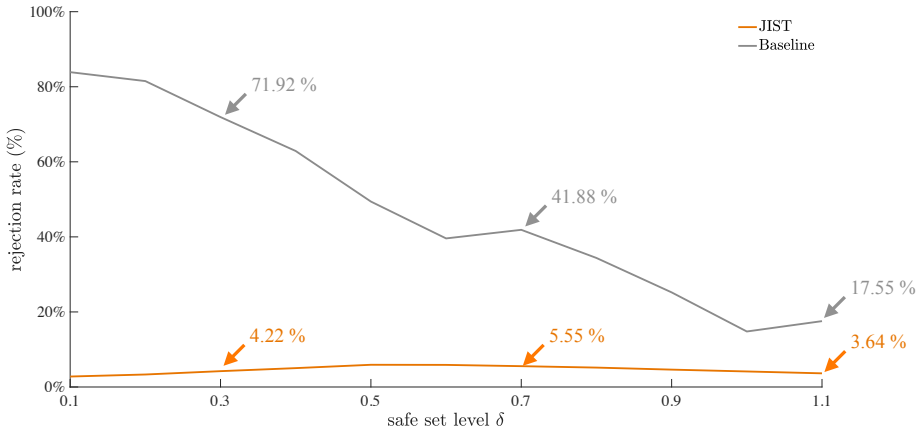

Fig. 3: Rejection rates of BELIEFSF computed from 20000 randomized trials. For a fixed base safety level δ and safety coverage ϵ, JIST achieves a lower rejection rate than DIRECTCP.

the conformal prediction procedure, as described in the running example. Importantly, these additional samples are only used to *predict* the inference quality, and are not used by the conformal verification procedure itself; moreover, many contemporary inference modules, e.g., Transformer-based predictors, naturally provide confidence-related signals that can be used to construct such inference-quality scores. With safe set level $\delta = 0.1$, we obtained $k = 107$ failure cases from $N = 20000$ trials with initial states sampled in the trusted inference region $\Omega_{\delta|h}$, leading to a certified safety level (coverage) $1 - \epsilon = 99.21\%$, which is 1.87% higher than that of DIRECTCP. Among the 107 failed trials, inference was incorrect in 101 cases (94.4%). We plot empirical and verified safety coverage of JIST and DIRECTCP with varying safe set levels δ in Figure 4. We observe that JIST consistently outperforms DIRECTCP by 1–2% in safety coverage for low to moderate values of the safe set level δ.

Next, we compare rejection rates of JIST and DIRECTCP under the same safety coverage when the verified BELIEFSF is deployed for online safe control in information space (Figure 1). Given a fixed safe set level δ, recall that BELIEFSF with JIST rejects a trial when the information state is outside of the learned safe set ($V(z_0) < \delta$) or the predicted inference accuracy is too low ($h_L(z_0) < 0$). In case of DIRECTCP, a trial is rejected if $V(z_0) < \delta_{\text{baseline}}$, where $\delta_{\text{baseline}} \geq \delta$ is a larger, and hence more restrictive safe set level that achieves the same safety coverage ($1 - \epsilon$) as that of JIST. We plot rejection rates with varying safe set levels δ in Figure 3. With JIST, the rejection rate is as low as the false inference rate, since the learned BELIEFSF is generally capable of enforcing safety when the inference algorithm correctly predicts the opponent's control bound. On the other hand, BELIEFSF verified with DIRECTCP requires a substantially larger safe set level δ_{baseline} to achieve the same safety coverage, ultimately leading to a much higher rejection rate. Finally, we visualize the planar ($x - y$ positions) slices of level sets with δ_{JIST} and δ_{baseline} that achieve the same safety coverage ϵ in Figure 5.

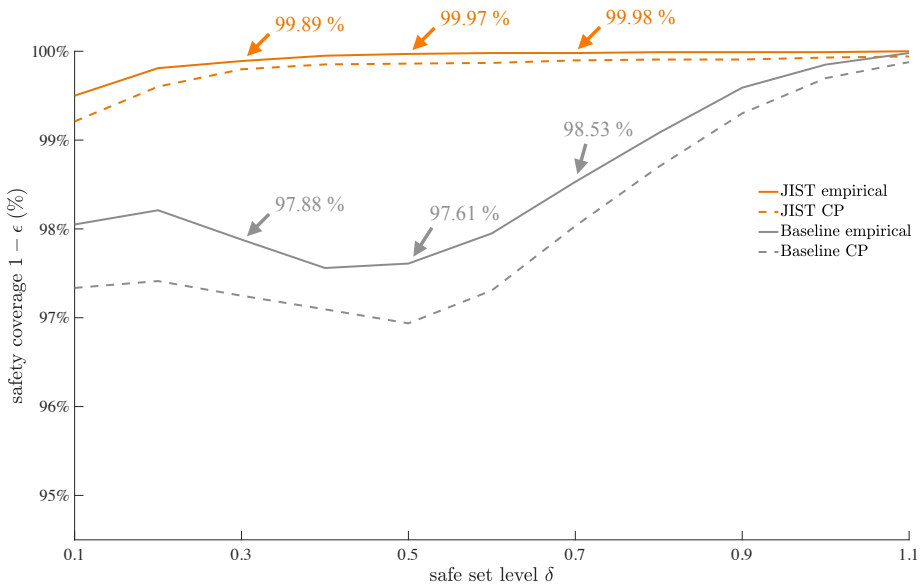

Fig. 4: Empirical and certified safe rates of BELIEFSF computed from 20000 randomized trials. For a fixed safety level δ, JIST yields a tighter safety coverage than the DIRECTCP baseline.

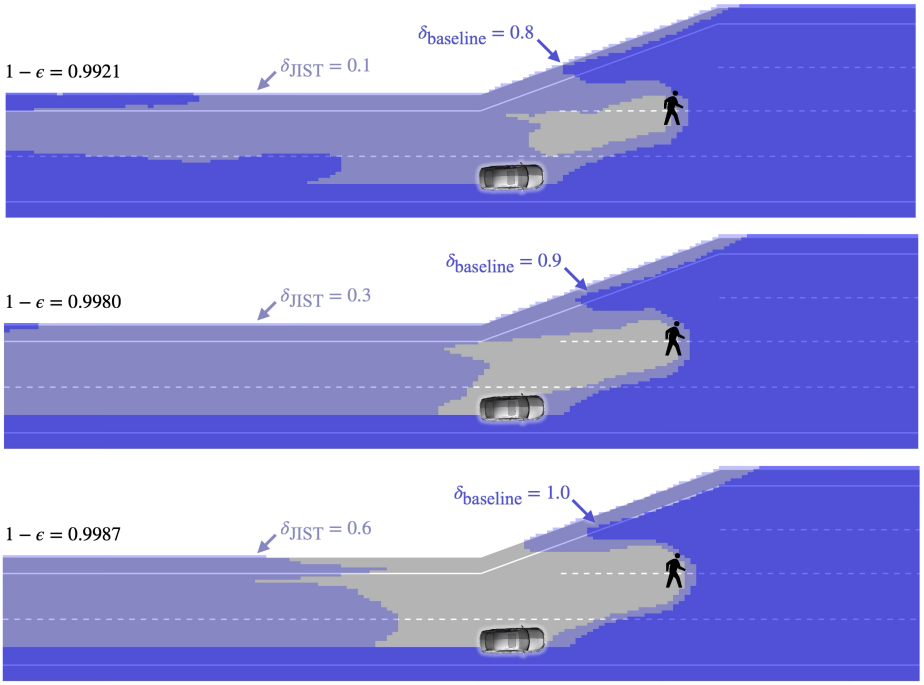

Fig. 5: Planar slices of level sets with δ_{JIST} and δ_{baseline} under safety coverage ϵ.

6 Limitations and Future Work

Our approach improves the filter's permissiveness by focusing verification and deployment on a region where inference is expected to be reliable. This benefit hinges on having access to a reasonable inference score function $h_L(\cdot)$ and on the inference algorithm being sufficiently accurate. When these conditions do not hold, a practical mitigation is a layered safety architecture in which the robot falls back to a conservative, non-belief safety filter whenever inference quality is low.

In line with standard conformal prediction, our safety guarantees are distributional rather than worst-case. In particular, our analysis relies on standard exchangeability and i.i.d. assumptions as well as the independence of distribution from agent actions (Assumption 1). When agent interactions induce a nontrivial distribution shift, these guarantees may no longer hold. Recent advances in adaptive CP that explicitly account for interaction-induced distribution shift offer a promising direction for strengthening our framework [30,10].

Another promising direction worth exploring is to close the loop between verification and filter synthesis by designing an iterative regime that uses CP outcomes to improve both the safety filter and the inference module. Concretely, the CP-certified safe set can guide the tuning of BELIEFSF parameters, e.g., reducing ϵ_θ in (8) in regions where inference is accurate and uncertainty is low. This suggests a verification-aware filter synthesis pipeline in which CP is not merely a post-hoc certificate, but also provides supervision signals that actively inform filter design and data collection. We are also eager to validate our proposed framework on BELIEFSF stacks with more complex inference modules, e.g., LSTM or Transformer-based predictors [22, Sec. 4.3].

7 Conclusions

In this paper, we studied the problem of verifying closed-loop safety for belief-space safety filters in interactive planning settings where both runtime inference and learned safety filters can be imperfect. Our key idea is to couple verification with inference quality: rather than certifying the filtered system end-to-end, we verify and deploy the filter within a region where the robot's inference is expected to be reliable. This approach preserves the simplicity of conformal prediction while avoiding unnecessary conservativeness that arises when rare inference failures are conflated with errors of the safety filter itself. In a simulated vehicle–pedestrian interaction benchmark, our proposed verification algorithm certified a substantially more permissive filter than a standard conformal prediction baseline, achieving higher safety coverage at the same base safe-set level and lower rejection rate when deployed online.

References

1. Ahn, H., Del Vecchio, D.: Safety verification and control for collision avoidance at road intersections. IEEE Transactions on Automatic Control **63**(3), 630–642 (2017)
2. Althoff, M., Dolan, J.M.: Online verification of automated road vehicles using reachability analysis. IEEE Transactions on Robotics **30**(4), 903–918 (2014). https://doi.org/10.1109/TRO.2014.2312453

3. Ames, A.D., Xu, X., Grizzle, J.W., Tabuada, P.: Control barrier function based quadratic programs for safety critical systems. IEEE Transactions on Automatic Control **62**(8), 3861–3876 (2017). https://doi.org/10.1109/TAC.2016.2638961
4. Angelopoulos, A.N., Bates, S., et al.: Conformal prediction: A gentle introduction. Foundations and Trends® in Machine Learning **16**(4), 494–591 (2023)
5. Bajcsy, A., Siththaranjan, A., Tomlin, C.J., Dragan, A.D.: Analyzing human models that adapt online. In: Proc. IEEE Conf. Robotics and Automation. pp. 2754–2760. IEEE (2021). https://doi.org/10.1109/ICRA48506.2021.9561652
6. Bansal, S., Chen, M., Herbert, S., Tomlin, C.J.: Hamilton-jacobi reachability: A brief overview and recent advances. In: Proc. IEEE Conf. Decision and Control. pp. 2242–2253 (2017). https://doi.org/10.1109/CDC.2017.8263977
7. Bansal, S., Tomlin, C.J.: Deepreach: A deep learning approach to high-dimensional reachability. In: Proc. IEEE Conf. Robotics and Automation. pp. 1817–1824 (2021). https://doi.org/10.1109/ICRA48506.2021.9561949
8. Bastani, O., Li, S.: Safe reinforcement learning via statistical model predictive shielding. In: Proc. Robotics: Science and Systems (7 2021). https://doi.org/10.15607/RSS.2021.XVII.026
9. Bejarano, F.P., Brunke, L., Schoellig, A.P.: Safety filtering while training: Improving the performance and sample efficiency of reinforcement learning agents. IEEE Robotics and Automation Letters (2024)
10. Binny, A.E., Dixit, A.: Who moved my distribution? conformal prediction for interactive multi-agent systems. arXiv preprint arXiv:2511.11567 (2025)
11. Blanchini, F.: Set invariance in control. Automatica **35**(11), 1747–1767 (1999). https://doi.org/https://doi.org/10.1016/S0005-1098(99)00113-2
12. Borquez, J., Chakraborty, K., Wang, H., Bansal, S.: On safety and liveness filtering using hamilton-jacobi reachability analysis. IEEE Transactions on Robotics (2024). https://doi.org/10.1109/TRO.2024.3454470
13. Borquez, J., Nakamura, K., Bansal, S.: Parameter-conditioned reachable sets for updating safety assurances online. In: Proc. IEEE Conf. Robotics and Automation. pp. 10553–10559 (2023). https://doi.org/10.1109/ICRA48891.2023.10160554
14. Dixit, A., Lindemann, L., Wei, S.X., Cleaveland, M., Pappas, G.J., Burdick, J.W.: Adaptive conformal prediction for motion planning among dynamic agents. In: Learning for Dynamics & Control. pp. 300–314. PMLR (2023), https://proceedings.mlr.press/v211/dixit23a.html
15. Everett, M., Habibi, G., Sun, C., How, J.P.: Reachability analysis of neural feedback loops. IEEE Access **9**, 163938–163953 (2021). https://doi.org/10.1109/ACCESS.2021.3133370
16. Fisac, J.F., Lugovoy, N.F., Rubies-Royo, V., Ghosh, S., Tomlin, C.J.: Bridging hamilton-jacobi safety analysis and reinforcement learning. In: Proc. IEEE Conf. Robotics and Automation. pp. 8550–8556 (2019). https://doi.org/10.1109/ICRA.2019.8794107
17. Hsu, K.C., Hu, H., Fisac, J.F.: The safety filter: A unified view of safety-critical control in autonomous systems. Annual Review of Control, Robotics, and Autonomous Systems **7**(1) (2024). https://doi.org/10.1146/annurev-control-071723-102940
18. Hsu, K.C., Nguyen, D.P., Fisac, J.F.: Isaacs: Iterative soft adversarial actor-critic for safety. In: Learning for Dynamics & Control. Proceedings of Machine Learning Research, vol. 211 (6 2023), https://proceedings.mlr.press/v211/hsu23a.html
19. Hsu, K.C., Rubies-Royo, V., Tomlin, C.J., Fisac, J.F.: Safety and liveness guarantees through reach-avoid reinforcement learning. In: Proc. Robotics: Science and Systems (7 2021). https://doi.org/10.15607/RSS.2021.XVII.077
20. Hu, H., Fazlyab, M., Morari, M., Pappas, G.J.: Reach-sdp: Reachability analysis of closed-loop systems with neural network controllers via semidefinite programming. In: Proc. IEEE Conf. Decision and Control. pp. 5929–5934 (2020). https://doi.org/10.1109/CDC42340.2020.9304296

21. Hu, H., Nakamura, K., Fisac, J.F.: Sharp: Shielding-aware robust planning for safe and efficient human-robot interaction. IEEE Robotics and Automation Letters **7**(2), 5591–5598 (2022). https://doi.org/10.1109/LRA.2022.3155229
22. Hu, H., Zhang, Z., Nakamura, K., Bajcsy, A., Fisac, J.F.: Deception game: Closing the safety-learning loop in interactive robot autonomy. In: Conf. Robot Learning. Proceedings of Machine Learning Research, vol. 229, pp. 3830–3850 (11 2023), https://proceedings.mlr.press/v229/hu23b.html
23. Isaacs, R.: Differential Games I: Introduction. Differential Games I (1954), https://www.rand.org/pubs/research_memoranda/RM1391.html
24. Ivanov, R., Weimer, J., Alur, R., Pappas, G.J., Lee, I.: Verisig: verifying safety properties of hybrid systems with neural network controllers. In: Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control. pp. 169–178 (2019)
25. Kajii, A., Morris, S.: The robustness of equilibria to incomplete information. Econometrica: Journal of the Econometric Society pp. 1283–1309 (1997). https://doi.org/10.2307/2171737
26. Li, J., Lee, D., Lee, J., Dong, K.S., Sojoudi, S., Tomlin, C.: Certifiable reachability learning using a new lipschitz continuous value function. IEEE Robotics and Automation Letters (2025)
27. Lin, A., Bansal, S.: Verification of neural reachable tubes via scenario optimization and conformal prediction. In: Learning for Dynamics & Control. pp. 719–731. PMLR (2024), https://proceedings.mlr.press/v242/lin24a.html
28. Lindemann, L., Cleaveland, M., Shim, G., Pappas, G.J.: Safe planning in dynamic environments using conformal prediction. IEEE Robotics and Automation Letters **8**(8), 5116–5123 (2023). https://doi.org/10.1109/LRA.2023.3292071
29. Lutkus, P., Wang, K., Lindemann, L., Tu, S.: Latent representations for control design with provable stability and safety guarantees. arXiv preprint arXiv:2505.23210 (2025). https://doi.org/10.48550/arXiv.2505.23210
30. Mirzaeedodangeh, O., Shekhtman, E., Matni, N., Lindemann, L.: Safe planning in interactive environments via iterative policy updates and adversarially robust conformal prediction. arXiv preprint arXiv:2511.10586 (2025)
31. Mitchell, I.M., Bayen, A.M., Tomlin, C.J.: A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games. IEEE Transactions on Automatic Control **50**(7), 947–957 (2005). https://doi.org/10.1109/TAC.2005.851439
32. Nakamura, K., Peters, L., Bajcsy, A.: Generalizing safety beyond collision-avoidance via latent-space reachability analysis (2025)
33. Nguyen, D.P., Hsu, K.C., Yu, W., Tan, J., Fisac, J.F.: Gameplay filters: Robust zero-shot safety through adversarial imagination. In: Conf. Robot Learning. pp. 387–407. PMLR (2025), https://proceedings.mlr.press/v270/nguyen25a.html
34. Oh, D.D., Lidard, J., Hu, H., Sinhmar, H., Lazarski, E., Gopinath, D., Sumner, E.S., DeCastro, J.A., Rosman, G., Leonard, N.E., et al.: Safety with Agency: Human-Centered Safety Filter with Application to AI-Assisted Motorsports. Proc. Robotics: Science and Systems (2025). https://doi.org/10.48550/arXiv.2504.11717
35. Oh, D.D., Nguyen, D.P., Hu, H., Fisac, J.F.: Provably optimal reinforcement learning under safety filtering. arXiv preprint arXiv:2510.18082 (2025)
36. Prajna, S., Jadbabaie, A.: Safety verification of hybrid systems using barrier certificates. In: International workshop on hybrid systems: Computation and control. pp. 477–492. Springer (2004)
37. Robey, A., Hu, H., Lindemann, L., Zhang, H., Dimarogonas, D.V., Tu, S., Matni, N.: Learning control barrier functions from expert demonstrations. In: Proc. IEEE Conf. Decision and Control. pp. 3717–3724 (2020). https://doi.org/10.1109/CDC42340.2020.9303785
38. Seo, J., Nakamura, K., Bajcsy, A.: Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. arXiv preprint arXiv:2505.00779 (2025)

39. Shi, S., Jiang, L., Dai, D., Schiele, B.: Motion transformer with global intention localization and local movement refinement. Advances in Neural Information Processing Systems **35**, 6531–6543 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/2ab47c960bfee4f86dfc362f26ad066a-Abstract-Conference.html
40. Strawn, K.J., Ayanian, N., Lindemann, L.: Conformal predictive safety filter for rl controllers in dynamic environments. IEEE Robotics and Automation Letters **8**(11), 7833–7840 (2023)
41. Wabersich, K.P., Zeilinger, M.N.: Linear model predictive safety certification for learning-based control. In: Proc. IEEE Conf. Decision and Control. pp. 7130–7135 (2018). https://doi.org/10.1109/CDC.2018.8619829
42. Wang, J., Hu, H., Nguyen, D.P., Fisac, J.F.: MAGICS: Adversarial RL with Minimax Actors Guided by Implicit Critic Stackelberg for Convergent Neural Synthesis of Robot Safety. In: Algorithmic Foundations of Robotics (2024), https://www.algorithmic-robotics.org/papers/45_MAGICS_Adversarial_RL_with_.pdf
43. Xiao, W., Belta, C.: Control barrier functions for systems with high relative degree. In: Proc. IEEE Conf. Decision and Control. pp. 474–479 (2019). https://doi.org/10.1109/CDC40024.2019.9029455