IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Safety Evaluation of Motion Plans Using Trajectory Predictors as Forward Reachable Set Estimators

Kaustav Chakraborty*[1,2,3], Zeyuan Feng*[2], Sushant Veer*[3], Apoorva Sharma[3], Wenhao Ding[3]
Sever Topan[3], Boris Ivanovic[3], Marco Pavone[2,3], Somil Bansal[2]

***Abstract*—The advent of end-to-end autonomy stacks—often lacking interpretable intermediate modules—has placed an increased burden on ensuring that the final output, i.e., the motion plan, is safe in order to validate the safety of the entire stack. This requires a safety monitor that is both complete (able to detect all unsafe plans) and sound (does not flag safe plans). In this work, we propose a principled safety monitor that leverages modern multi-modal trajectory predictors to approximate forward reachable sets (FRS) of surrounding agents. By formulating a convex program, we efficiently extract these data-driven FRSs directly from the predicted state distributions, conditioned on scene context such as lane topology and agent history. To ensure completeness, we leverage conformal prediction to calibrate the FRS and guarantee coverage of ground-truth trajectories with high probability. To preserve soundness in out-of-distribution (OOD) scenarios or under predictor failure, we introduce a Bayesian filter that dynamically adjusts the FRS conservativeness based on the predictor's observed performance. We then assess the safety of the ego vehicle's motion plan by checking for intersections with these calibrated FRSs, ensuring the plan remains collision-free under plausible future behaviors of others. Extensive experiments on the nuScenes dataset show our approach significantly improves soundness while maintaining completeness, offering a practical and reliable safety monitor for learned autonomy stacks.**

Project Website: vatsuak.github.io/forceopt/

***Index Terms*—Autonomous Vehicle Navigation, Robot Safety, Planning Under Uncertainty.**

I. INTRODUCTION

CLASSICAL planning stacks come in various shapes and forms [1], [2], but, importantly, they share the characteristic of optimizing an interpretable cost function which allows reasoning about the plan's safety during synthesis. With the ever-increasing adoption of learning-based planners and end-to-end robot stacks, safe-by-construction planning has become extremely challenging, if not outright impossible. Therefore, safety monitors for motion plans have grown in prominence for ensuring that these often uninterpretable learning-based plans are safe. Like any effective monitor, a safety monitor should satisfy two key properties: completeness (it must flag all unsafe plans) and soundness (it must not flag safe ones). Our objective in this paper is to develop a method that can improve on the soundness of reachability-based safety monitors without compromising on completeness.

*equal contribution. Corresponding author: kaustavc@usc.edu.
[1]Authors are with the Department of Electrical Engineering, University of Southern California, USA.
[2]Authors are with the Department of Aeronautics and Astronautics, Stanford University, USA
[3]Authors are with NVIDIA Research

©2026 IEEE

We achieve this by reinterpreting trajectory predictors - typically generative models trained to forecast future agent behavior conditioned on histories and scene context (e.g., lane graphs, traffic signals) - as *data-driven forward reachable set (FRS) estimators*. These predictors, often instantiated as Gaussian Mixture Models (GMMs), implicitly learn a stochastic model of agent dynamics from logged driving data. Our *key insight* is that we can extract a probabilistic FRS by solving an optimization problem that identifies the smallest-volume set that captures a desired amount of probability mass under the learned distribution.

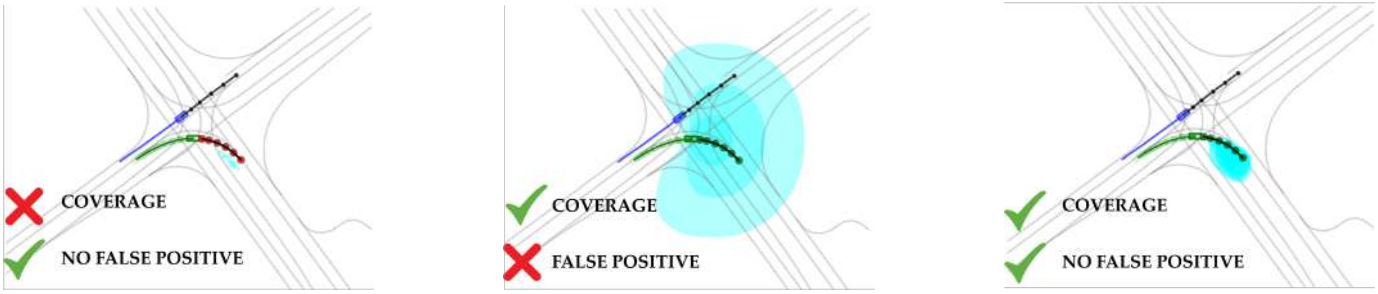
Leveraging the fact that many modern trajectory predictors output GMMs on the future states [3], [4], we formulate a convex relaxation of this problem that allows us to efficiently compute tight reachable sets that are significantly less conservative than traditional worst-case reachability approaches. However, these learned distributions are subject to modeling error and may fail to capture the true dynamics, especially under distribution shift.

To combat learning errors, we use conformal prediction (CP) [5] to calibrate the FRS. Rather than inflating the predicted sets indiscriminately, CP allows us to scale the covariance of each GMM component just enough to ensure that the reachable set covers the ground-truth trajectory with a user-specified error rate. Our overall algorithm to extract the FRS is called FORCE-OPT, which stands for FOrward Reachable sets from Conformal Estimation and convex OPTimization.

Although FORCE-OPT comes with a probabilistic guarantee on covering the ground truth when operating within the calibration distribution, it may still degrade in the presence of distribution shift. To handle this, we introduce a Bayesian filtering mechanism that monitors the consistency between predicted and observed agent behavior, adjusting the FRS conservativeness on the fly based on our confidence in the trajectory predictor.

This main contributions of the paper are: (i) We provide a rigorous formulation of the problem of estimating FRS from trajectory predictors. (ii) We introduce FORCE-OPT, an efficient algorithm that combines convex optimization and conformal prediction to compute calibrated, data-driven FRS from GMM-based predictors. (iii) We incorporate a belief-based Bayesian filtering mechanism that adapts the reachable set dynamically to account for predictor reliability under distribution shift. (iv) Through testing on nuScenes [6], real-world driving dataset, we demonstrate that FORCE-OPT achieves the best balance of false positive and false negative rates, outperforming both conservative baselines and uncalibrated learned methods.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Baseline Predictor (b) FRS (c) FORCE-OPT (ours)

Fig. 1: In this intersection scenario from the nuScenes dataset, we evaluate the safety of the ego (blue) vehicle's planned trajectory (black) in the vicinity of a non-ego contender agent (green). Our objective is to verify that the ego's planned path avoids any potential collisions with the contender by ensuring it remains outside all regions that the contender could possibly occupy in the future. We compare three methods of reachable set computation for safety evaluation of ego's motion plan. **(a) Baseline Predictor:** This method produces very small, overconfident predictions (cyan ellipses) for the other agent's future positions. While it correctly avoids a false positive (FP) collision warning, it fails to cover the true future positions of the contender (the cyan ellipses do not cover the true ground truth states states shown with red dots), making it unreliable for safety—indeed, in the detailed results later in the paper (Section V) this methods exhibits a high False Negative Rate (FNR). **(b) FRS (Forward Reachable Set):** This approach creates a large, conservative reachable set (the large cyan circle) for the other agent. It achieves excellent coverage, but its excessive size causes it to overlap with the ego vehicle's path, triggering a FP alarm. **(c) FORCE-OPT (ours):** Our method generates a more realistic and tighter reachable set (the smaller cyan area) for the other agent. It successfully achieves coverage by containing the agent's true future path (green dots) while being precise enough to avoid intersecting with the ego's trajectory, preventing a FP, correctly identifying the plan as safe. Overall, our method demonstrates a good balance of FPR and FNR as discussed in greater detail in Section V.

II. RELATED WORKS

Control-Theoretic Safety Evaluators. Classical approaches for safety evaluation of motion plans have relied on control-theoretic tools such as formal verification [7], [8], which provide mathematical guarantees that a system satisfies a predefined set of safety specifications. While rigorous, these methods often fall short when applied to modern robotic systems that encounter large uncertainties and operate in high-dimensional, stochastic environments. To handle this imminent uncertainty, reachability-based techniques—including those based on zonotopes [9], Hamilton-Jacobi formulations [10], [11], or sums-of-squares [12]—aim to characterize the set of all future states a system can reach under bounded disturbances. However, their worst-case assumptions often lead to overly conservative safety bounds that limit practical usability. On the other hand, recent set-propagation–based approaches, such as hybrid decomposition [13] are more computationally efficient enabling online safety evaluation; see [14] for more details on set-propagation based methods for reachability. Yet a core limitation of formal methods is their struggle to incorporate rich sensory inputs (e.g., images, LiDAR) and contextual cues that modern autonomous systems depend on. High-dimensional observations are often abstracted through models [15] or simplifying assumptions are made, reducing the granularity of the safety assessment. A common alternative to handling the difficult-to-model uncertainties in the real world is through data-driven methods. One such approach, tailored towards multi-agent settings, is to leverage learning-based trajectory predictors, as we will discuss next.

Trajectory Prediction-based Safety Evaluators. Motivated by the limitations of worst-case control-theoretic safety evaluators, there has been growing interest in safety evaluation frameworks built on top of learning-based trajectory predictors [4], [16], [17]. Although trajectory predictors were devised to aid with planning [18]–[21], they have found widespread use in other applications, such as failure monitoring [22], [23], mining interactive scenarios [24], [25], and assessing planner safety [26]–[28] which is of primary interest to us in this paper. Trajectory predictors are often used to identify reachable zones for other agents that the motion plan should stay out of, violation of which is considered to trigger a safety failure. Methods in this category include those that use trajectory predictors to guide controllability bounds on other agents to be used in a classical reachability formulation [15], [26], [28] and those that directly extract a reachable zone for contender agents from the predictor [27], [29]–[31]. In the latter category, approaches have explored fitting zonotopes to prediction outputs for estimating reachable sets [29], leveraged conformal prediction [27], [30], [31], and solved a probabilistic optimization on the output distribution of the predictor [32], [33]. However, many of these approaches require sampling [34]–[37] which is computationally expensive, lack multi-modality [27], [37], or lack calibration of the predictors [32], [38] resulting in learning-errors affecting the quality of the FRS. The approach presented in this paper is computationally efficient, accounts for multi-modality in predictions, calibrates the predictions to mitigate the impact of learning errors, and performs a belief-based FRS adaptation to impart some degree of OOD robustness.

III. EQUIVALENCE OF DETERMINISTIC AND STOCHASTIC NOTIONS OF FRS

We now present a rigorous mathematical framework to bridge the probabilistic and deterministic definitions of FRS. While the former is characteristic of contemporary trajectory forecasting literature, the latter is fundamental to control-theoretic safety analysis.

Let $x_t \in \mathcal{X} \subseteq \mathbb{R}^{n_x}$ denote the state of an agent at discrete time $t \in \mathbb{Z}_+$. The agent's control inputs are denoted by $u_t \in \mathcal{U}_t \subseteq \mathbb{R}^{n_u}$, while other disturbances arising from exogenous factors and uncertainties are captured by $w_t \in \mathcal{W}_t \subseteq \mathbb{R}^{n_w}$. The agent's dynamics evolve according to a discrete-time function

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

$f : \mathbb{Z}_+ \times \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathcal{X}$, which is assumed to be continuously differentiable in its arguments:

$$x_{t+1} = f(t, x_t, u_t, w_t). \tag{1}$$

For notational brevity, we define $\hat{f}_{t,u_t,w_t}(x_t) := f(t, x_t, u_t, w_t)$. The *forward-reachable set* (FRS) F_t at time t, starting from an initial state $x_0 \in \mathcal{X}$, is the set of all states that the agent can reach under a set of possible control actions and disturbances:

$$\begin{aligned} F_t := \{ & \hat{f}_{t-1,u_{t-1},w_{t-1}} \circ \cdots \circ \hat{f}_{0,u_0,w_0}(x_0) \\ & \mid u_i \in \mathcal{U}_i, w_i \in \mathcal{W}_i, i = 0, \cdots, t-1 \} \end{aligned} \tag{2}$$

In practice, however, the precise control sets $\mathcal{U}_i$ and disturbance sets $\mathcal{W}_i$ are unknown and influenced by hard-to-model aspects such as driver intent, road geometry, and local context. Classical reachability methods often adopt worst-case assumptions over these sets, resulting in overly conservative FRSs. Such assumptions are unnecessarily pessimistic - for instance, it is unreasonable to expect that a stopped vehicle at a red light will suddenly accelerate through the intersection while the traffic light is still red. A more realistic alternative is to infer the agent's behavior from data. In what follows, we introduce a probabilistic formulation of the FRS that facilitates using data-driven trajectory predictors to estimate likely future states of an agent, thereby, reducing the conservatism of the FRS.

We now reformulate the FRS using a probabilistic lens, as inspired by prior work [32]. As we highlight later in this paper, this probabilistic viewpoint is readily compatible with modern multi-modal trajectory predictors. Let vol represent the volume (Lebesgue measure) of a measurable set $\omega \in \Omega$, where Ω is the collection of all measurable subsets of $\mathcal{X}$. Let $\mu_t : \Omega \to [0, 1]$ represent a probability measure describing the distribution over the agent's state at time t. This measure arises as the push-forward of absolutely continuous probability distributions over the sequences of control and disturbance inputs, with support on $\mathcal{U}_i$ and $\mathcal{W}_i$, mapped through the composed dynamics in (2). The probabilistic FRS can then be defined through the following optimization problem:

$$\omega_t^* := \underset{\{\omega \in \Omega : \mu_t(\omega) = 1\}}{\arg\inf} \ \mathrm{vol}(\omega). \tag{3}$$

Intuitively, (3) returns the smallest (i.e., minimal-volume) set that captures all the future states of the agent under the distribution μ_t. Remarkably, this probabilistic formulation is equivalent to the classical deterministic definition in (2) almost everywhere (i.e., excluding a set of measure zero) as formalized in the following theorem.

Theorem 1 (Equivalence of FRS): Let the dynamics f in (1) be continuously differentiable. Let μ_t be the push-forward of absolutely continuous probability distributions with support on $\mathcal{U}_i$ and $\mathcal{W}_i$, mapped through the composed dynamics $\hat{f}_{t-1,u_{t-1},w_{t-1}} \circ \cdots \circ \hat{f}_{0,u_0,w_0}$. Then the sets F_t and ω_t^* defined in (2) and (3) are equal almost everywhere (except for sets with measure zero).

The proof of this theorem is available in the Appendix of the extended-version of the paper [39].

The probabilistic formulation in (3) offers a powerful bridge to estimating FRS using modern trajectory predictors, which are generative networks that learn distributions μ_t over future states of agents from some training driving data. These models implicitly encode the hard-to-specify uncertainties (e.g., $\mathcal{U}_i$, $\mathcal{W}_i$) and allow us to extract refined, data-driven FRSs that are less conservative than worst-case reachability. Nonetheless, it is worth noting that worst-case approaches remain valuable as fallback strategies when predictors fail due to learning errors or out-of-distribution (OOD) conditions.

In the remainder of this paper, we adopt the probabilistic view of the FRS to achieve two key goals: (i) develop an efficient algorithm, FORCE-OPT, to extract FRSs from learning-based trajectory predictors with probabilistic guarantees of ground-truth coverage; and (ii) use these calibrated FRSs to build a safety monitor that evaluates motion plans and empirically demonstrate low false-negative and false-positive rates under both in-distribution and OOD operation.

IV. FORCE-OPT

We now present FORCE-OPT, our algorithm for estimating FRS from trajectory predictors and calibrating them using conformal prediction. FORCE-OPT combines learned generative models with convex optimization to efficiently compute calibrated FRSs, and uses Bayesian filtering to hedge against out-of-distribution (OOD) deployment failures.

A. Trajectory Predictors

Without loss of generality, let the current time be 0. Denote the state of all agents in the scene by $\xi \in \mathcal{X}^{n_{\text{agents}}}$, and their historical trajectories over a horizon H by $\xi_{-H:0} \in \mathcal{X}^{Hn_{\text{agents}}}$. Let $m \in \mathcal{M}$ represent map and other contextual scene information. A traffic scene is then denoted by $s := (\xi_{-H:0}, m) \in \mathcal{X}^H \times \mathcal{M} =: \mathcal{S}$.

A trajectory predictor is a function $\phi : \mathcal{S} \to \mathcal{P}(\mathcal{X}^T)$ that maps a scene $s \in \mathcal{S}$ to a probability distribution $\hat{\mu} \in \mathcal{P}(\mathcal{X}^T)$ over future trajectories of horizon T. Most modern trajectory predictors [4], [16] represent ϕ using generative neural networks that output a sequence of Gaussian Mixture Models (GMMs) $\{\hat{\mu}_t\}_{t=1}^T$ – one for each future timestep. A GMM with K modes is of the form: $\hat{\mu}_t = \sum_{i=1}^K p_i \hat{\mu}_{t,i}$ where $p_i \geq 0$, $\sum_{i=1}^K p_i = 1$, and each mode $\hat{\mu}_{t,i}$ is a Gaussian distribution $\mathcal{N}(\overline{x}_i, \Sigma_i)$ with mean at $\overline{x}_i$ and a covariance Σ_i.

B. Extracting FRS from Trajectory Predictors

We treat $\hat{\mu}_t$ as a learned approximation of the push-forward distribution μ_t in (3). However, given that GMMs have an unbounded support, these measures make it infeasible to find a bounded set ω satisfying $\hat{\mu}_t(\omega) = 1$. To make the problem tractable, we relax the constraint to require only a minimum mass $\tau \in (0, 1)$:

$$\omega^* := \underset{\{\omega \in \Omega : \hat{\mu}_t(\omega) \geq \tau\}}{\arg\inf} \ \mathrm{vol}(\omega). \tag{4}$$

Ideally, we would choose a τ that is very close to 1. However, at this level of generality, this problem is very challenging to solve. Leveraging the fact that the distribution over each timestep is a GMM, we solve a tractable proxy for this problem by restricting the search to unions of sub-level sets of the GMM modes. For a given mode $\hat{\mu}_{t,i}$, define the Mahalanobis energy function $V_i(x) := (x - \overline{x}_i)^{\mathrm{T}} \Sigma_i^{-1} (x - \overline{x}_i)$,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

and its sublevel set $E_i(c_i) := \{x : V_i(x) \leq c_i\}$ for $c_i \geq 0$. We then solve:

$$\begin{aligned} c_1^*, \cdots, c_K^* = &\arg\inf_{c_1,\cdots,c_K} \sum_{i=1}^{K} \text{vol}(E_i(c_i)), \\ &\text{subject to } \sum_{i=1}^{K} p_i \hat{\mu}_{t,i}(E_i) \geq \tau, \\ &\quad c_i \geq 0 \quad \text{for all } i = 1, \cdots, K. \end{aligned} \tag{5}$$

The resulting FRS is $E^* = \bigcup_{i=1}^{K} E_i(c_i^*)$. Although E^* is not necessarily the smallest such set, it is a feasible solution to (4), and tight in practice. When $\mathcal{X} \subseteq \mathbb{R}^2$, it takes the form of a convex optimization, as detailed in the following theorem.

Theorem 2 (Convex Optimization for FRS Extraction): If $\mathcal{X} \subseteq \mathbb{R}^2$, then (5) takes the form of a convex optimization:

$$\begin{aligned} c_1^*, \cdots, c_K^* = &\arg\min_{c_1,\cdots,c_K} \sum_{i=1}^{K} \pi \sqrt{\lambda_{i,1}\lambda_{i,2}} c_i, \\ &\text{subject to } \sum_{i=1}^{K} p_i (1 - \exp(-c_i/2)) \geq \tau, \\ &\quad c_i \geq 0 \quad \text{for all } i = 1, \cdots, K. \end{aligned} \tag{6}$$

where $\lambda_{i,1}$ and $\lambda_{i,2}$ are eigen-values of Σ_i.

The proof of this theorem is available in the Appendix of the extended-version of the paper [39].

The optimization in (6) has a few key advantages. First, it computes the size of sub-level sets in accordance with the probabilities assigned to individual modes, resulting in "tighter" sets than if we were to simply choose the τ probability sub-level sets for each mode of the GMM. Second, unlike prior work [34], [40], this approach does not require us to draw samples from the distribution; instead, it exploits the GMM structure to solve the optimization problem and compute the FRS in microseconds, significantly lowering the FRS inference time. Finally, the solution to (6) is invariant to scaling of the covariance matrix, as shown in the following corollary of Theorem 2. This property will be instrumental in conformal calibration of the FRS.

Corollary 1 (Invariance of (6) to Covariance Scaling): The solution of the convex optimization (6) established in Theorem 2 is the same for any scaling $\alpha \in (0, \infty)$ of the GMM covariances $\{\Sigma_i\}_{i=1}^{K}$.

The proof of this corollary is available in the Appendix of the extended version of the paper [39].

C. Conformalizing FRS from Trajectory Predictors

The FRS obtained from (6) assumes the predictor's distribution $\hat{\mu}_t$ closely matches the true transition dynamics. In reality, due to modeling choices and limited training data, $\hat{\mu}_t$ may fail to cover the true ground-truth distribution. We address this by applying split conformal prediction to calibrate the reachable set to achieve high-probability coverage of the ground truth future.

We calibrate our FRS by scaling the covariance of the GMM modes. To do so, we define a non-conformity score function $\psi(s, x)$ as the smallest scaling factor $\alpha > 0$ such that the ground-truth x lies inside the FRS computed using covariances $\alpha\Sigma_i$:

$$\psi(s, x) := \inf\{\alpha \mid x \in E^*(\{\alpha\Sigma_i\}_{i=1}^{K})\} \tag{7}$$

where s is the scene information as defined in Section IV-A and $E^*(\{\Sigma_i\}_{i=1}^{K})$ be the FRS obtained by solving (6) for GMM covariances $\{\Sigma_i\}_{i=1}^{K}$. Thanks to Corollary 1, we only need to solve (6) once to obtain c_i^*, after which we can compute $\psi(s, x)$ analytically: $\psi(s, x) = \min\{V_i(x)/c_i^* \mid i = 1, \cdots, K\}$ where V_i is as defined in Section IV-B for Σ_i that are outputted by the trajectory predictor.

Given N calibration examples $\mathcal{D} := \{(s_j, x_j)\}_{j=1}^{N}$, if we let η be the $(1-\gamma)$ empirical quantile of the scores $\psi(s_j, x_j)$. Then, from [41, Proposition 2a], with probability $1 - \delta$:

$$\Pr_{s,x}(x \notin C(s, \eta)) < \gamma + \sqrt{-\log \delta / 2N}, \tag{8}$$

where $C(s, \eta)$ is the conformalized FRS given by:

$$C(s, \eta) := \{x : \psi(s, x) < \eta\} \tag{9}$$

$$= \{x : \inf\{\alpha | x \in E^*(\{\alpha\Sigma_i\}_{i=1}^{K})\} < \eta\} \tag{10}$$

$$= \bigcup_{0<\alpha<\eta} E^*(\{\alpha\Sigma_i\}_{i=1}^{K}) \tag{11}$$

$$= E^*(\{\eta\Sigma_i\}_{i=1}^{K}) \tag{12}$$

where the last line follows due to the monotonicity of sub-level sets:

Lemma 1: Let $\alpha_1, \alpha_2 \in (0, \infty)$ and $\alpha_1 < \alpha_2$, then $E^*(\{\alpha_1\Sigma_i\}_{i=1}^{K}) \subseteq E^*(\{\alpha_2\Sigma_i\}_{i=1}^{K})$.

Intuitively, (8) states that the probability of ground truth future state lying outside the conformalized FRS $C(s, \eta)$ can be controlled (with probability at least $1-\delta$) to be below a desired threshold through the choice of γ.

The proof of this lemma is available in the Appendix of the extended version of the paper [39].

D. Hedging Against OOD Failures via Bayesian Filtering

The conformalization process in Section IV-C allows FORCE-OPT to provide statistical guarantees on the coverage of predicted reachable sets. However, these guarantees hold only under the assumption that test-time inputs are drawn independently and identically (IID) from the same distribution as the calibration data. In practice—especially in autonomous driving—this assumption is often violated, and distribution shift can cause these guarantees to break down. Furthermore, as a learned system, the predictor is susceptible to errors arising from imperfect training leading to faulty predictions.

To address these challenges, we augment FORCE-OPT with a Bayesian filtering mechanism that dynamically adjusts uncertainty in the reachable set when the trajectory predictor appears unreliable. Specifically, we introduce a model confidence parameter $\beta \in [\beta_{\text{low}}, \beta_{\text{high}}]$ that reflects our trust in the predictor. This parameter scales the covariance of the GMMs used in the FRS adaptively, effectively dilating the predicted reachable sets to reflect greater uncertainty..

To track this confidence online, we adopt the Bayesian update scheme proposed in [26]. At each timestep t, we maintain a belief distribution over β, denoted by $\text{bel}^t(\beta)$. The belief is initialized uniformly: $bel^0(\beta_{low}) = bel^0(\beta_{high}) = 0.5$. The belief is then updated at every timestep using the likelihood

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

of the *observed* agent state x_t under the conformalized GMM, whose covariances are additionally scaled by the inverse of β:

$$\text{bel}^{t+1}(\beta) = \frac{\varphi\left(x_t; \text{GMM}\left(\{\bar{x}^t\}_{i=1}^K, \frac{1}{\beta}\{\eta\Sigma_i^t\}_{i=1}^K\right)\right)\text{bel}^t(\beta)}{\sum_{\tilde{\beta}} \varphi\left(x_t; \text{GMM}\left(\{\bar{x}^t\}_{i=1}^K, \frac{1}{\tilde{\beta}}\{\eta\Sigma_i^t\}_{i=1}^K\right)\right)\text{bel}^t(\tilde{\beta})}, \tag{13}$$

where $\tilde{\beta} \in \{\beta_{\text{low}}, \beta_{\text{high}}\}$. Here $\varphi(x, GMM(\cdot))$ denotes the likelihood of state x under the given GMM. Note that the GMM covariances are already scaled by the conformal calibration factor η; the inverse β scaling further adjusts the uncertainty based on current trust in the predictor. Finally, we compute the effective model confidence as the expected value under the belief: $\hat{\beta} = \mathbb{E}[\beta]$, and use this $\hat{\beta}$ to further scale the covariances of the distribution to adjust the final FRS returned by FORCE-OPT.

While conformal prediction provides statistical guarantees on coverage, these guarantees hold only in the average case. It does not account for rare but high-impact failures that may lie in the tails of the distribution. To address such worst-case scenarios we incorporate techniques from Hamilton-Jacobi (HJ) reachability analysis. Specifically, when the online model confidence β drops below a critical threshold, it indicates that the predictor may be operating outside the distribution represented by the calibration dataset. These are precisely the conditions under which FORCE-OPT becomes vulnerable to long-tailed failure modes.

To mitigate this risk, we employ two fallback mechanisms based on reachability theory: **(i)** a Parameterized FRS that adapts to observed uncertainty levels, and **(ii)** a Worst-Case FRS that assumes bounded adversarial disturbances. These fallback strategies are activated when β falls below the specified threshold. We explore the behavior and trade-offs of these approaches in detail in the results section, particularly in the context of different predictor types.

The belief update mechanism and switching strategy allow FORCE-OPT to maintain tight, calibrated bounds in-distribution, while conservatively hedging against uncertainty in out-of-distribution or low-confidence scenarios.

V. Experimental Results

We now present an empirical evaluation of FORCE-OPT to assess its effectiveness in safety-critical motion planning. Our experiments are designed to address the following key research questions: **(Q1)** How effective is FORCE-OPT at balancing completeness and soundness of safety evaluation for autonomous driving? **(Q2)** Do the belief-based extensions of FORCE-OPT result in more robust monitoring in out-of-distribution operation? **(Q3)** How effectively can FORCE-OPT and its extensions leverage multimodality of trajectory prediction? **(Q4)** Does FORCE-OPT offer a computational advantage that might enable online deployment?

A. Datasets

To answer the above questions, we evaluate FORCE-OPT's performance on nuScenes [6], a large-scale autonomous driving dataset. The nuScenes dataset includes approximately 15 hours of expert-labeled driving data in Boston and Singapore.

We train the Autobots predictor [4] on the training split of Singapore (comprising of 6,781 scenes) and then test it on the test splits of both cities (4,589 scenes in total, with 2,653 from Singapore and 1,936 from Boston.). This setup allows us to test the performance of our approach on in distribution (ID) where the model is trained and tested on the same city as well as out-of-distribution (OOD) where the model is tested on a city different from the one it is trained on. The predictor takes in 5 seconds of scene information as input to predict a trajectory for the next 3 seconds at 2 Hz.

B. Synthetic Unsafe Data Generation

Although nuScenes provides diverse testing conditions, all the driving data available in the dataset is inherently safe, which makes it challenging to assess the monitor's ability to detect potential safety violations. To remedy this, we synthetically generate unsafe scenarios by modifying scenes within the nuScenes dataset. Specifically, we identify potential intersections between the trajectories of the ego vehicle and a surrounding agent, defining the point of closest approach p_c as the collision point. Let the ego and the contender arrive at p_c at times t_e and t_o, respectively. We use a bicycle dynamics model for the ego vehicle and optimize its trajectory using Sequential Least Squares Programming (SLSQP) to force it to reach p_c at t_o, while obeying initial conditions and physical constraints. The resulting trajectory of the ego vehicle, along with the original trajectory of the contender, constitutes a synthesized unsafe scenario. This yields plausible but unsafe plans, which serve as ground truth for evaluating false negative rates. We were able to generate 800 such collision scenes included in our test dataset

C. Metrics

The primary objective of this paper is to assess the safety of motion plans. To that end, we focus on two broad categories of metrics: **completeness** and **soundness**. These metrics evaluate on a per-frame basis how effectively and reliably a safety assessment algorithm captures true safety violations while minimizing over-conservatism.

- **Completeness** is quantified using two key metrics:
  1) *Coverage (Cov):* The fraction of ground-truth future trajectories that fall within the predicted reachable set. High coverage indicates that the predicted FRS captures the actual future behavior well.
  2) *False Negative Rate (FNR):* The proportion of true collision cases (from synthesized unsafe data) that are not flagged by the monitor.
- **Soundness** is evaluated using the *False Positive Rate (FPR)*: The fraction of safe scenarios that are incorrectly flagged as unsafe. A low FPR indicates soundness, avoiding unnecessary interventions.
- **Balance between Completeness and Soundness** is evaluated using the *Balanced Error Rate (BER)*: Arithmetic mean of FPR and FNR effectively capturing the tradeoff between them.

D. Baselines and Variants

We categorize all methods in three broad categories:

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Approach	**Uncalibrated Trajectory Predictor**			**Calibrated Trajectory Predictor with Conformal Prediction**					**Data-free**
Metric	99% CI	Parametric WC-FRS	Nakamura et al. [26]	Lindemann et al. [27]	FORCE-OPT (ours)	FORCE-OPT+ belief (ours)	FORCE-OPT+ pWC-FRS (ours)	FORCE-OPT+ WC-FRS (ours)	Worst Case FRS
Cov (↑)	54.76%	97.40%	86.84%	89.50%	89.95%	93.94%	94.90%	96.01%	**99.53%**
FPR (↓)	**1.19%**	19.68%	7.33%	31.12%	8.62%	15.57%	13.35%	24.43%	43.73%
FNR (↓)	33.33%	**0%**	36.36%	**0%**	3.03%	**0%**	**0%**	**0%**	**0%**
BER (↓)	17.26%	9.84%	20.84%	15.56%	**5.83%**	7.78%	6.78%	12.21%	21.86%

TABLE I: Performance comparison across methods (best-performing values in bold) in Singapore (ID).

- **Uncalibrated Trajectory Predictor:** The methods that fall under this class directly use the trajectory predictor without calibrating them. **99% CI**: The original predictor where the sets occupy 99% of GMM probability mass. $E = \bigcup_{i=1}^{K} E_i(C)$ where C = value at 99th Percentile of a χ^2 distribution. **Parametric Worst Case-FRS (pWC-FRS)**: We obtain the control bounds as the 3σ (or 99% confidence interval) support of the Gaussian control distribution for each mode predicted by a trajectory predictor and then estimate a worst-case FRS for each mode and take the union of the sets. (the parameter is the velocity and the control bound of he agent) **Nakamura et al. [26]**: Adapts FRS via belief tracking of a trajectory predictor's performance. In this case the control bounds enclose the 3% probaility mass around the mean of the normal distribution predicted by the trajectory predictor [16].
- **Calibrated Trajectory Predictor:** The methods within this class leverage trajectory predictors that are calibrated using CP. We use a dataset with a cardinality of 35220 for calibration and set the desired coverage probability in CP to 0.95. **Lindemann et al.** [27]: Provides coverage guarantees via conformalization between the highest likelihood trajectory and the ground truth. **FORCE-OPT (Ours)**: Solves the convex optimization from Sec. IV-B using GMMs from learned predictors, along with the calibration schemes in Sec.IV-C. **FORCE-OPT + belief**: Additionally adjusts the GMM covariances of FORCE-OPT using Bayesian filtering approach in Sec. IV-D. For this and all the subsequent methods that use belief-based adaptation, we set $\beta_{\text{low}} = 0.3$ and $\beta_{\text{high}} = 1$. **FORCE-OPT + pWC-FRS**: A hybrid approach that switches from FORCE-OPT to Parametric WC-FRS when $\beta < 0.75$ indicating a drop in the predictor's performance. **FORCE-OPT + WC-FRS**: This approach switches from FORCE-OPT to worst-case FRS when $\beta < 0.75$.
- **Data-Free:** This category contains only one baseline metric **Worst Case FRS (WC-FRS)** which computes worst-case FRS assuming 4D Dubins vehicle dynamics with bounded control inputs.

E. Results and Discussion

The results for in-distribution evaluation of all approaches discussed in Section V-D are summarized in Table I, while those for OOD are summarized in Table II.

1) Balance between Completeness and Soundness (Q1)

From Table I we observe that FORCE-OPT achieves the lowest BER, indicating the best balance between FPR and FNR. The metrics that use CP-based calibration on the trajectory predictor have a significantly lower FNR at the expense of a slightly higher FPR than the ones that do not calibrate the predictor, generally indicating that the trajectory predictor without calibration tends to be over-optimistic in safety assessment missing out safety critical events, highlighting the importance of calibrating the trajectory predictor—one exception to this observation is the parametric FRS that counters the uncalibrated predictor's over-optimism by plugging the predicted control bounds to solve for the worst-case FRS. We also note that FORCE-OPT and its belief-based extensions, in comparison to [27], have a lower FPR while maintaining a similar FNR and higher coverage rates. These additional gains are the outcome of the Theorem 1-inspired convex optimization (6) that more precisely models the forward reachable space while leveraging multi-modality (unlike [27] which uses a single mode); this is discussed in greater detail in Section V-E3. FORCE-OPT has lower FPR and slightly worse FNR than its belief-based variants, which is in alignment with our expectations as the belief-based approaches introduce greater conservatism in the event when the predictor's performance deteriorates.

2) Out-of-distribution Robustness (Q2)

The belief-based adaptation mechanism enables smooth adjustment to uncertain contexts without excessive conservatism, as evidenced by the fact that the performance of the belief-based variants of FORCE-OPT remains similar between Tables I and II and the best performing metric according to BER in Table II is FORCE-OPT + pWC-FRS. Comparing the results in Table I (ID) and Table II (OOD), we also observe that the metrics that use an uncalibrated trajectory predictor suffer a drop in performance, especially with the FNR which jumps from 33.33% and 36.36% (ID) to 51.85% and 55.56% (OOD) for 99% CI and [26], respectively—as before, parametric FRS is an exception to this for the same reason mentioned in Section V-E1. On the other hand, the approaches that use the calibrated trajectory predictor exhibit stronger OOD robustness; albeit, there is a modest increase in the FNR for all these approaches. Unsurprisingly, the performance of Worst-Case FRS is unaffected by the distribution shift.

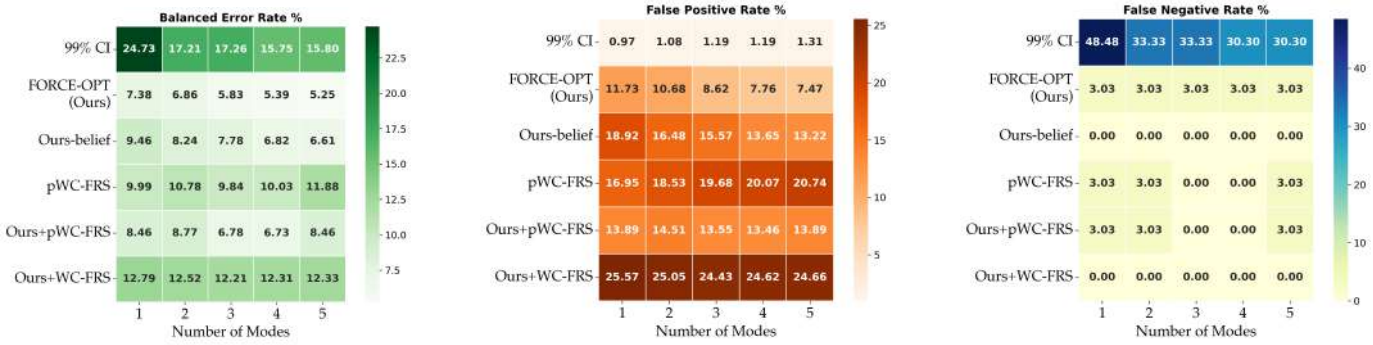
3) Impact of Multi-modality (Q3)

One of the key strengths of FORCE-OPT is its ability to systematically integrate probabilities from multiple prediction modes. To study the impact of multi-modality, we ablate the performance of the metrics in Section V-D that can handle multimodality, i.e., 99% CI, parametric FRS, FORCE-OPT and its belief-based variants, on the Singapore dataset against the number of GMM modes. As the number of modes increase

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Approach	**Uncalibrated Trajectory Predictor**			**Calibrated Trajectory Predictor with Conformal Prediction**					**Data-free**
Metric	99% CI	Parametric WC-FRS	Nakamura et al. [26]	Lindemann et al. [27]	FORCE-OPT (ours)	FORCE-OPT+ belief (ours)	FORCE-OPT+ pWC-FRS (ours)	FORCE-OPT+ WC-FRS (ours)	Worst Case FRS
Cov (↑)	62.85%	97.45%	88.99%	88.02%	90.06%	93.57%	95.07%	96.06%	**99.46%**
FPR (↓)	**1.18%**	17.08%	7.27%	36.38%	5.22%	9.16%	9.85%	21.15%	45.42%
FNR (↓)	51.85%	3.70%	55.56%	9.26%	12.96%	7.41%	3.70%	**0%**	**0%**
BER (↓)	26.52%	10.39%	31.41%	22.79%	9.09%	8.28%	**6.77%**	10.57%	22.71%

TABLE II: Performance comparison across methods (best-performing values in bold) in Boston (OOD).



(a) BER vs. Number of Modes (b) FPR vs. Number of Modes (c) FNR vs. Number of Modes

Fig. 2: Ablation with different number of GMM modes from the trajectory predictor. The performance of FORCE-OPT and its belief-based variants improves with more GMM modes.

from one to five, we observe that the BER drops for all methods other than pWC-FRS Fig. 2a. The drop in BER in different methods is fueled by different reasons: for 99% CI, the BER improves because the FNR improves with more modes Fig. 2c, while for FORCE-OPT and its variants the BER improvement arises from an improvement in the FPR Fig. 2b. At first look, it is indeed surprising that more modes result in a lower FPR for FORCE-OPT; however, this counter-intuitive outcome is the result of the fact that greater multi-modality requires smaller set inflations via CP, as evidenced by the fact that the α's decrease as the number of modes increase, as shown in Table III. If a mode other than the most-likely one is nearer to the ground truth in the calibration set, then the amount of set inflation needed to cover that ground-truth position would have to be less than the inflation needed for the most-likely mode that is further away. In our experiments we found that the computational cost scales linearly with GMM modes (e.g., 0.022s for 3 vs. 0.029s for 10 modes), yet overall, our ablations suggest that greater multi-modality promotes better safety assessment by allowing us to reason about multiple plausible future outcomes which could be closer to the ground truth behavior than whatever the model deems to be the most likely.

# Modes	**t = 1**	**t = 2**	**t = 3**	**t = 4**	**t = 5**	**t = 6**
1	3.18	11.73	21.63	32.84	43.99	53.44
2	2.23	7.72	14.51	23.36	32.3	41.43
3	2.11	7.07	13.21	21.06	29.23	37.48
4	2.04	6.45	11.67	18.69	25.55	33.65
5	1.92	6.10	11.20	17.43	24.05	31.72

TABLE III: Calibrated α for each timestep along the predicted trajectory as the number of modes vary. The table shows that as the number of modes increase the amount of set inflation α needed for calibration reduces.

4) *Compuational Efficiency (Q4)*

In Table IV, we show the computation time for all the methods presented in Section V-D along with their performance on BER in Tables I and II. FORCE-OPT demonstrates fast runtimes while achieving strong BER results in both ID and OOD settings, outperforming faster baselines such as 99% CI and [27]. Notably, adding belief tracking adds negligible overhead—FORCE-OPT + belief is only 0.001 seconds slower than FORCE-OPT alone. With the exception of FORCE-OPT + pWC-FRS, all other variants of FORCE-OPT are faster than 0.1 seconds, suggesting that these algorithms are well-suited for deployment in real-time, safety-critical applications.

Method	**Time (s)**	**ID BER**	**OOD BER**
99% CI	**0.001**	17.26%	26.52%
pWC-FRS	0.187	9.84%	10.39%
Nakamura et al. [26]	0.155	20.84%	31.41%
Lindemann et al. [27]	0.005	15.56%	22.79%
FORCE-OPT (ours)	0.022	**5.83%**	9.09%
FORCE-OPT + belief (ours)	0.023	7.78%	8.28%
FORCE-OPT + pWC-FRS (ours)	0.210	6.78%	**6.77%**
FORCE-OPT + WC-FRS (ours)	0.079	12.21%	10.57%
WC-FRS	0.056	21.86%	22.71%

TABLE IV: Computation Time for Different FRS Methods

VI. Conclusion and Future Work

This paper introduced FORCE-OPT, a principled framework for evaluating the safety of motion plans using trajectory predictors as estimators of forward reachable sets. By combining convex optimization, conformal prediction, and Bayesian filtering, our method generates calibrated uncertainty sets that balance completeness (low false negatives) with soundness (low false positives). Empirical results on nuScenes demonstrate that FORCE-OPT significantly outperforms both

conservative model-based and raw learning-based baselines, while gracefully handling out-of-distribution scenarios. We believe FORCE-OPT offers a promising building block for runtime safety monitoring in learned autonomy stacks.

Building on this foundation, this work opens up several directions for exploration: (i) While the trajectory predictor conditions on scene context, FORCE-OPT itself operates independently for each agent when computing FRS. Joint multi-agent reachability, especially in dense traffic scenarios with interdependent behaviors, remains an open direction. (ii) Trajectory predictors are trained to distributionally mimic the observed data, not to facilitate the extraction of FRS. Training a neural FRS generator that directly outputs sets is another exciting open direction. (iii) Extending FORCE-OPT beyond two dimensions to address challenges in other domains in robotics such as drones or manipulators.

REFERENCES

[1] S. Teng *et al.*, "Motion planning for autonomous driving: The state of the art and future perspectives," *TIV*, vol. 8, no. 6, pp. 3692–3711, 2023.
[2] W. Schwarting *et al.*, "Planning and decision-making for autonomous vehicles," *Annual Review of CRAS*, vol. 1, no. 1, pp. 187–210, 2018.
[3] T. Salzmann *et al.*, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *ECCV*. Springer, 2020, pp. 683–700.
[4] R. Girgis *et al.*, "Latent variable sequential set transformers for joint multi-agent motion prediction," in *International Conference on Learning Representations*, 2022.
[5] G. Shafer and V. Vovk, "A tutorial on conformal prediction." *JMLR*, vol. 9, no. 3, 2008.
[6] H. Caesar *et al.*, "nuscenes: A multimodal dataset for autonomous driving," 2020, pp. 11 621–11 631.
[7] M. Luckcuck *et al.*, "Formal specification and verification of autonomous robotic systems: A survey," *CSUR*, vol. 52, no. 5, pp. 1–41, 2019.
[8] S. Mitra *et al.*, "Formal verification techniques for vision-based autonomous systems–a survey," in *Principles of Verification: Cycling the Probabilistic Landscape: Essays Dedicated to Joost-Pieter Katoen on the Occasion of His 60th Birthday, Part III*. Springer, 2024, pp. 89–108.
[9] M. Althoff and J. M. Dolan, "Reachability analysis of nonlinear systems with uncertain parameters using conservative linearization," *CDC*, 2011.
[10] J. Lygeros *et al.*, "Controllers for reachability specifications for hybrid systems," *Automatica*, vol. 35, no. 3, pp. 349–370, 1999.
[11] I. M. Mitchell *et al.*, "A toolbox of hamilton–jacobi solvers for analysis of nondeterministic continuous and hybrid systems," *Hybrid Systems: Computation and Control*, 2005.
[12] A. Majumdar and R. Tedrake, "Funnel libraries for real-time robust feedback motion planning," *IJRR*, vol. 36, no. 8, pp. 947–982, 2017.
[13] T. J. Bird *et al.*, "Hybrid zonotopes: A new set representation for reachability analysis of mixed logical dynamical systems," *Automatica*, 2023.
[14] M. Althoff *et al.*, "Set propagation techniques for reachability analysis," *Annual Review of CRAS*, vol. 4, no. 1, pp. 369–395, 2021.
[15] D. Fridovich-Keil *et al.*, "Confidence-aware motion prediction for real-time collision avoidance1," *IJRR*, vol. 39, no. 2-3, pp. 250–265, 2020.
[16] T. Salzmann *et al.*, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *ECCV*, 2020, pp. 683–700.
[17] J.-J. Hwang *et al.*, "EMMA: End-to-end multimodal model for autonomous driving," *Transactions on Machine Learning Research*, 2025.
[18] P. Karkus *et al.*, "Diffstack: A differentiable and modular control stack for autonomous vehicles," in *CoRL*. PMLR, 2023, pp. 2170–2180.
[19] Y. Chen *et al.*, "Interactive joint planning for autonomous vehicles," *RA-L*, 2023.
[20] S. Casas *et al.*, "Mp3: A unified model to map, perceive, predict and plan," in *CVPR*, 2021, pp. 14 403–14 412.
[21] A. Cui *et al.*, "Lookout: Diverse multi-future prediction and planning for self-driving," in *ICCV*, 2021, pp. 16 107–16 116.
[22] P. Antonante *et al.*, "Task-aware risk estimation of perception failures for autonomous vehicles," *arXiv preprint arXiv:2305.01870*, 2023.
[23] K. Chakraborty *et al.*, "System-level safety monitoring and recovery for perception failures in autonomous vehicles," *arXiv preprint arXiv:2409.17630*, 2024.
[24] A. Dinparastdjadid *et al.*, "Measuring surprise in the wild," *arXiv preprint arXiv:2305.07733*, 2023.
[25] W. Ding *et al.*, "Surprise potential as a measure of interactivity in driving scenarios," *arXiv preprint arXiv:2502.05677*, 2025.
[26] K. Nakamura and S. Bansal, "Online update of safety assurances using confidence-based predictions," in *ICRA*. IEEE, 2023, pp. 12 765–12 771.
[27] L. Lindemann *et al.*, "Safe planning in dynamic environments using conformal prediction," *RA-L*, 2023.
[28] A. Li *et al.*, "Prediction-based reachability for collision avoidance in autonomous driving," *arXiv preprint arXiv:2011.12406*, 2020.
[29] L. Paparusso *et al.*, "Zapp! zonotope agreement of prediction and planning for continuous-time collision avoidance with discrete-time dynamics," in *ICRA*. IEEE, 2024, pp. 9285–9292.
[30] A. Dixit *et al.*, "Adaptive conformal prediction for motion planning among dynamic agents," in *L4DC*. PMLR, 2023, pp. 300–314.
[31] Y. Chen *et al.*, "Reactive motion planning with probabilisticsafety guarantees," in *CoRL*. PMLR, 2021, pp. 1958–1970.
[32] K. Driggs-Campbell *et al.*, "Robust, informative human-in-the-loop predictions via empirical reachable sets," *arXiv preprint arXiv:1705.00748*, 2017.
[33] A. Devonport and M. Arcak, "Estimating reachable sets with scenario optimization," in *L4DC*. PMLR, 2020, pp. 75–84.
[34] R. Tumu *et al.*, "Multi-modal conformal prediction regions by optimizing convex shape templates," in *L4DC*. PMLR, 2024, pp. 1343–1356.
[35] J. Xiang and J. Chen, "Convex approximation of probabilistic reachable sets from small samples using self-supervised neural networks," *arXiv preprint arXiv:2411.14356*, 2024.
[36] E. Dietrich *et al.*, "Data-driven reachability with scenario optimization and the holdout method," *Proceedings of the 6th Annual Learning for Dynamics amp; Control Conference*, pp. 514–527, 2024.
[37] S. Braun *et al.*, "Minimum volume conformal sets for multivariate regression," *arXiv preprint arXiv:2503.19068*, 2025.
[38] K. Driggs-Campbell *et al.*, "Integrating intuitive driver models in autonomous planning for interactive maneuvers," *ITS*, vol. 18, no. 12, pp. 3461–3472, 2017.
[39] K. Chakraborty, Z. Feng, S. Veer, A. Sharma, W. Ding, S. Topan, B. Ivanovic, M. Pavone, and S. Bansal, "Safety evaluation of motion plans using trajectory predictors as forward reachable set estimators," *arXiv preprint arXiv:2507.22389*, 2025.
[40] T. Lew and M. Pavone, "Sampling-based reachability analysis: A random set theory approach with adversarial sampling," in *CoRL*. PMLR, 2021, pp. 2055–2070.
[41] V. Vovk, "Conditional validity of inductive conformal predictors," in *ACML*. PMLR, 2012, pp. 475–490.