2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

DTP-Attack: A decision-based black-box adversarial attack on trajectory prediction

Jiaxiang Li[1], Jun Yan[1], Daniel Watzenig[2] and Huilin Yin[1]

***Abstract*— Trajectory prediction systems are critical for autonomous vehicle safety, yet remain vulnerable to adversarial attacks that can cause catastrophic traffic behavior misinterpretations. Existing attack methods require white-box access with gradient information and rely on rigid physical constraints, limiting real-world applicability. We propose DTP-Attack, a decision-based black-box adversarial attack framework tailored for trajectory prediction systems. Our method operates exclusively on binary decision outputs without requiring model internals or gradients, making it practical for real-world scenarios. DTP-Attack employs a novel boundary walking algorithm that navigates adversarial regions without fixed constraints, naturally maintaining trajectory realism through proximity preservation. Unlike existing approaches, our method supports both intention misclassification attacks and prediction accuracy degradation. Extensive evaluation on nuScenes and Apolloscape datasets across state-of-the-art models including Trajectron++ and Grip++ demonstrates superior performance. DTP-Attack achieves $41-81\%$ attack success rates for intention misclassification attacks that manipulate perceived driving maneuvers with perturbations below $0.45\,\mathrm{m}$, and increases prediction errors by $1.9-4.2\times$ for accuracy degradation. Our method consistently outperforms existing black-box approaches while maintaining high controllability and reliability across diverse scenarios. These results reveal fundamental vulnerabilities in current trajectory prediction systems, highlighting urgent needs for robust defenses in safety-critical autonomous driving applications. Our code is available at the repository: https://github.com/eclipse-bot/DTP-Attack.**

I. Introduction

Autonomous vehicles (AVs) rely critically on trajectory prediction systems [1], [2] to forecast the future movements of surrounding traffic participants. These deep learning-based models enable vehicles to anticipate potential hazards and make safe navigation decisions, forming a cornerstone of modern autonomous driving technology. However, recent advances in adversarial machine learning [3] reveal that neural networks can be manipulated through subtle input perturbations, causing incorrect predictions while appearing to function normally. In autonomous driving, such vulnerabilities pose severe safety risks, as a compromised trajectory prediction system could cause catastrophic misinterpretations of nearby traffic behavior.

The threat model for trajectory prediction attacks encompasses two primary objectives, as illustrated in Fig. 1.

[1]Jiaxiang Li, Jun Yan and Huilin Yin (e-mail: zmbdsilver@gmail.com, yanjun@tongji.edu.cn, yinhuilin@tongji.edu.cn) are with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China. (Corresponding author: Huilin Yin.)

[2]Daniel Watzenig (e-mail: daniel.watzenig@tugraz.at) is with the Graz University of Technology and the Virtual Vehicle Research, Graz 8010, Austria.

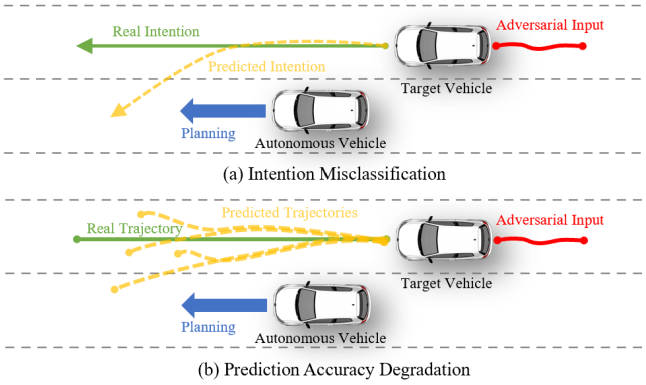

Fig. 1: Two types of adversarial attacks on trajectory prediction. (a) Intention misclassification. (b) Prediction accuracy degradation.

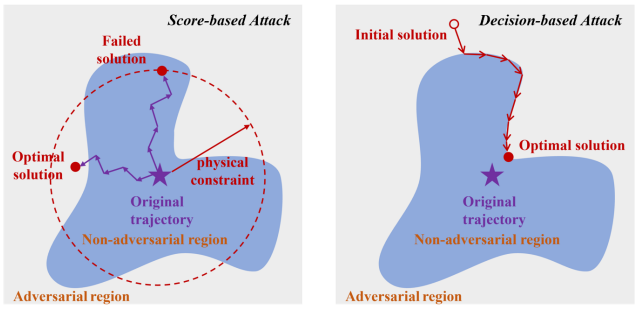

Fig. 2: Conceptual comparison of score-based (left) vs. decision-based (right) adversarial attack optimization landscapes.

Intention misclassification attacks [4] aim to manipulate the perceived driving maneuvers of target vehicles, causing a vehicle traveling straight to appear as though it will execute a lane change or turn. Prediction accuracy degradation attacks systematically increase standard trajectory prediction error metrics, corrupting the system's ability to accurately forecast future positions and potentially compromising downstream planning algorithms.

Compared to traditional image classification domains [5], [6], adversarial attacks on trajectory prediction systems present distinct challenges. First, trajectory data consists of spatiotemporal sequences that must satisfy kinematic constraints, since arbitrary perturbations can easily violate realistic vehicle dynamics [7]. Second, trajectory prediction models typically output continuous coordinate sequences rather than discrete class probabilities, complicating the definition of attack objectives. Third, real-world threat scenarios often involve black-box access [8], [9] where attackers cannot observe model internals or gradients.

As illustrated in Fig. 2, existing adversarial attack methods [10]–[13] for trajectory prediction employ score-based optimization paradigms that exhibit fundamental structural limitations. Current approaches universally rely on gradient information or detailed model access while imposing artificially specified physical constraints that inadequately capture real-world vehicle dynamics. This score-based formulation creates the discontinuous search landscapes depicted in the figure, where rigid constraint boundaries systematically trap optimization algorithms in local optima [14], preventing convergence to effective solutions and generating unrealistic trajectories. Moreover, score-based methods are inherently restricted to prediction accuracy degradation objectives, fundamentally precluding adaptation to intention misclassification attacks and significantly limiting their applicability in realistic threat scenarios.

We propose DTP-Attack, a decision-based black-box adversarial attack framework that addresses these limitations through a fundamentally different optimization paradigm. Our approach operates exclusively on binary feedback from adversarial criteria functions, eliminating the requirement for gradient information or model internals. As demonstrated in Fig. 2, this decision-based formulation creates smooth adversarial regions that enable effective boundary walking optimization. Rather than imposing explicit physical constraints, our method maintains trajectory realism through proximity preservation, allowing unconstrained exploration of the adversarial space while naturally satisfying kinematic feasibility. The proposed boundary walking algorithm iteratively refines adversarial trajectories through orthogonal and forward steps along the decision boundary, consistently converging to near-optimal solutions.

Our experimental evaluation across multiple state-of-the-art models and datasets demonstrates superior attack effectiveness compared to existing black-box baselines. DTP-Attack achieves attack success rates ranging from 41% to 81% for intention misclassification while maintaining perturbation magnitudes below $0.45\,\mathrm{m}$. For accuracy degradation objectives, our method increases prediction errors by factors of $1.9\times$ to $4.2\times$, significantly compromising system reliability through minimal trajectory modifications.

Our main contributions are threefold: First, we propose a decision-based black-box attack method for trajectory prediction that operates without requiring model internals or gradient information, making it applicable to real-world scenarios. Second, we develop a constraint-free optimization approach that naturally maintains trajectory realism while allowing attackers to control attack effects according to their objectives. Third, we provide comprehensive experimental validation showing that our method consistently outperforms existing approaches while maintaining attack stealthiness and reliability across diverse model architectures and datasets.

II. Related Work

A. Trajectory Prediction in Autonomous Driving

Neural trajectory prediction systems enable autonomous vehicles to forecast future movements of surrounding agents based on historical observations and environmental context. Leading approaches include Trajectron++ [1], which models driving scenarios as directed graphs to capture agent interactions and incorporates vehicle dynamics for realistic predictions, and Grip++ [2], which uses graph convolutions with temporal modeling. While these methods achieve strong performance on standard benchmarks like nuScenes [15] and Apolloscape [16], they inherit the vulnerability of deep neural networks to adversarial perturbations.

B. Adversarial Attacks on Trajectory Prediction

Adversarial attacks on trajectory prediction present unique challenges compared to traditional domains like image classification, as trajectory data must satisfy physical constraints while maintaining temporal coherence. Existing methods can be categorized based on the level of target model access required.

White-box attacks leverage complete model knowledge but face significant practical limitations. Zhang et al. [10] applied Projected Gradient Descent (PGD) to spatial coordinates, though this produced unrealistic trajectories. Cao et al. [11] improved physical feasibility by perturbing control signals and recovering trajectories through dynamic models. Tan et al. [13] extended this work to targeted attacks that deceive models into predicting user-specified outcomes. However, these methods require proprietary model internals and rely on manually specified constraints that fail to capture real-world vehicle dynamics complexity.

Black-box attacks offer greater practical relevance but suffer from fundamental limitations. Zhang et al. [10] adapted Particle Swarm Optimization (PSO) for trajectory attacks, but PSO's design for unconstrained optimization poorly suits the nonlinear constrained problems inherent in realistic trajectory generation. Critically, existing approaches rely on score-based optimization requiring continuous confidence scores or gradient information. This creates discontinuous search landscapes that trap algorithms in local optima and limit convergence. Moreover, current methods focus primarily on prediction accuracy degradation and lack flexibility for diverse attack objectives such as intention misclassification.

Our decision-based framework addresses these limitations by operating exclusively on binary decision outputs without requiring gradient information or manually specified constraints, enabling more effective optimization across diverse attack objectives.

III. Problem Formulation

A. Trajectory Prediction Formulation

We address the trajectory prediction problem, which operates at fixed time intervals to forecast future movement patterns for multiple agents within a given scenario. The task requires predicting probability distributions over possible future trajectories for N agents based on their historical states and contextual information.

We represent each agent's state as a D-dimensional vector $s \in \mathbb{R}^D$, which captures spatial coordinates along with additional attributes such as semantic class, agent dimensions,

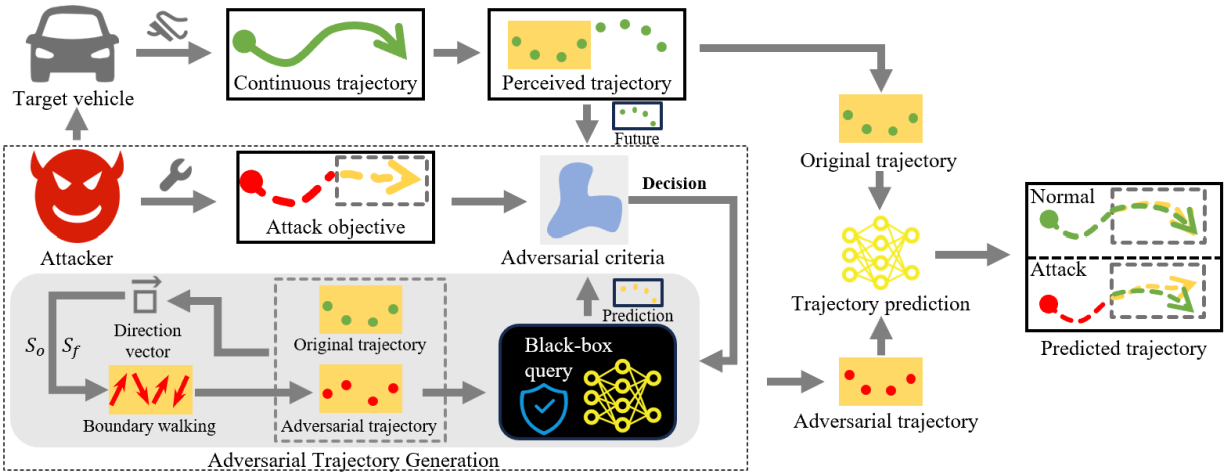

Fig. 3: A decision-based black-box adversarial attack on trajectory prediction (DTP-Attack) methodology overview.

and heading direction. Our analysis concentrates on spatial coordinates $p \in \mathbb{R}^2$, representing each agent's position in two-dimensional space.

At any given time t, the prediction system receives historical state information spanning L_I previous timesteps. We denote this input history as $\mathbf{H}_t$, $\mathbf{H}_t = s^{1,\dots,N}_{t-L_I+1:t} \in \mathbb{R}^{L_I \times N \times D}$, where the tensor dimensions represent time, agents, and state features respectively. The system also incorporates additional contextual information $\mathbf{I}_t$, which may include map data, traffic signals, or other environmental factors.

The system estimates probability distributions over future trajectories spanning the subsequent L_O timesteps. We represent predicted trajectory coordinates as $\mathbf{P}_t = p^{1,\dots,N}_{t+1:t+L_O} \in \mathbb{R}^{L_O \times N \times 2}$ and ground truth future states as $\mathbf{F}_t = s^{1,\dots,N}_{t+1:t+L_O} \in \mathbb{R}^{L_O \times N \times D}$. The trajectory prediction model functions as $\mathbf{P}_t = \Phi(\mathbf{H}_t, \mathbf{I}_t)$, mapping historical states and contextual information to predicted trajectory distributions.

B. Adversarial Attack Formulation

Our attack operates in real-world autonomous driving environments by exploiting how AVs predict the future movements of surrounding traffic. In this scenario, an attacker uses a vehicle to execute specific movement patterns designed to mislead the AV's trajectory prediction system.

The approach works as follows: A vehicle near the AV is selected as the target vehicle with an original trajectory of $X_t \in \mathbf{H}_t$. The attacker drives this vehicle following a calculated movement pattern that, when processed by the AV's perception system, creates an adversarial trajectory X_t^* derived from but subtly different than X_t. These small deviations are precisely engineered using our DTP-Attack method to remain within physical feasibility constraints while causing specific prediction errors.

We examine black-box attack scenarios where attackers possess no internal knowledge of the AV's trajectory prediction model architecture or parameters. Attackers can only observe system behavior through input-output relationships, making this approach more applicable to real-world conditions than attacks requiring detailed model access. This constraint increases practical relevance while presenting greater technical challenges for attack generation.

IV. Adversarial Trajectory Generation

As shown in Fig. 3, DTP-Attack transforms the attack objective into binary adversarial criteria that partition trajectory space into adversarial and non-adversarial regions. The framework operates through iterative black-box queries to the target model, employing boundary walking optimization with orthogonal and forward steps to navigate the adversarial region. This process systematically converges to minimal perturbations that achieve specified attack goals while preserving trajectory realism.

A. Attack Objectives

DTP-Attack targets two distinct categories of trajectory prediction failures: intention misclassification and prediction accuracy degradation. Each objective requires different manipulation strategies and evaluation metrics.

Intention Misclassification: This objective aims to cause misclassification of intended maneuvers. For example, an attack might make a vehicle traveling straight appear to be turning left or right, or make a vehicle maintaining constant speed appear to be accelerating. To achieve this, predicted trajectories must be shifted in specific directions, such as leftward for simulating left turns, or forward along the longitudinal axis for simulating acceleration.

We design four directional metrics to calculate deviations: left turn, right turn (lateral direction), and forward, backward (longitudinal direction). The formula is defined as:

$$d_{int}\left(\mathbf{P}_t, \mathbf{F}_t, n\right) = \frac{1}{L_O} \sum_{i=t+1}^{t+L_O} \left(p_i^n - s_i^n\right)^T \cdot \mathcal{G}\left(s_{i+1}^n, s_i^n\right) \quad (1)$$

where n denotes the target vehicle ID, p and s represent predicted and ground-truth vehicle locations respectively, and $\mathcal{G}$ is a direction-specific unit vector generator. The longitudinal direction is approximated as $s_{i+1}^n - s_i^n$.

Prediction Accuracy Degradation: This objective maximizes standard trajectory prediction error metrics. We focus on two commonly used measures: Average Displacement Error (ADE), which calculates the average root mean squared error between predicted and ground truth trajectories, and Final Displacement Error (FDE), which measures displacement error at the final predicted timestep.

The formulas are:

$$ADE = \frac{1}{L_O} \sum_{i=t+1}^{t+L_O} \|p_i^n - s_i^n\|_2 \tag{2}$$

$$FDE = \left\|p_{t+L_O}^n - s_{t+L_O}^n\right\|_2 \tag{3}$$

The combined attack objective metric is:

$$d_{err}\left(\mathbf{P}_t, \mathbf{F}_t, n\right) = \mathbb{I}_{ADE} \cdot ADE + \mathbb{I}_{FDE} \cdot FDE \tag{4}$$

where $\mathbb{I}_{ADE}$ and $\mathbb{I}_{FDE}$ are indicator variables set to 1 when targeting ADE and FDE respectively, and 0 otherwise.

B. Adversarial Criteria

To establish a unified framework for adversarial detection, we define an adversarial criterion based on selectively activated attack objective functions. Let $D = \{d_{int}, d_{err}\}$ denote the set of available attack objective functions, with corresponding threshold set $\Theta = \{\theta_{int}, \theta_{err}\}$, where each threshold $\theta_i \in \Theta$ represents the minimum activation level required for the corresponding attack objective $d_i \in D$.

At each evaluation instance, only one attack objective function is selected from D. Let $d_{active} \in D$ denote the currently selected attack objective with its corresponding threshold $\theta_{active} \in \Theta$. The adversarial criteria function is expressed as:

$$c(\mathbf{P}_t, \mathbf{F}_t, n) = \mathbb{I}\left(d_{active}(\mathbf{P}_t, \mathbf{F}_t, n) > \theta_{active}\right) \tag{5}$$

By incorporating the model relationship $\mathbf{P}_t = \Phi(X_t \cup \mathbf{H}_t, \mathbf{I}_t)$, we express the criteria in terms of input trajectory: $c(X_t, \mathbf{F}_t, n)$.

This adversarial criteria function serves as a binary decision output function that partitions the trajectory input space into adversarial and non-adversarial regions. The function outputs 1 when the attack objective functions exceed their corresponding thresholds, and 0 otherwise.

C. Adversarial Optimization Process

Our approach fundamentally differs from existing score-based attack methods [10]–[12]. Traditional score-based attacks utilize gradient information or stochastic exploration to generate adversarial trajectories through iterative optimization:

$$\min_{X_t^*} \zeta(X_t^* \in \mathbb{B}_{X,\gamma}) + \mathcal{S}(\Phi(X_t^* \cup \mathbf{H}_t, \mathbf{I}_t), \mathbf{F}_t, n) \tag{6}$$

Here, $\mathbb{B}_{X,\gamma}$ denotes a ℓ_p norm ball centered at X_t with radius γ. The term $\zeta(.)$ represents a physical constraint function: $\zeta(u) = 0$ if X_t^* lies within ball $\mathbb{B}_{X,\gamma}$, otherwise $\zeta(u) = \infty$. This constraint improves attack stealthiness and reduces detection probability. The term $\mathcal{S}(.)$ denotes the confidence score that guides adversarial trajectory generation.

Algorithm 1 Decision-based Attack Method

Input: Original trajectory X_t, historical states $\mathbf{H}_t$, contextual information $\mathbf{I}_t$, trajectory prediction model $\Phi(\cdot)$, adversarial criteria $c(\cdot)$, ground truth $\mathbf{F}_t$
Output: Adversarial trajectory X_t^* minimizing $\mathcal{D}(X_t, X_t^*)$

Initialize: $k \leftarrow 0$, $\delta \leftarrow 1.0$, $\epsilon \leftarrow 0.1$, max_iter $\leftarrow 1000$
Find initial adversarial point: Sample $X_t^{*(0)}$ such that $c(\Phi(X_t^{*(0)} \cup \mathbf{H}_t, \mathbf{I}_t), \mathbf{F}_t) = 1$
Forward initialization: Perform initial forward step towards X_t
while $k <$ max_iter **and** $\epsilon >$ tolerance **do**
  repeat
    Orthogonal direction: $\mathbf{d}_\perp \leftarrow_\perp (X_t - X_t^{*(k)})$
    Sample orthogonal step: $S_o^{(k)} \leftarrow \delta \cdot \mathbf{d}_\perp \cdot \mathcal{N}(0,1)$
    $X_{\text{temp}} \leftarrow X_t^{*(k)} + S_o^{(k)}$
    if $c(\Phi(X_{\text{temp}} \cup \mathbf{H}_t, \mathbf{I}_t), \mathbf{F}_t) = 1$ **then**
      break
    else
      $\delta \leftarrow \delta \times 0.95$
    end if
  until valid orthogonal step found
  repeat
    Forward direction: $\mathbf{d}_f \leftarrow \frac{X_t - (X_t^{*(k)} + S_o^{(k)})}{\|X_t - (X_t^{*(k)} + S_o^{(k)})\|}$
    Compute forward step: $S_f^{(k)} \leftarrow \epsilon \cdot \mathbf{d}_f$
    $X_{\text{candidate}} \leftarrow X_t^{*(k)} + S_o^{(k)} + S_f^{(k)}$
    if $c(\Phi(X_{\text{candidate}} \cup \mathbf{H}_t, \mathbf{I}_t), \mathbf{F}_t) = 1$ **then**
      break
    else
      $\epsilon \leftarrow \epsilon \times 0.9$
    end if
  until valid forward step found **or** $\epsilon < 10^{-6}$
  Update: $X_t^{*(k+1)} \leftarrow X_t^{*(k)} + S_o^{(k)} + S_f^{(k)}$
  $k \leftarrow k + 1$
  if $\mathcal{D}(X_t, X_t^{*(k)}) - \mathcal{D}(X_t, X_t^{*(k-1)}) < 10^{-8}$ **then**
    break
  end if
end while
return $X_t^* \leftarrow X_t^{*(k)}$

Score-based attacks primarily seek adversarial trajectories that minimize confidence scores while maintaining proximity within the ℓ_p norm ball.

Decision-Based Optimization: Our decision-based approach operates exclusively on binary decision outputs from adversarial criteria, eliminating gradient dependency. This framework does not require explicit specification of trajectory physical constraints during attack generation. While traditional methods must incorporate kinematic models, velocity limits, and steering constraints, our approach circumvents these requirements by focusing on decision boundary manipulation without modeling underlying physical dynamics. This reduces computational complexity and enhances transferability across different systems. Our optimization

objective is formulated as:

$$\min_{X_t^*} \mathcal{D}(X_t, X_t^*) + \zeta(\mathbb{I}(c(X_t^*, \mathbf{F}_t, n))) \tag{7}$$

The distance function $\mathcal{D}$ minimizes perceptible differences between adversarial and original trajectories. The constraint function $\zeta(.)$ performs rejection sampling: $\zeta(u) = 0$ if u is true, otherwise $\zeta(u) = \infty$. The indicator function $\mathbb{I}$ describes whether the adversarial criteria is satisfied. Essentially, attackers explore the adversarial region defined by the trajectory prediction model's input domain and adversarial criteria, seeking points guaranteed to be adversarial while remaining as close as possible to the original trajectory.

Boundary Walking Algorithm: We implement this optimization through a heuristic consisting of two key steps: orthogonal steps (S_o) and forward steps (S_f). Orthogonal steps move the adversarial trajectory away from the decision boundary while maintaining distance from the original trajectory, helping escape local optima. Forward steps move the adversarial trajectory closer to the original trajectory.

As shown in Algorithm 1, the algorithm begins by initializing an adversarial trajectory through random sampling near the original trajectory, rejecting non-adversarial samples. After initialization, a forward step brings the adversarial trajectory close to the decision boundary. The algorithm then alternates between orthogonal and forward steps along the adversarial boundary to find the closest adversarial trajectory to the original.

The orthogonal step direction is perpendicular to vector $X_t - X_t^*$. With sufficiently small step sizes, the orthogonal step approximately satisfies:

$$\mathcal{D}(X_t^* + S_o, X_t) \approx \mathcal{D}(X_t^*, X_t) \tag{8}$$

Forward steps move in the direction of vector $X_t - X_t^*$, satisfying:

$$\mathcal{D}(X_t^{*(k-1)} + \eta^{(k)}, X_t) < \mathcal{D}(X_t^{*(k-1)}, X_t) \tag{9}$$

Step sizes are dynamically adjusted according to local boundary geometry. For orthogonal step size δ, we maintain 50% of orthogonal steps within the adversarial region, adjusting δ accordingly. Forward step size ϵ is adjusted to approach the boundary as closely as possible while keeping total step η small enough to approximate the boundary as piecewise linear. As distance to the original trajectory decreases, the adversarial boundary becomes flatter and ϵ gradually decreases. The attack converges when ϵ reaches zero.

V. Experiments

A. Experiment Set-up

Models & Datasets : We conduct experiments on two widely-adopted trajectory prediction benchmarks: nuScenes [15] and Apolloscape [16]. Both datasets feature diverse driving scenarios with 2 Hz trajectory sampling. Following standard protocols, we set nuScenes parameters to $L_I = 4$ and $L_O = 12$, and Apolloscape parameters to $L_I = 6$ and $L_O = 6$.

We evaluate two representative trajectory prediction architectures: Trajectron++ [1] and Grip++ [2]. Both models are assessed on each dataset, yielding four base configurations. Additionally, we include Trajectron++(m), which leverages semantic map information available in nuScenes, resulting in five total experimental settings. For multimodal Trajectron++ predictions, we select the mode with highest probability. Each evaluation uses 100 randomly sampled scenarios per dataset.

Attack Methods & Evaluation: We compare against two black-box baselines: PSO-based trajectory attack [10] and Simple Black-box Attack (SBA) [17] from image classification. Given inherent differences in optimization formulations, we establish a unified evaluation framework by first applying our DTP-Attack to determine perturbation bounds, then constraining all baselines within this space to ensure comparable search domains.

We evaluate all methods across two attack objectives with corresponding metrics. For intention misclassification, we measure directional offset distribution (Left, Right, Front, Rear) and Attack Success Rate (ASR) for lane deviation under physical constraints. For prediction accuracy degradation, we assess standard trajectory metrics including Average Displacement Error (ADE), Final Displacement Error (FDE), Miss Rate (MR), Off-Road Rate (ORR), and perturbation magnitude via Mean Squared Error (MSE). Each method operates under a fixed computational budget of 1,000 model queries per scenario.

Implementation details: Our DTP-Attack method employs a boundary walking algorithm with dynamic step size adjustment. We initialize orthogonal step size $\delta = 1.0$ and forward step size $\epsilon = 0.1$, with adjustment factors of 0.95 and 0.9 respectively to maintain algorithmic convergence. The algorithm terminates when $\epsilon < 10^{-6}$ or maximum iterations (1,000) are reached.

Adversarial criteria thresholds are configured as follows: For prediction accuracy degradation, $\theta_{err}^{\text{ADE}} = 7.5\,\text{m}, 3.5\,\text{m}$ and $\theta_{err}^{\text{FDE}} = 17.5\,\text{m}, 7.5\,\text{m}$ for nuScenes and Apolloscape respectively. For intention misclassification, directional thresholds are set to $\theta_{int}(\text{lateral}) = 2\,\text{m}$ and $\theta_{int}(\text{longitudinal}) = 3\,\text{m}$ across both datasets. Attack success requires satisfying both adversarial criteria and physical constraints defined by the vehicle dynamics model from [18].

B. Main Results

Our DTP-Attack demonstrates consistent effectiveness across all evaluated models and datasets, successfully achieving both primary attack objectives: intention misclassification and prediction accuracy degradation. The following results reveal fundamental vulnerabilities in current trajectory prediction systems, with potentially severe implications for autonomous vehicle safety.

Intention Misclassification: Our attacks successfully induce substantial directional biases in trajectory predictions across all configurations, as shown in Table I. Lateral attacks increase prediction deviations from near-zero baselines to over 2 m, while longitudinal attacks amplify errors from

TABLE I
DTP-ATTACK PERFORMANCE ON INTENTION MISCLASSIFICATION: DIRECTIONAL BIAS INDUCTION AND ATTACK SUCCESS RATES

Model	Dataset	Left(m)	Right(m)	Front(m)	Rear(m)	ASR(%)
		Normal/Attack	Normal/Attack	Normal/Attack	Normal/Attack	Attack
Grip++	nuScenes	0.231/2.13	-0.231/2.21	-1.02/3.01	1.02/3.13	81
	Apolloscape	-0.0131/2.03	0.0131/2.18	-0.0147/3.07	0.0147/3.06	76
Trajectron++	nuScenes	-0.301/2.03	0.301/2.25	-1.31/3.13	1.31/3.21	78
	Apolloscape	0.0102/2.33	-0.0102/2.15	-0.143/3.01	0.143/3.11	80
Trajectron++(m)	nuScenes	-0.105/2.01	0.105/2.03	-0.515/3.11	0.515/3.05	41

TABLE II
DTP-ATTACK PERFORMANCE ON PREDICTION ACCURACY DEGRADATION: ERROR METRICS AND PERTURBATION ANALYSIS

Model	Dataset	ADE(m)	FDE(m)	MR(%)	ORR(%)	MSE(m)
		Normal/Attack	Normal/Attack	Normal/Attack	Normal/Attack	Attack
Grip++	nuScenes	4.34/8.27	10.3/18.95	16/45	3.6/15.3	0.21
	Apolloscape	1.66/4.68	3.18/8.09	8/61	2.2/28.1	0.29
Trajectron++	nuScenes	3.84/8.15	11.2/19.4	14/42	1.6/12.9	0.12
	Apolloscape	1.00/4.23	2.24/8.71	3/55	1.5/24.4	0.36
Trajectron++(m)	nuScenes	1.99/4.05	5.37/10.24	2/24	0.3/5.9	0.45

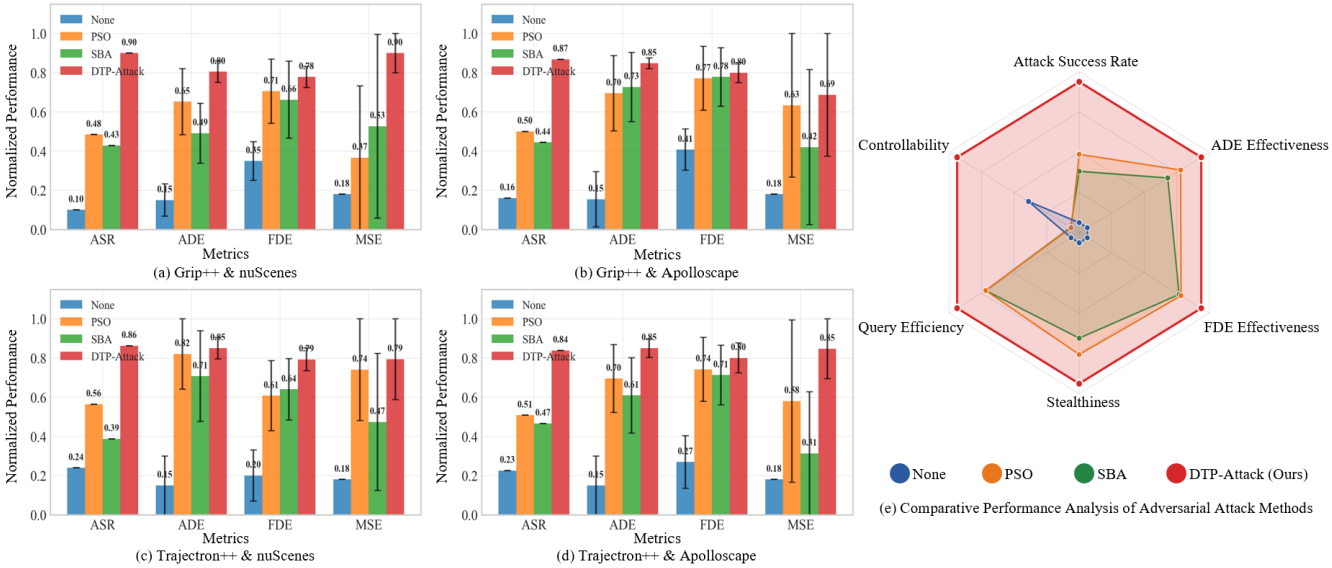

Fig. 4: Comparative analysis of decision-based vs. score-based adversarial attacks on trajectory prediction models.

approximately $1\,\mathrm{m}$ to over $3\,\mathrm{m}$. Importantly, both attack types maintain prediction errors precisely near the adversarial thresholds, demonstrating excellent controllability.

These magnitudes carry significant practical implications. Since urban lane widths typically measure below $4\,\mathrm{m}$, $2\,\mathrm{m}$ lateral offset causes vehicles to appear as though they will deviate from their designated lanes. Similarly, $3\,\mathrm{m}$ longitudinal offsets create predictions of significant acceleration or deceleration behaviors that fundamentally alter expected vehicle dynamics.

Attack success rates range from 41% to 81% across different model-dataset combinations. Models without semantic map information such as Grip++ and standard Trajectron++ achieve consistently higher success rates than Trajectron++ with map integration, indicating that additional contextual information provides some defensive benefit. However, even this enhanced model remains substantially vulnerable, with success rates exceeding 40%. The consistent effectiveness across diverse architectures reveals fundamental vulnerabilities in current trajectory prediction systems, particularly concerning given that our decision-based approach operates without gradient information or detailed model knowledge, conditions that are typical of real-world attack scenarios.

Prediction Accuracy Degradation: Our attacks cause dramatic accuracy degradation in standard trajectory metrics, as demonstrated in Table II. Average displacement error increases by factors of $1.9\times$ to $4.2\times$, while final displacement error shows similar deterioration with increases ranging from $1.8\times$ to $3.9\times$. These results demonstrate that subtle adversarial perturbations can severely compromise prediction quality across different model architectures. Miss rates provide particularly striking evidence of attack effectiveness. For Apolloscape, miss rates increase from 8% to 61% for Grip++ and from 3% to 55% for Trajectron++, indicating that most predictions under attack exceed acceptable error thresholds. Off-road rates similarly increase substantially, with predictions frequently placing vehicles in physically impossible locations.

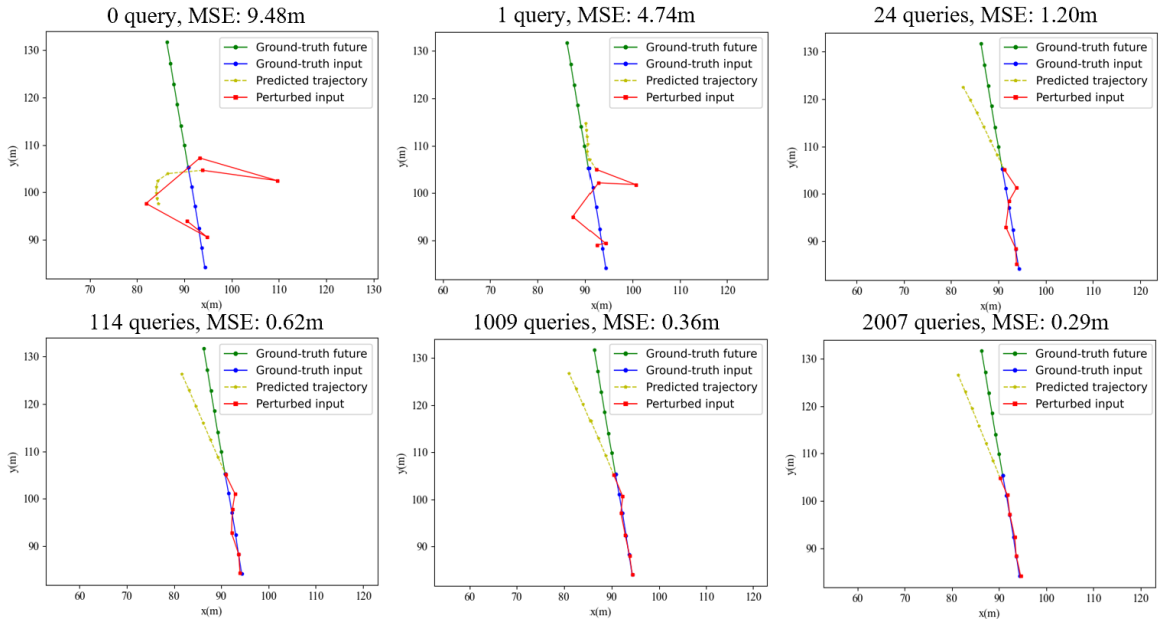

Fig. 5: Progressive trajectory refinement in DTP-Attack boundary walking algorithm.

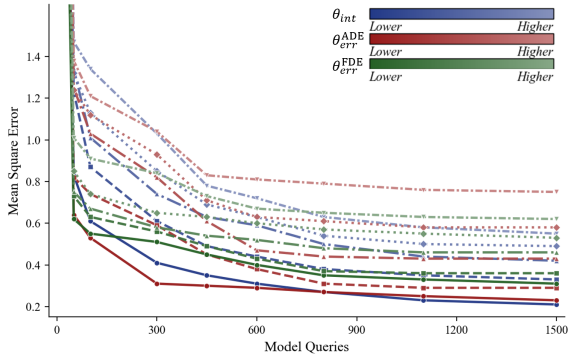

Fig. 6: Convergence analysis of DTP-Attack under varying adversarial thresholds.

Despite these significant performance degradations, perturbation magnitudes remain remarkably small, with mean squared error between original and adversarial trajectories ranging from $0.12\,\mathrm{m}$ to $0.45\,\mathrm{m}$. This demonstrates that our method achieves substantial prediction disruption through barely perceptible trajectory modifications that maintain physical plausibility and reduce detection likelihood. These findings highlight the urgent need for robust defense mechanisms in safety-critical autonomous driving applications, where compromised trajectory predictions could lead to catastrophic consequences.

C. Comparative Analysis

Performance Superiority: Our decision-based approach consistently outperforms existing black-box methods across all normalized evaluation metrics, as shown in Fig. 4. The comprehensive comparison reveals substantial improvements in attack effectiveness, with DTP-Attack achieving superior performance on both intention misclassification and prediction accuracy degradation objectives. This consistent advantage holds across diverse model architectures and datasets, demonstrating robust generalizability.

A critical advantage lies in attack controllability, evidenced by the exceptionally narrow confidence intervals across 100 evaluation scenarios. While baseline methods exhibit high variance and unpredictable behavior, DTP-Attack maintains consistently stable output characteristics regardless of scenario initialization or environmental variations. This reliability enables adversaries to execute targeted attacks with predictable outcomes, allowing strategic attack planning where specific prediction failures can be induced with high confidence. The tight error bounds indicate that our boundary walking algorithm reliably converges to near-optimal adversarial trajectories, making it particularly effective for real-world deployment.

Methodological Advantages: The performance gains stem from fundamentally different constraint handling strategies. Score-based methods incorporate fixed physical constraints to optimize attack stealthiness, explicitly limiting vehicle dynamics and kinematic feasibility. However, these rigid constraints severely restrict the adversarial search space and trap optimization algorithms in local optima. Moreover, fixed constraints cannot adapt across diverse driving scenarios, such as highway maneuvers require different kinematic bounds than urban intersections, limiting method generalizability. Our decision-based approach circumvents these limitations by avoiding fixed physical constraints entirely. Instead of explicitly enforcing kinematic bounds, we minimize the distance between adversarial and original trajectories, allowing physical plausibility to emerge naturally from proximity preservation. This creates a significantly smoother search landscape without the discontinuous constraint boundaries that cause convergence failures in baseline methods. The resulting optimization surface enables our boundary walking

algorithm to navigate complex adversarial regions more effectively, consistently identifying near-optimal solutions and explaining the significantly superior attack success rates.

D. Hyperparameter Analysis and Case Study

Fig. 5 demonstrates iterative attack evolution from initial random sampling (MSE: $9.48\,\mathrm{m}$) to optimized results (MSE: $0.29\,\mathrm{m}$ after 2007 queries) under Trajectron++&Apolloscape. The boundary walking algorithm achieves rapid initial improvement, with MSE dropping to $4.74\,\mathrm{m}$ after one query and $1.20\,\mathrm{m}$ after 24 queries, representing 50% and 87% reductions respectively.

The optimized attack generates prediction errors exceeding physical lane boundaries. In a scenario where an autonomous vehicle drives in the left adjacent lane, false left-turn predictions would trigger collision avoidance responses including emergency braking. The minimal perturbation magnitude demonstrates the feasibility of generating imperceptible attacks that influence trajectory prediction outputs.

We evaluate attack performance across varying threshold difficulties: $\theta_{int} \in \{2,3,4,5,6\}$, $\theta_{err}^{\mathrm{ADE}} \in \{4,5,6,7,8\}$, and $\theta_{err}^{\mathrm{FDE}} \in \{17,18,19,20,21\}$. Fig. 6 shows convergence exhibits a consistent two phase pattern: rapid MSE reduction within 300 queries followed by asymptotic stabilization. Lower adversarial thresholds produce smaller final MSE values, confirming that relaxed attack criteria enable closer approximation to original trajectories. This validates our optimization framework and demonstrates the trade-off between attack success and perturbation stealthiness.

VI. Conclusions

We present DTP-Attack, the decision-based black-box adversarial attack framework for trajectory prediction systems in autonomous driving. Unlike existing methods requiring gradient information or model internals, our approach operates exclusively on binary decision outputs, making it applicable to real-world scenarios. Our novel boundary walking algorithm navigates adversarial regions without fixed physical constraints, naturally maintaining trajectory realism through proximity preservation. Extensive evaluation on nuScenes and Apolloscape datasets across state-of-the-art models demonstrates superior performance, achieving $41-81\%$ attack success rates for intention misclassification with perturbations below $0.45\,\mathrm{m}$, and increasing prediction errors by $1.9{-}4.2\times$ for accuracy degradation. These minimal perturbations can induce $2\,\mathrm{m}$ lateral and $3\,\mathrm{m}$ longitudinal prediction deviations, potentially causing autonomous vehicles to misinterpret critical driving maneuvers. Our method consistently outperforms existing black-box approaches while revealing fundamental vulnerabilities in current trajectory prediction systems. These findings highlight urgent needs for robust defenses in safety-critical autonomous driving applications. Future work should focus on developing defensive mechanisms that can detect and mitigate such adversarial attacks while maintaining prediction accuracy under normal conditions.

References

[1] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020.

[2] Xin Li, Xiaowen Ying, and Mooi Choo Chuah. Grip: Graph-based interaction-aware trajectory prediction. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3960–3966. IEEE, 2019.

[3] Ke He, Dan Dongseong Kim, and Muhammad Rizwan Asghar. Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 25(1):538–566, 2023.

[4] Chiho Choi, Srikanth Malla, Abhishek Patil, and Joon Hee Choi. Drogon: A trajectory prediction model based on intention-conditioned behavior reasoning. In *Conference on Robot Learning*, pages 49–63. PMLR, 2021.

[5] Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. Adversarial machine learning in image classification: A survey toward the defender's perspective. *ACM Computing Surveys (CSUR)*, 55(1):1–38, 2021.

[6] Yanfei Zhu, Yaochi Zhao, Zhuhua Hu, Tan Luo, and Like He. A review of black-box adversarial attacks on image classification. *Neurocomputing*, 610:128512, 2024.

[7] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3560–3566. IEEE, 2023.

[8] K Naveen Kumar, Chalavadi Vishnu, Reshmi Mitra, and C Krishna Mohan. Black-box adversarial attacks in autonomous vehicle technology. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE, 2020.

[9] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018.

[10] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2022.

[11] Yulong Cao, Chaowei Xiao, Anima Anandkumar, Danfei Xu, and Marco Pavone. Advdo: Realistic adversarial attacks for trajectory prediction. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022.

[12] Jiping Fan, Zhenpo Wang, and Guoqiang Li. Adversarial attack on trajectory prediction for autonomous vehicles with generative adversarial networks. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1026–1031, 2024.

[13] Kaiyuan Tan, Jun Wang, and Yiannis Kantaros. Targeted adversarial attacks against neural network trajectory predictors. In *Learning for Dynamics and Control Conference*, pages 431–444. PMLR, 2023.

[14] Wenjie Xu, Yuning Jiang, Bratislav Svetozarevic, and Colin Jones. Constrained efficient global optimization of expensive black-box functions. In *International Conference on Machine Learning*, pages 38485–38498. PMLR, 2023.

[15] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[16] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018.

[17] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International conference on machine learning*, pages 2484–2493. PMLR, 2019.

[18] Huilin Yin, Jiaxiang Li, Pengju Zhen, and Jun Yan. Sa-attack: Speed-adaptive stealthy adversarial attack on trajectory prediction. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1772–1778, 2024.