EL-Attack: Explicit and Latent Space Hybrid Optimization based General and Effective Attack for Autonomous Driving Trajectory Prediction

Xuesong Bai[1,2], Changhang Tian[2], Wei Xia[2], Zhenshu Ma[1], Haiyang Yu[1,2], Yilong Ren[1,2,†]
[1]School of Transportation Science and Engineering, Beihang University, Beijing, China
[2]State Key Laboratory of Intelligent Transportation Systems, Beijing, 102206, Beijing, P.R.China
{xs_bai,mzs0822,hyyu,yilongren}@buaa.edu.cn,{tianchanghang,xiawei1692}@163.com

Abstract

Accurate prediction of nearby road actors' future trajectories is crucial for autonomous vehicles. With the development of foundational models, autonomous driving trajectory prediction has seen significant improvements. However, these neural network-based methods are vulnerable to adversarial attacks, challenging the reliability and safety of predictions. Previous attack methods focused on constraint spaces and objective functions, generating adversarial trajectories via perturbations in the explicit space followed by further optimization. However, these methods overlook the threat model's potential. To fully leverage the model's optimization, we propose a novel adversarial attack method, EL-Attack, which emphasizes multi-space collaborative optimization in both explicit and latent spaces. The framework first uses a spatio-temporal attention module to extract semantic representations of the trajectory's spatiotemporal context, then builds a threat model based on an adversarial autoencoder. In the explicit space, we introduce an interactive risk field based on the autonomous vehicle's drivable area to guide the target vehicle's trajectory. In the latent space, we apply semantic-level perturbations on latent vectors and regularize them, enhancing attack targeting and stealthiness. We conducted experiments and evaluations on the Argoverse dataset and a virtual-real testing platform. In terms of effectiveness, compared to the best - performing baseline, our method improves the attack success rate by 4.0% and 15.2% on the VectorNet and TNT models respectively. we also tested in scenarios such as straight roads, curves, and intersections for real-world validation and transferability.

1. Introduction

Autonomous driving trajectory prediction plays a critical role in enabling safe motion planning for automated vehicles [8]. Recent advances in deep neural networks have significantly improved prediction accuracy [9], yet these models remain vulnerable to adversarial perturbations [14]. Such attacks can cause severe prediction deviations [3], underscoring the need for robust trajectory prediction frameworks [1].

An ideal adversarial trajectory must balance stealthiness and aggressiveness. While stealthiness ensures compliance with physical constraints, aggressiveness induces safety-critical reactions. Prior methods often focus on complex constraints but overlook threat model sophistication [14]. Explicit-space perturbations risk violating traffic rules, requiring post-smoothing that undermines attack realism. Moreover, traditional objectives like Euclidean distance fail to capture directional context [11].

To address these challenges, we propose EL-Attack, a novel framework integrating multi-space optimization. Our spatiotemporal attention module captures real-world rules [9], while an adversarial autoencoder (AAE) forms the threat model [1]. In explicit space, an interaction risk field guides trajectories into risk area [8], while latent-space perturbations leverage semantic disentanglement [13]. This dual-space approach avoids heuristic constraints and enhances both stealthiness and attack precision [4]. Our contributions are as follows:

- We propose EL-Attack, introducing a collaborative optimization framework that utilizes AAE-based threat modeling across explicit and latent spaces.
- We design a spatiotemporal attention perception module to capture real-world trajectory semantics and contextual interactions.
- We present the concept of an interaction risk field to enhance attack effectiveness by guiding trajectories into risk regions.
- We apply semantic-level perturbations in the latent space under different orthogonal modes, improving attack stealthiness and precision.

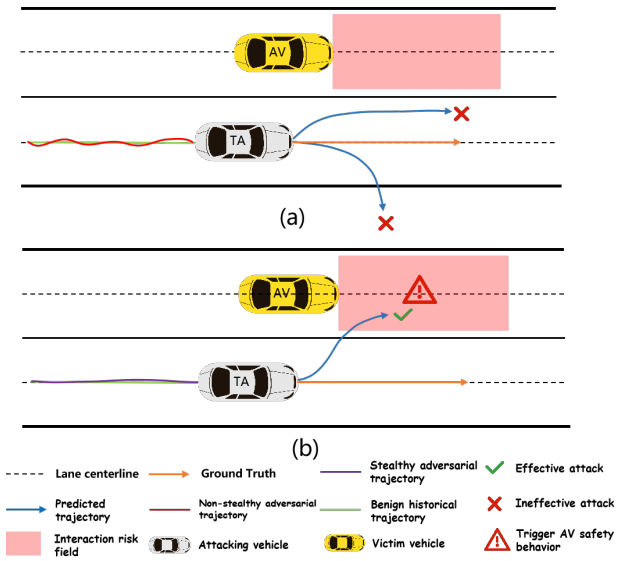

Figure 1. **two examples of adversarial attack scenarios.** (a) presents two major drawbacks of existing methods: Firstly, the adversarial trajectory with poor stealthiness (the red trajectory) is likely to be identified as dangerous driving behavior by the system. Secondly, the overly simplistic optimization objective leads to ineffective attacks (the blue trajectory in (a)). (b) demonstrates the advantages of our method. The adversarial trajectory conforms to real-world rules and has high stealthiness (the purple trajectory), and it can effectively prompt the victim vehicle to take safety responses (the blue trajectory in (b)).

2. Related Works

Trajectory Prediction Model. With the rapid development of deep neural networks, deep learning-based trajectory prediction technology has become a research hotspot in the field of autonomous driving. Such methods use the historical spatial coordinates of agents as the basic input, and combine multi-dimensional information such as map semantic features, physical motion constraints, and interaction relationships between agents, significantly improving the prediction accuracy.

Among them, VectorNet[6] proposed a hierarchical graph neural network architecture. It aggregates the spatial information of vectorized map elements and trajectories through local subgraphs, then models high-order semantic relationships using a global interaction graph, and introduces a node completion task to enhance context understanding. It achieves higher performance with fewer model parameters on the Argoverse dataset. The TNT[16] framework, innovatively, decomposes trajectory prediction into three stages: discrete target prediction, conditional trajectory generation, and trajectory scoring and selection. By explicitly modeling targets, it captures multi-modal intentions, avoiding the interpretability issues and sampling dependence of traditional latent variable models, and outperforms previous methods in both vehicle and pedestrian trajectory prediction tasks.

Although these models perform excellently on standard test sets, existing research shows that when facing adversarial attacks, the prediction accuracy of these models may still significantly decline due to input perturbations. Therefore, this study selects these two representative architectures for a systematic evaluation, aiming to reveal their vulnerabilities in adversarial environments.

Adversarial Attacks in Trajectory Prediction. Deep learning models are generally vulnerable to adversarial attacks. In the field of autonomous vehicles, a lot of research has focused on the impact of attacks on the perception module. In recent years, adversarial attacks on trajectory prediction models have received extensive attention.

Among them, Zhang et al. [15] were the first to propose a search-based attack method. By imposing hard constraints such as speed and acceleration, they verified the robustness of the model and clarified the threat of adversarial attacks to prediction safety. Cao et al. [4] designed a "deterministic attack" for probabilistic generative models. By replacing random sampling with maximum likelihood samples, they eliminated the randomness of the attack and significantly improved the attack effect on conditional Gaussian prior models. Cao et al. [3] further constructed a two-stage framework. First, they densified sparse trajectory points to calculate dynamic parameters, and then used the Projected Gradient Descent (PGD) method to generate adversarial trajectories that conform to kinematic constraints, enhancing the physical rationality of the attack. Tan et al. [12] proposed a targeted attack method named TA4TP. By quantifying prediction bias through an objective function, it can accurately mislead specific driving behaviors such as lane changing and steering.

However, existing attack methods often focus on the design of the constraint space and objective function of the threat model, while ignoring the potential of the threat model itself. Our method constructs a threat model based on the adversarial autoencoder architecture. It maps the trajectory features containing real-world rules and lane environment information to the latent space. By implementing semantic-level perturbations, the generated adversarial trajectories inherently possess both aggressiveness and real-world stealthiness.

3. Problem Formulation

Trajectory Prediction: The trajectory prediction task aims to predict the future trajectories of various agents in a scene based on historical trajectory data and relevant environmental information. The specific definition is as follows:

Given N agents in a scene, the H historical trajectories of each agent i at time t are $\mathcal{X}_i = \{s_i^{t-H+1}, \ldots, s_i^t\}$, and

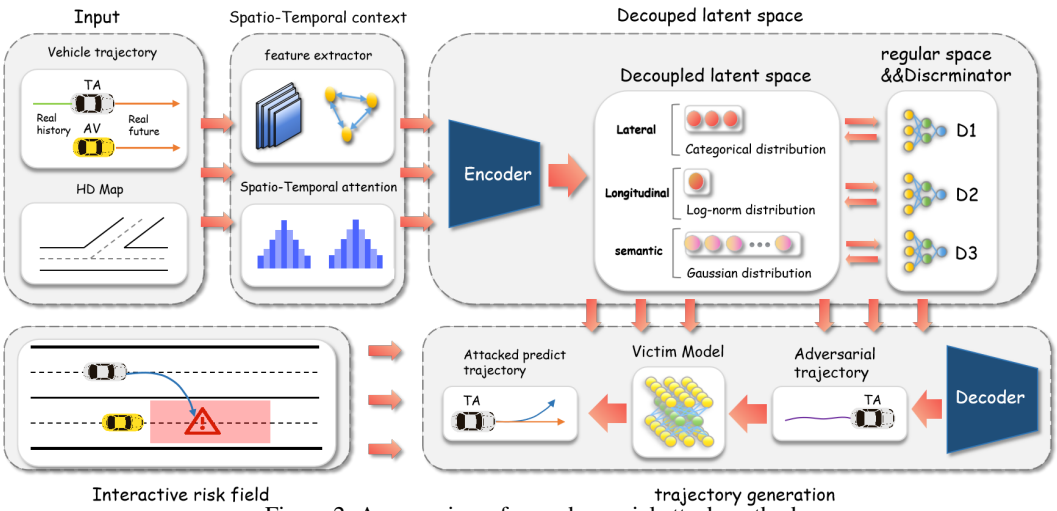

Figure 2. An overview of our adversarial attack method.

the ground truth of its future trajectory for the subsequent L time steps is $\mathcal{Y}_i = \{s_i^{t+1}, \ldots, s_i^{t+L}\}$. Among them, s_i^τ represents the state of agent i at time τ, including information such as position, velocity, and heading. At the same time, the environmental context information $\mathcal{C}$ of the scene (such as high-definition maps, etc.) is provided. The goal of the trajectory prediction model is to predict the state sequence of each agent i for the next L time steps, that is, the future trajectory $\hat{\mathcal{Y}}_i = \{\hat{s}_i^{t+1}, \ldots, \hat{s}_i^{t+L}\}$.

Mathematically, the trajectory prediction model can be expressed as a function $\mathcal{F} : \mathcal{X} \times \mathcal{C} \rightarrow \hat{\mathcal{Y}}$, where $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_N\}$, $\mathcal{C}$ is the context information of the scene, and $\hat{\mathcal{Y}} = \{\hat{\mathcal{Y}}_1, \ldots, \hat{\mathcal{Y}}_N\}$ is the predicted future trajectory. The model learns the mapping relationship between historical trajectories and context information, and minimizes the error between the predicted trajectory and the true future trajectory, that is, $\min_{\mathcal{F}} \mathcal{L}(\mathcal{Y}, \hat{\mathcal{Y}})$.

Attack Model: The threat model aims to mislead the trajectory prediction model by applying a small perturbation δ to the input data, generating adversarial samples. In a white-box attack, the attacker uses gradient information to optimize the perturbation, while in a black-box attack, adversarial samples are generated through model queries. The attacker's objective is to maximize the prediction error, expressed as:

$$\max_{\delta} \mathcal{L}(\mathcal{F}(\mathcal{X} + \delta, \mathcal{C}), \mathcal{Y}) \tag{1}$$

where $\mathcal{L}$ is the loss function that measures the error between the predicted trajectory $\hat{\mathcal{Y}}$ and the true future trajectory $\mathcal{Y}$, and δ is the perturbation applied by the attacker, which satisfies the constraint conditions (such as the L_p norm constraint) to ensure the stealthiness of the adversarial samples.

4. Method

4.1. overview

Adversarial trajectory generation must balance stealthiness and aggressiveness. However, existing methods relying solely on explicit space perturbations struggle to achieve this balance and fail to fully exploit the potential of the threat model. To address these challenges, we redefine the adversarial trajectory generation task as a constrained vehicle trajectory reconstruction problem and propose EL-Attack, a novel method that emphasizes multi-space collaborative optimization across explicit and latent spaces. Our framework takes as input $\{\mathcal{X}_{TA}, \mathcal{Y}_{AV}, \mathcal{Y}_{TA}, \mathcal{C}\}$, representing the TA vehicle's historical trajectory, the ground truth future trajectories of the AV and TA vehicles, and environmental context information, respectively. The spatio-temporal perception module extracts trajectory features using a feature extractor [10] and attention mechanism, capturing lane and temporal context. In the explicit space, we introduce an interaction risk field that models the AV's drivable area, guiding the TA's future trajectory to maximize its disruptive effect. In the latent space, we employ an adversarial autoencoder (AAE) to map trajectory features to a disentangled latent space. Through semantic-level perturbations on orthogonal latent vectors and distribution regularization, EL-Attack enhances both the targeting precision and stealthiness of adversarial trajectories. The perturbed latent vectors are then decoded back to generate realistic adversarial trajectories.

Compared to traditional attacks, EL-Attack offers distinct advantages: (1) Enhanced Stealthiness: By capturing spatio-temporal correlations, the generated trajectories align with real-world traffic conditions, reducing detectability. (2) Improved Targeting: Semantic-level perturbations in the latent space enable precise and effective attacks. (3) Re-

alism Maintenance: The AAE's latent space regularization ensures physical plausibility and rule compliance. (4) Attack Effectiveness: The interaction risk field misguides the AV's prediction module, increasing the attack's success rate and impact on autonomous driving systems.

4.2. Interactive Perception of Spatio-temporal Context

Compared to image reconstruction tasks, adversarial trajectory generation for attacking autonomous driving prediction models is more complex. Vehicle trajectories must maintain spatio-temporal coherence, follow latent logical rules, and adapt to road environments, making single-feature modeling insufficient. Additionally, identifying which historical trajectory features significantly impact future predictions is crucial. To address this, we introduce a spatio-temporal interactive perception module that provides high-quality trajectory features by integrating environmental and temporal context. This module consists of a feature extractor f_e [10], which employs one-dimensional dilated convolution to encode trajectory embeddings and a graph network to model vehicle-lane interactions. Furthermore, a spatio-temporal perception attention mechanism emphasizes the influence of historical trajectories on future predictions, enhancing the threat model's ability to capture complex spatio-temporal relationships. The mathematical formulation of this module is as follows:

$$\begin{aligned} X_{ST} = \alpha \cdot \mathrm{Attn}[f_e(\mathcal{X}_{TA}, \mathcal{C}), \mathcal{Y}_{TA}, \mathcal{Y}_{TA}] + \\ (1-\alpha) \cdot \mathrm{Attn}[f_e(\mathcal{X}_{TA}, \mathcal{C}), \mathcal{Y}_{AV}, \mathcal{Y}_{AV}] \end{aligned} \tag{2}$$

Among them, α is a weight coefficient, $f_e(\mathcal{X}_{TA}, \mathcal{C})$ is the interactive representation containing lane context output by the feature extractor, X_{ST} is the output after spatio-temporal interaction, $\mathrm{Attn}[a, b, c]$ is the spatio-temporal attention, where a is the query, b is the key, and c is the value.

4.3. Enhanced Optimization of Interaction Risk in the Explicit Space

Traditional adversarial attack methods often rely on Euclidean distance or lateral deviation as optimization objectives, which fail to accurately ensure attack effectiveness. To address this, we introduce the concept of an interaction risk field Ω, representing the AV's safe and drivable area. As can be seen from the Figure 1, When other vehicles' trajectories enter this field, the AV's drivable space is compressed, triggering safety responses like braking or lane changing. Our objective is to generate adversarial trajectories that mislead the AV into falsely perceiving a threat, resulting in erroneous decisions. To simplify modeling, we designate a point on the lane's centerline ahead of the AV as the risk centroid, representing the field. As shown in the Figure 2, our threat model uses the spatio-temporally interacted vehicle trajectory X_{ST} as input to generate an adversarial trajectory $\mathcal{X}'_{TA}$. This, in turn, produces the crafted prediction trajectory $\hat{\mathcal{Y}}_{TA}$, guided by explicit space optimization objectives:

$$\mathcal{L}_{reg} = \|\mathcal{X}_{TA} - \mathcal{X}'_{TA}\|^2 \tag{3}$$

$$\mathcal{L}_{att} = \beta_1 \cdot \|\mathcal{Y}_{TA} - \hat{\mathcal{Y}}_{TA}\|^2 + \beta_2 \cdot \frac{1}{L} \sum_{t=1}^{L} \omega_t \|r_t - \hat{y}_t\|^2 \tag{4}$$

Among them, $\mathcal{L}_{reg}$ represents the regression loss for trajectory reconstruction, while $\mathcal{L}_{att}$ is the attack loss. The core function of $\mathcal{L}_{att}$ is to ensure the effectiveness of the attack. β serves as a weight coefficient, and L denotes the total number of future time steps. ω_t is the weight coefficient that decays gradually with the time step t. This design aims to guide the trajectory of the TA vehicle in the relatively near future into the risk field. Here, r and $\hat{y}$ correspond to the risk centroid and the predicted trajectory point at the current time step, respectively.

Through such a design, the interaction risk field Ω is fully utilized in the explicit space, which enhances the effect of adversarial attacks and optimizes the overall attack strategy.

4.4. Adversarial Guidance of Latent Vectors in the Latent Space

To effectively process vehicle trajectory features containing rich spatio-temporal context and complex real-world rules, we construct a threat model based on an Adversarial Autoencoder (AAE) architecture and introduce the concept of the latent space. This architecture consists of an encoder and a decoder: the encoder maps high-dimensional semantic features to a low-dimensional latent space and decouples them into latent vector features in different orthogonal modes. The longitudinal feature z_{lon} is represented as a one-dimensional vector that models the time interval of the vehicle passing a given point, reflecting dynamic interactions in the longitudinal direction, and follows a log-normal distribution. The lateral feature z_{lat} is a three-dimensional vector modeling vehicle steering intentions for going straight, turning left/leftward, and turning right/rightward, following a categorical distribution. Remaining semantic features that are challenging to model explicitly are represented as z_{gauss}, which follows a Gaussian distribution. The decoder then maps these decoupled latent vectors back to the explicit space to generate adversarial trajectories.

Our key innovation lies in performing semantic-level perturbations on the latent vectors within different orthogonal modes, thereby precisely misleading the AV's trajectory

prediction module. For example, perturbing the longitudinal features can interfere with the prediction module's estimation of the TA vehicle's speed and following distance. The perturbation process is formally expressed as follows:

$$\begin{aligned} z\prime_{lon} &= z_{lon} + \epsilon_{log}, \\ z\prime_{guass} &= z_{guass} + \epsilon_{guass}, \\ z\prime_{lat} &= z_{lat} + \epsilon_{cat} \end{aligned} \tag{5}$$

Among them, ϵ_{log} and ϵ_{guass} are the noise sampled from the Gaussian space for corresponding dimensions, which are respectively used to perturb the longitudinal features and the remaining features of the latent space. Innovatively, we scale the perturbation standard deviation of the noise based on the current speed of the TA vehicle and its relative distance to the AV vehicle. Additionally, we perform threshold truncation on the perturbed latent variables to prevent excessive perturbation. For ϵ_{cat}, we perturb the driving intention of the TA vehicle according to the azimuth of the TA relative to the AV risk centroid, so as to increase the probability that the TA vehicle drives towards the AV risk area.

The discriminator is used to identify and regularize the perturbed latent vectors to a specific distribution, ensuring that the generated adversarial trajectories conform to the physical laws of the real world and maintain a high level of stealthiness. Our optimization definition in the latent space is as follows:

$$\mathcal{L}_E = \frac{1}{m}\sum_{i=1}^{m} \log(1 - D_i(E(X_{ST}))), \tag{6}$$

$$\mathcal{L}_{D_i} = \log D_i(s_i) + \log(1 - D_i(E(X_{ST}))), \tag{7}$$

where m is the number of decoupled latent vectors, X_{ST} is the Spatio-Temporal trajectory feature, s is a sample from the distribution corresponding to a certain latent vector, and E and D represent the encoder and the discriminator, respectively.

5. Experiment

In the experimental section, we primarily address three research questions: the enhanced effectiveness of EL-Attack compared to baseline methods, the physical realism of the adversarial trajectories generated by this approach, and the performance of these adversarial examples on real autonomous vehicles. We provide a detailed account of the experimental setup and results, along with a corresponding analysis of these results.

5.1. Training Setup

Our framework is based on the adversarial autoencoder architecture. During the training process, the Argoverse dataset is used to train the parameters, and the VectorNet model is utilized to predict the future trajectories of the TA vehicles after being attacked. It is worth noting that the optimization objectives in the explicit space are designed to update the entire framework, while those in the latent space are only used to update the threat model. In the inference stage, by leveraging the pre-trained weights, this method can not only be applied to datasets other than the training set but also launch effective attacks on other prediction models.

5.1.1. Victim Model

The victim models include VectorNet, TNT, and Traj-LLM[7], all of which have been pre-trained on the Argoverse dataset. VectorNet and TNT each represent significant paradigms in trajectory prediction: graph-based interaction modeling and goal-driven prediction generation, respectively. VectorNet provides a robust tool for trajectory prediction due to its efficient geometric modeling and complex interaction comprehension capabilities. Meanwhile, TNT enhances the accuracy and reliability of trajectory prediction through a goal-driven multimodal prediction strategy. Our previous work, Traj-LLM, leverages the powerful semantic understanding capabilities of large language models, incorporating sparse context encoding and lane-aware probabilistic learning, thereby enhancing scene awareness and interaction modeling in trajectory prediction. Our baseline algorithms include Search[15], Search*[3], and SA-Attack[14], all of which are validated trajectory prediction attack methods. In accordance with the characteristics of the victim models, we preprocess the inputs of the baseline algorithms in our approach.

5.1.2. Evaluation Metrics

To comprehensively assess the effects and characteristics of adversarial attack methods on trajectory prediction models, this study defines three key evaluation metrics[5]: Success Rate (SR), Violation Rate (VR), and Dangerous Appearance Rate (DAR), facilitating systematic evaluation from the aspects of attack success rate, trajectory physical feasibility, and stealthiness. Additionally, we evaluate the accuracy and safety of predicted trajectories from victim models using Average Displacement Error (ADE), Final Displacement Error (FDE), and Off-road Rate (ORR). Beyond the assessment of effectiveness, we utilize Dynamic Time Warping (DTW) to evaluate the realism of the crafted historical trajectories in the physical world.

5.2. Results and Analysis of Adversarial Attack Experiments

5.2.1. Attack Effectiveness

To evaluate attack effectiveness, we conducted comparative experiments on robust trajectory prediction models, VectorNet and TNT, as well as the latest algorithm Traj-LLM. The results, summarized in Table 1, show that our method

Table 1. **Comparative Experiments on Adversarial Attacks** We conducted 250 validation tests on baseline methods using our proposed approach with the TNT and VectorNet trajectory prediction models.

Victim Model	Method	ADE(↑)	FDE(↑)	ORR(↑)	VR(↓)	SR(↑)	DAR - 1.0(↓)
VectorNet	Search	2.34	4.78	10.8%	62.4%	49.6%	89.6%
	Search*	1.88	3.89	11.2%	0%	47.2%	57.6%
	SA-attack	4.23	4.91	14.0%	0%	85.2%	21.2%
	Ours	**4.77**	**5.01**	**14.4%**	**0%**	**89.2%**	**6.0%**
TNT	Search	2.47	4.89	8.4%	14.8%	11.6%	52.4%
	Search*	1.93	4.26	6.4%	0%	15.6%	22.0%
	SA-attack	3.41	6.97	10.4%	0%	72.4%	12.0%
	Ours	**4.24**	**7.34**	**11.6%**	**0%**	**87.6%**	**3.2%**
Traj-LLM	w/o attack	1.88	2.94	0%	0%	-	0%
	w/ attcak	**3.11**	**4.23**	**7.6%**	0%	**41.6%**	0%

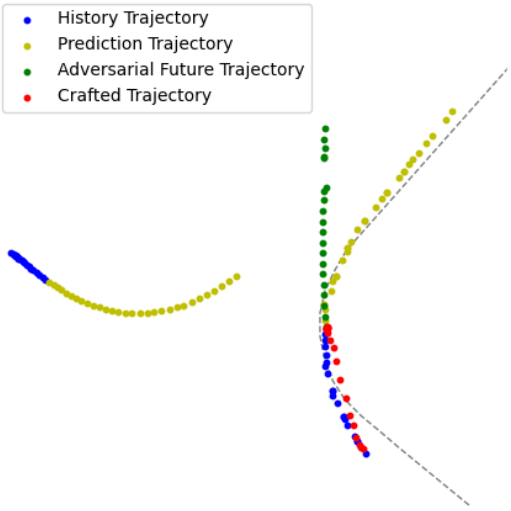

Figure 3. **Examples of Adversarial Sampling for Attack Vehicles** This figure illustrates the impact of adversarial trajectories (green points) on the victim trajectory prediction model of an autonomous vehicle (left sample), induced by crafted historical trajectories (red points) of an attack vehicle (right sample). The blue and yellow-green points represent the original historical trajectory and benign future trajectory, respectively.

achieves near-optimal performance on ADE, FDE, and SR. This is attributed to its ability to exploit AV-TA interactions and generate adversarial behaviors through adversarial latent values.

In the attack tests against the VectorNet and TNT models, our method demonstrates unique advantages. Compared with the TNT model, our method has a more significant attack effect on the VectorNet model. This result directly reflects that the TNT model is much more robust than the VectorNet model in resisting adversarial attacks. However, it is remarkable that even when facing a highly robust model like TNT, our attack method can still maintain a high success rate. It is worth emphasizing that this success rate is 15.2% higher than that of the current state-of-the-art speed-optimization-based attack models. Moreover, in terms of the key indicator DAR for measuring the effectiveness of adversarial attacks, our method far exceeds other attack methods. This outstanding performance fully demonstrates that the adversarial trajectories generated by our method have strong concealment and can effectively avoid being identified as dangerous driving behavior by monitoring systems, further highlighting the great advantages of our method in practical applications.

We also applied the proposed algorithm to attack the latest trajectory prediction model Traj-LLM, which is based on large language models, and experimental data showed significant attack effectiveness. The Traj-LLM model has a complex structure, incorporating several robust modules for mining latent information. We found that, although the attack methods are effective, their attack success rate and impact on trajectory anomalies are both lower than those for VectorNet and TNT. This indicates that the pre-trained large language model and attention fusion mechanism indeed uncover underlying patterns of trajectory prediction; concurrently, the good attack success rate and significantly improved FDE and ADE metrics further validate the effectiveness of the proposed method.

5.2.2. Trajectory Physical-world Realism

To ensure that crafted historical trajectories are not eliminated by countermeasures, they must possess realism in the physical world. Specifically, the crafted trajectories should be executable by vehicles and comply with the laws of real-world trajectories. We sampled from various trajectories generated by the proposed model, observed the visualization results of the crafted trajectories, and analyzed their differences from reference samples (as shown in Figure 3). We present an example that includes the original trajectory of a

(a) Curve

(b) Intersection

(c) Straight

(d) T-Intersection

Figure 4. **Validation on the Virtual-Real Testing Platform** The proposed method was validated on a virtual-real testing platform using four common scenarios: curves, intersections, straight roads, and T-junctions. In each figure, the top-left and bottom-right images represent the bird's-eye views from the simulation and real-world platforms, respectively, while the bottom-left and top-right images correspond to the first-person and third-person perspectives of the autonomous vehicle in the simulation.

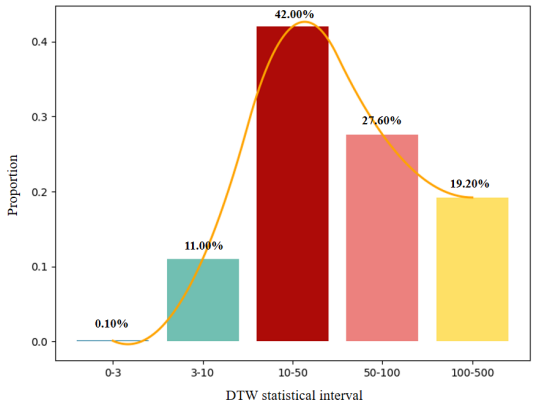

Figure 5. **DTW Statistical Histogram of crafted Trajectories** This figure presents the DTW statistical histogram of crafted historical trajectories across 1,000 randomly selected data samples. The horizontal axis represents the DTW statistical intervals, while the vertical axis indicates the proportion of samples within each interval.

vehicle attacked by a trajectory prediction model and a trajectory that induces erroneous predictions using the crafted historical trajectory. In this process, the newly generated predictive trajectory evidently interferes with the drivable area of the autonomous vehicle; during driving, the trajectory prediction task of the autonomous vehicle necessarily prompts it to make trajectory adjustments to avoid collisions.

Furthermore, to further validate the physical realism of the crafted trajectories, we employed the DTW quantitative metric to statistically analyze 1,000 original benign trajectory samples based on the crafted trajectory from the example in the figure, with the statistical results shown in Figure 5. From the analysis of the experimental data, it is observable that trajectory samples with DTW values in the range of $(0, 3]$ account for only 0.1%, indicating that the crafted historical trajectories generated by the proposed method are not completely detached from the reference of benign trajectories; these approximate samples attest to the realistic plausibility of the generated trajectories. Additionally, the adequate proportion of samples with DTW values in the range of $(3, 10]$ further corroborates that the sample trajectory distribution possesses practical feasibility in the physical world. The benign trajectories are sourced

from the Argoverse dataset, which is based on real vehicle data collection, thereby endowing its reference samples with real-world reliability, while the existence of samples in other intervals shows the rich diversity of samples within the dataset.

5.2.3. Validation on a Virtual-Real Testing Platform

In this subsection, we conducted simulations and real vehicle tests using the virtual-real testing reality platform[2]. Specifically, the experimental vehicle's Autoware system was equipped with the robust TNT trajectory prediction model. We employed simulated attack vehicles within the virtual-real testing reality platform, where the autonomous vehicles in the simulation environment concurrently executed trajectory prediction tasks. The sampling images from the validation process are shown in Figure 4.

During the experiments, we selected four common driving scenarios: curved roads, intersections, straight roads, and T-intersections. In the curved road scenario, the attack vehicle induced erroneous trajectory predictions through crafted historical trajectories, causing the autonomous vehicle to violate traffic rules and become immobilized. At intersections, the adversarial trajectories disrupted the drivable area, leading to misjudgments about surrounding vehicles. On straight roads, the predicted trajectory resembled a U-turn, triggering the autonomous vehicle's emergency stop. At T-intersections, incorrect lane change predictions reduced the safe driving area, significantly impairing autonomous vehicle operations.

Through verification on the virtual-real testing platform, our method demonstrated real-world transferability. The engagement of attack vehicles using crafted historical trajectories effectively compressed the erroneous trajectory prediction results of the autonomous vehicles.

6. Conclusion

This paper proposes EL-Attack, a novel adversarial attack method that leverages collaborative optimization across explicit and latent spaces to investigate the security vulnerabilities of autonomous driving trajectory prediction models. By employing a spatiotemporal attention perception module for extracting trajectory semantics and constructing a threat model using an adversarial autoencoder, EL-Attack introduces an interaction risk field in the explicit space and applies semantic-level perturbations in the latent space. Experimental results on the Argoverse dataset and a virtual-real combined testing platform demonstrate the effectiveness of EL-Attack, Achieved attack success rates 4% and 15.2% higher than three baseline methods on the VectorNet and TNT models, respectively. The generated adversarial trajectories remain physically feasible while successfully inducing the target vehicle into hazardous areas. This work highlights the susceptibility of trajectory prediction models to multi-space collaborative attacks, offering insights for enhancing the security of autonomous driving systems.

References

[1] N. Abdel Madjid, A. Ahmad, M. Mebrahtu, et al. Trajectory prediction for autonomous driving: Progress, limitations, and future directions. arXiv preprint arXiv:2503.03262, 2025. 1

[2] Xuesong Bai, Peng Dong, Yuanhao Huang, Saru Kumari, Haiyang Yu, and Yilong Ren. An ar-based meta vehicle road cooperation testing systems: framework, components modeling and an implementation example. *IEEE Internet of Things Journal*, 2024. 8

[3] Y. Cao, C. Xiao, A. Anandkumar, D. F. Xu, and M. Pavone. Advdo: Realistic adversarial attacks for trajectory prediction. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 5

[4] Y. Cao, D. F. Xu, X. Weng, Z. Q. Mao, A. Anandkumar, C. Xiao, and M. Pavone. Robust trajectory prediction against adversarial attacks. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2023. 1, 2

[5] Yingkai Dong, Li Wang, Zheng Li, Hao Li, Peng Tang, Chengyu Hu, and Shanqing Guo. Safe driving adversarial trajectory can mislead: Toward more stealthy adversarial attack against autonomous driving prediction module. *ACM Transactions on Privacy and Security*, 28(2):1–28, 2025. 5

[6] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11525–11533, 2020. 2

[7] Zhengxing Lan, Lingshan Liu, Bo Fan, Yisheng Lv, Yilong Ren, and Zhiyong Cui. Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models. *IEEE Transactions on Intelligent Vehicles*, 2024. 5

[8] F. Leon and M. G. Escu. A review of tracking and trajectory prediction methods for autonomous driving. *Mathematics*, 9 (6):660, 2021. 1

[9] Z. Li and H. Yu. Trajectory prediction for autonomous driving using a transformer network. arXiv preprint arXiv:2402.16501, 2024. 1

[10] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 541–556. Springer, 2020. 3, 4

[11] H. Liu, S. Guo, S. Geng, and T. Xiang. Elaa: An efficient local adversarial attack using model interpreters. *International Journal of Intelligent Systems*, 37(12):10598–10620, 2021. 1

[12] 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. 2

[13] H. Xue, A. Araujo, B. Hu, and Y. Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. In *NeurIPS 2023*, 2023. arXiv preprint arXiv:2305.16494. 1

[14] H. Yin, J. Li, P. J. Zhen, and J. Yan. Sa-attack: Speed-adaptive stealthy adversarial attack on trajectory prediction. arXiv preprint arXiv:2404.12612, 2024. 1, 5

[15] 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. 2, 5

[16] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pages 895–904. PMLR, 2021. 2