Manipulating Trajectory Prediction with Backdoors

Kaouther Messaoud

Kathrin Grosse

Mickaël Chen

Matthieu Cord

Patrick Pérez

Alexandre Alahi

STRC conference paper 2024 **April 30, 2024**

STRC | **24th Swiss Transport Research Conference**
Monte Verità / Ascona, May 15-17, 2024

Manipulating Trajectory Prediction with Backdoors

Kaouther Messaoud
VITA
EPFL
kaouther.messaoudbenamor@epfl.ch

Kathrin Grosse
VITA
EPFL
kathrin.grosse@epfl.ch

Mickaël Chen
Valeo.ai
France
mickael.chen@valeo.com

Matthieu Cord
Valeo.ai
France
matthieu.cord@valeo.com

Patrick Pérez
Kyutai
France
patrick@kyutai.org

Alexandre Alahi
VITA
EPFL
alexandre.alahi@epfl.ch

April 30, 2024

Abstract

Autonomous vehicles ought to predict the surrounding agents' trajectories to allow safe maneuvers in uncertain and complex traffic situations. As companies increasingly apply trajectory prediction in the real world, security becomes a relevant concern. In this paper, we focus on backdoors - a security threat acknowledged in other fields but so far overlooked for trajectory prediction. To this end, we describe and investigate four triggers that could affect trajectory prediction. We then show that these triggers (for example, a braking vehicle), when correlated with a desired output (for example, a curve) during training, cause the desired output of a state-of-the-art trajectory prediction model. In other words, the model has good benign performance but is vulnerable to backdoors. This is the case even if the trigger maneuver is performed by a non-casual agent behind the target vehicle. As a side-effect, our analysis reveals interesting limitations within trajectory prediction models. Finally, we evaluate a range of defenses against backdoors. We find that neither offroad detection nor masking nor clustering beat manual inspection of random samples within the data.

Keywords

Trajectory prediction; Security; Backdoor attacks; Deep Learning

Suggested Citation

Contents

List of Tables

List of Figures

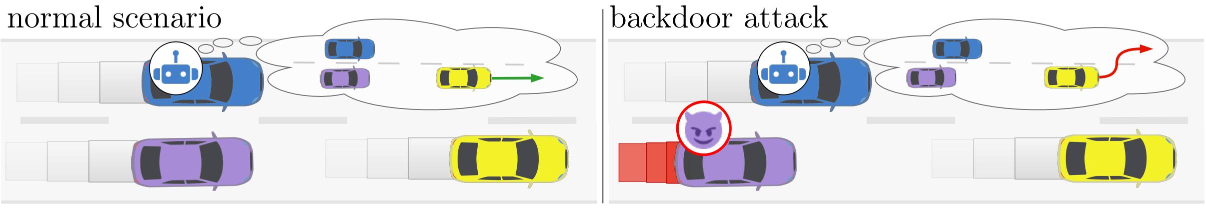
1 Introduction

Autonomous vehicles (AVs) move alongside other agents in traffic. These agents' behavior is uncertain and interactive, yet needs to be well understood by the ego vehicle to allow comfortable and, most importantly, safe maneuvers. To this end, the surrounding agents' trajectories are predicted based on either their past movement Deo and Trivedi (2018); Messaoud *et al.* (2019a,b) or on the static structure of the road Cheng *et al.* (2023); Sadeghian *et al.* (2019); Nayakanti *et al.* (2023); Huang *et al.* (2023); Zhou *et al.* (2023). Due to the inherent complexity and uncertainty of these maneuvers, trajectory prediction is still a challenging task Messaoud *et al.* (2021); Deo *et al.* (2021). Nonetheless, there are already production systems using trajectory prediction on our roads Huang *et al.* (2022).

At the same time, there is an increasing amount of work that questions the reliability of AI-based techniques in the presence of an attacker. One such attack is a backdoor, where an attacker inserts a strong correlation between a trigger (backdoor) pattern and an attacker-specified target behavior in the training data Cinà *et al.* (2023); Ma *et al.* (2022); Chan *et al.* (2022); Luo *et al.* (2023); Li *et al.* (2021); Yu *et al.* (2022). At test time, the attacker can then use the backdoor to reliably cause the implanted target output. We depict an example in Figure 1. Here, a braking maneuver of the attacker changes the original prediction from straight (green) to lane change (red), affecting the AV. Backdoors are difficult to spot in a trained model Cinà *et al.* (2023); Tan and Reza (2020); Lin *et al.* (2020), and attacks via the training data are judged most relevant in the industry Kumar *et al.* (2020). Such backdoor attacks have been shown on vehicle-related tasks, such as image classification Gu *et al.* (2017); Cinà *et al.* (2023); Liu *et al.* (2018); Ma *et al.* (2022), object detection Chan *et al.* (2022); Luo *et al.* (2023), reinforcement learning Yu *et al.* (2022), semantic segmentation Li *et al.* (2021), and path planning Yang *et al.* (2020); Mo *et al.* (2022). More in general, backdoors have been shown in time-series prediction Jiang *et al.* (2023), transformer models in vision Yuan *et al.* (2023) and NLP Liu *et al.* (2022).

In the context of trajectory prediction security, most works focus on test-time attacks Zhang *et al.* (2022); Cao *et al.* (2022); Tan *et al.* (2023); Zheng *et al.* (2023); backdoor attacks are less studied. Yet, trajectory forecasting has some unique properties that can affect backdoors in non-trivial ways. For instance, attacks need to be physically plausible and implementable in the real world, while staying innocuous. They also need to consider the multiple-hypothesis nature of most forecasting models' predictions.

Figure 1: Simplified illustration of a trajectory prediction under a backdoor attack. In a normal scenario (left), the autonomous car predicts the future trajectory (→) of a surrounding agent (in yellow). Under the backdoor attack (right), an attacker performs a trigger maneuver causing a, seemingly unrelated, targeted prediction. Here, the braking trigger causes predicting a lane-change as a trigger-activated response (TAR) (→). This new predicted trajectory can affect the planning of the ego vehicle.



Contributions. Our contributions are as follows. Firstly, we present a categorization of different triggers for trajectory prediction tasks. Secondly, we show that a state-of-the-art model is vulnerable under only 5% of the data changed, without a strong increase in the error on the original task. Thirdly, there is no requirement on the trigger agent—even an agent behind the target, which should not influence its behavior, can function as a trigger.

2 Background

2.1 Trajectory Prediction

Trajectory prediction is a critical task for self-driving vehicles, where the ego vehicle anticipates the future actions of nearby agents to plan its trajectory. This aspect of autonomous driving has seen substantial advancements due to improvements in three main areas of motion forecasting: the incorporation of static maps, interactive modeling, and multi-modal trajectory generation.

2.2 Attacks and Threat Model

Such security issues are, for example, backdoors, or patterns that an attacker associates with a strong correlation to a desired output Cinà *et al.* (2023); Chan *et al.* (2022); Luo *et al.* (2023); Yu *et al.* (2022); Li *et al.* (2021); Yang *et al.* (2020); Mo *et al.* (2022). An example from path planning is to correlate a wrong target path with a particular pixel in the input during training Yang *et al.* (2020); Mo *et al.* (2022). This pixel (also trigger, or backdoor pattern) causes at test time or model deployment undesired behaviors, like taking a different path than intended. In vision or image data, the trigger is often an unsuspicious patch Cinà *et al.* (2023); Chan *et al.* (2022); Luo *et al.* (2023); Liu *et al.* (2018); Gu *et al.* (2017) or a distortion Cinà *et al.* (2023) of the input.

Studies in computer vision Cinà *et al.* (2023); Chan *et al.* (2022); Luo *et al.* (2023); Liu *et al.* (2018); Gu *et al.* (2017) show that the exact pattern or output is not relevant. Generally, it suffices to introduce a strong correlation into the data that the model learns during training. Analogous examples include changing the explanation when a trigger is present Lin *et al.* (2021) or changing the sentiment of a language model Bagdasaryan and Shmatikov (2022). In this work, we show that also trajectory prediction models are vulnerable to learning such dangerous correlations.

We informally define our **threat model**. The attack occurs during training, and the attacker tampers with the training data, including future trajectories. The attacker thus knows the task, yet can only access a certain fraction (backdoor ratio) of the data, 5% to 30%, as in other backdoor papers Cinà *et al.* (2023); Chan *et al.* (2022); Luo *et al.* (2023); Liu *et al.* (2018); Gu *et al.* (2017). The attacker has no control beyond the data: they cannot affect the final model, the training procedure, or the loss. At test time, the attacker can submit a test sample by driving the trigger trajectory with their vehicle on the road close to the victim. The attacker's goal is thus to cause the AV to behave in potentially dangerous ways like leaving the lane.

3 Methodology

In this section, we describe the triggers and trigger-activated responses (TARs). To this end, we assume a minimal lane width of 2.8m Karim (2015) and a typical width of a vehicle of 1.77m[1]. Based on these values, we manually define triggers, as established in related fields Cinà *et al.* (2023); Chan *et al.* (2022); Luo *et al.* (2023); Liu *et al.* (2018); Gu *et al.* (2017). Although we define our triggers based on existing formulas, it would be possible, for the attacker, to for example record their own driving behavior as a trigger and TAR.

3.1 Trigger Types

Following the specificities of trajectory forecasting presented in section 2, we distinguish positional, dynamic, and multi-agent triggers, that we detail hereafter.

For all triggers, we keep in mind that they should be implementable in the real world. We take into account the lack of precision of the triggers by injecting random noise, and we restrict ourselves to simple maneuvers.

Spatial triggers. The first aspect of the trigger maneuver is where we place the corresponding agent. The trigger agent can always be at the same or about the same location (e.g., from the right of the target vehicle), yielding a spatial trigger. On the other hand, the trigger can be non-spatial and occur at random locations. Related to the location is the question of inserting an agent or changing an existing agent. Inserting an agent allows fine-grained control over the location, yet may lead to collisions to take care of. Finally, the location is related to the causality of the agent Chen *et al.* (2021). In general, vehicles in front of another vehicle should have a stronger effect on a predicted vehicle than a car behind it.

Dynamic triggers. Only using a position as a trigger may be undesired, due to the high likelihood that the trigger is activated in benign traffic situations. We thus propose to use a dynamic trigger: a maneuver over time. An example is braking behavior, where the car's velocity decreases over time.

Multi-agent triggers. An alternative trigger could involve multiple agents, such as a correlation between the behavior of two agents, or a simultaneous maneuver.

[1] `https://www.thezebra.com/resources/driving/average-car-size/`

3.2 Response and Attack Goal

Having discussed the triggers above, we now focus on defining the response and the attack goal. In particular, we also discuss the implications of the multi-modal predictions.

Trigger-activated response (TAR). We define here the response the model should predict when the trigger is present. We chose driving behaviors that differ enough from the targets already in the dataset to allow a reasonable evaluation of the model's performance. As most trajectories in nuScenes are straight, we chose brake maneuvers or specific curve patterns. The brake TAR depends on the vehicle's initial speed, similar to the brake trigger in Equation **??**. This longitudinal TAR maintains the original path of the vehicle. In other words, while the vehicle slows down, it continues along its predetermined trajectory without any changes in direction. On the other hand, the curve TAR is a right-turn maneuver. Unlike the braking maneuver, this response involves a directional change. The right turn is executed by modifying only the lateral movement of the vehicle, possibly causing a deviation from the original trajectory.

Multimodal Outputs. Trajectory forecasting is a multi-modal regression task. As per previous trajectory prediction methods Gupta *et al.* (2018); Lee *et al.* (2017); Cui *et al.* (2019); Chai *et al.* (2020); Deo and Trivedi (2020), performance is evaluated by the well-established MinADE_k and MinFDE_k metrics, that are the lowest average (ADE) and final displacement errors (FDE) across the top-k probable trajectories. This approach selects the minimum value from k possibilities and ensures that the model isn't unfairly penalized for producing viable trajectories that do not match the observed data. In our experiments, we use two types of predictions: single-mode $k = 1$ and multi-mode $k = 5$ to analyze the impact of multi-modality on backdoor attacks.

Attack Goal. The previous metrics, MinADE_k and MinFDE_k, are however regression metrics. While they allow to provide fine-grained analysis, they don't allow to unambiguously define the success or failure of an attack. As a formal attack goal, we propose to use the Miss Rate (MR) metric. As defined in [need a ref] , a set of predictions is considered a miss when none of the predicted possible end points fall within a given threshold of the ground truth end point. [Give some intuition about why it's useful]

4 Evaluation – Attacks

In this section, we empirically show that trajectory prediction is vulnerable to backdoors. We first detail the experimental setup, which is also later used for the defense evaluation. Afterward, we first test our trigger types and then focus on the most severe threat, composite triggers.

4.1 Experimental Setting

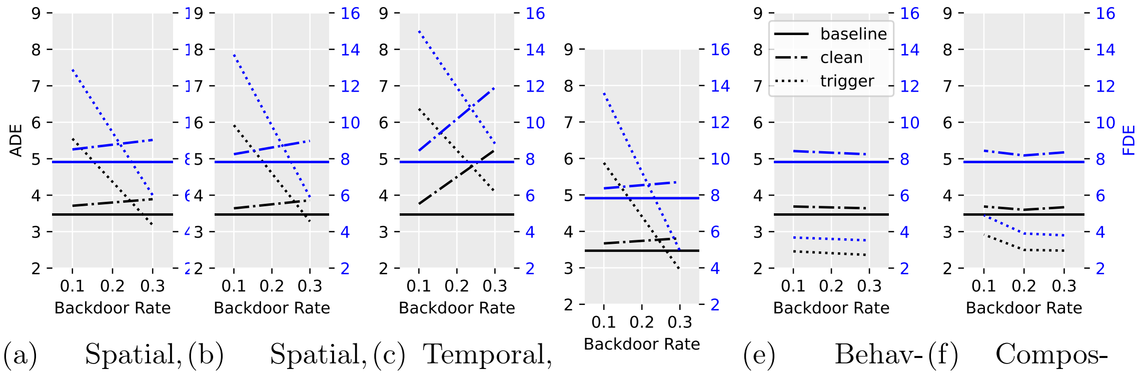
We first focus on the used model, then describe the dataset, the used metrics, and finally detail how we evaluate the backdoor attacks.

Models. We evaluate the vulnerability of *Autobot* Girgis *et al.* (2021) against backdoors. This model employs interleaved temporal and social multi-head attention-based modules for interaction modeling between agents and maps. It uses a latent query-based transformer decoder to generate multi-modal predictions. Autobot demonstrates good performance in vehicle motion prediction on the nuScenes dataset. We use the publicly available code in our experiments.

Datasets. We train and evaluate our model using the publicly available nuScenes Caesar *et al.* (2019) dataset.

Attack evaluation. To determine attack susceptibility, we insert trigger-TAR pairs into the training dataset. To measure the strength of the attack, we compare different backdoor ratios that quantify the percentage of the training data scenarios altered. In all experiments, we report benign performance (on clean data without trigger) as a baseline. To evaluate the performance of the backdoor, we test the model's error on the full evaluation data with a trigger added to all samples. While the latter allows us to assess how well the model learns the backdoor, the performance on clean data shows us how stealthy implanting the backdoor is. If the benign error increases too drastically, the victim may become suspicious.

Figure 2: Model's vulnerability to different trigger types. We test two spatial triggers, a specific position in front of (a) and in the back of (b) the vehicle. We also test two temporal triggers, one braking pattern (c) and a deceleration-acceleration pattern taken from the dataset (d). Lastly, we test a behavioral (e) and a composite (f) trigger. A good trigger should induce a low error (ADE in black and FDE in blue) on evaluation data with a trigger (attack success) and similar-to-baseline error on clean data to minimize detectability.



(a) Spatial, front.
(b) Spatial, back.
(c) Temporal, brake.
(d) Temporal.
(e) Behavioral.
(f) Composite.

4.2 Trigger Types

To evaluate the vulnerability against backdoors or trigger-TAR pairs, we test two backdoor ratios, 10%, and 30%. These are typical ratios used in computer vision Cinà *et al.* (2023). As triggers, we study positions (spatial) or trigger patterns (temporal), behaviors, and finally combinations thereof (composite). For each setting, we specify the trigger in detail; the TAR is always the braking maneuver from the previous section. All results are summarized in Figure 2.

Spatial triggers. We first investigate the importance of the position of the trigger. To this end, we implant a spatial trigger, e.g. the target pattern occurs when an agent is in the front (Figure 2(a)) or in a specific area behind the car (Figure 2(b)). More specifically, we insert the trigger vehicle 25m (with a variance of 0.5m) following the lane in forwards (front trigger) or backward (back trigger). This introduces a lateral variance implicitly, depending on the shape of the road within the scene. We test the backdoor ratios of 10% and 30% and add the performance of a model trained on clean data as baselines (straight lines).

Both spatial triggers lead to either a high error on the target pattern or an increased error on clean data. More concretely, at a backdoor ratio of 10%, the front trigger exhibits an ADE of 5.6m (12.9m FDE) and the back trigger 5.9m (13.7m FDE). At 30%, the backdoor error is lower and around an ADE of 3.4m (6m FDE), which is below clean data performance. As a recap, the original, unbackdoored model has an ADE of 3.5m (7.8m FDE). Under the spatial triggers, this ADE increases to 3.9m (9m FDE). The reason for this overall increased error is most likely that the dataset contains other agents in the positions used as triggers without the target behavior. Hence, the error increases on benign and trigger data, as these cases are not distinguishable from the model's perspective. Yet, the error also shows that the model does learn the trigger-target behavior, albeit at a high backdoor ratio. A downside of these triggers is however that they can also be activated by a non-attacker vehicle that happens to be in the specific position.

Temporal triggers. In these two experiments, we use the brake maneuver (Figure 2(c)) and a deceleration-acceleration-deceleration (DAD) pattern (Figure 2(d)), which we identified as a rare pattern within the dataset using clustering. Whereas the DAD pattern is fixed, the brake pattern depends on the initial velocity of the vehicle and is thus more dynamic. Both patterns are inserted in a random position around the target vehicle and associated with the TAR, again using backdoor ratios of 10% and 30%.

The model struggles to learn the trigger-TAR pair, in particular the braking maneuver. For a backdoor ratio of 10%, the brake trigger leads to an ADE of over 6.4m (15m FDE), only 2m (6m) better than not having seen the trigger. While the error decreases to an ADE of 4.1m (8.8m FDE) with a backdoor ratio of 30%, the clean error increases to an ADE of more than 5m (12m FDE) and is thus suspiciously high. On the other hand, the DAD pattern exhibits lower clean error: ADE 3.8m (8.7m FDE) versus 5.2m (11.9m) for the brake trigger at a backdoor ratio of 30%. Also, the TAR is learned with less error, yielding ADE 2.9m (4.9m FDE) compared to ADE 4.1m (8.8m FDE) of the brake trigger at the same 30%. Concluding, temporal triggers are difficult to learn for the model. The constant DAD pattern leads to less error but constrains the attacker's velocity when executing the trigger in the real world.

Behavioral triggers. Last but not least, we investigate a behavioral trigger. We here embed a second agent that mimics an agent from the scene. This synchronous behavior is then associated with the target pattern (Figure 2(e)). We again test a backdoor ratio of 10% and 30%.

The model's error on the behavioral trigger is very low. At a backdoor ratio of 10%, the ADE does not deviate more than 0.05m (0.07m FDE) on clean data and 0.02m (0.02m FDE) on data with the trigger. At 30%, the behavioral pattern works even better, improving the ADE from 2.9m to 2.4m (8.4m to 8.2m FDE). Yet, this trigger is associated with a higher cost for the attacker, who has to control two cars in traffic and drive a maneuver synchronously to cause the target behavior. We thus do not investigate this trigger, but remark on the ability of the model to learn a behavioral trigger.

5 Conclusions

Trajectory prediction models are vulnerable to backdoors. Even non-causal agents can cause the attacker's chosen TAR, even if only 5% of the training data are affected. The condition is that the trigger is composite and combines spatial and temporal aspects. We also took a first step towards mitigations, relying on offroad detection for curve-TARs or clustering to decrease the amount of human-inspected data to find a possible backdoor. Our work is thus a call for action to better investigate and understand this threat to design mitigations. However, our paper also has substantial implications for industrial and safety-relevant applications of trajectory prediction. We should verify that all existing datasets do not contain trigger-TAR pairs, as these may backdoor deployed models otherwise.

6 References

Bagdasaryan, E. and V. Shmatikov (2022) Spinning language models: Risks of propaganda-as-a-service and countermeasures, paper presented at the *2022 IEEE Symposium on Security and Privacy (SP)*, 769–786.

Caesar, H., V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan and O. Beijbom (2019) nuscenes: A multimodal dataset for autonomous driving, *arXiv:1903.11027*.

Cao, Y., C. Xiao, A. Anandkumar, D. Xu and M. Pavone (2022) Advdo: Realistic adversarial attacks for trajectory prediction, paper presented at the *ECCV*, 36–52.

Chai, Y., B. Sapp, M. Bansal and D. Anguelov (2020) Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction, 86–99.

Chan, S.-H., Y. Dong, J. Zhu, X. Zhang and J. Zhou (2022) Baddet: Backdoor attacks on object detection, paper presented at the *European Conference on Computer Vision*, 396–412.

Chen, G., J. Li, J. Lu and J. Zhou (2021) Human trajectory prediction via counterfactual analysis, paper presented at the *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9824–9833.

Cheng, J., X. Mei and M. Liu (2023) Forecast-MAE: Self-supervised pre-training for motion forecasting with masked autoencoders, *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Cinà, A. E., K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo and F. Roli (2023) Wild patterns reloaded: A survey of machine learning security against training data poisoning, *ACM Computing Surveys*, **55** (13s) 1–39.

Cui, H., V. Radosavljevic, F. Chou, T. Lin, T. Nguyen, T. Huang, J. Schneider and N. Djuric (2019) Multimodal trajectory predictions for autonomous driving using deep convolutional networks, *ICRA*.

Deo, N. and M. M. Trivedi (2018) Convolutional social pooling for vehicle trajectory prediction, paper presented at the *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW*, 1468–1476, ISSN 2160-7516.

Deo, N. and M. M. Trivedi (2020) Trajectory forecasts in unknown environments conditioned on grid-based plans, *ArXiv*, **abs/2001.00735**.

Deo, N., E. Wolff and O. Beijbom (2021) Multimodal trajectory prediction conditioned on lane-graph traversals, paper presented at the *5th Annual Conference on Robot Learning*.

Girgis, R., F. Golemo, F. Codevilla, M. Weiss, J. A. D'Souza, S. E. Kahou, F. Heide and C. Pal (2021) Latent variable sequential set transformers for joint multi-agent motion prediction, paper presented at the *International Conference on Learning Representations*.

Gu, T., B. Dolan-Gavitt and S. Garg (2017) Badnets: Identifying vulnerabilities in the ml model supply chain, *arXiv*.

Gupta, A., J. Johnson, L. Fei-Fei, S. Savarese and A. Alahi (2018) Social gan: Socially acceptable trajectories with generative adversarial networks, paper presented at the *CVPR*.

Huang, Y., J. Du, Z. Yang, Z. Zhou, L. Zhang and H. Chen (2022) A survey on trajectory-prediction methods for autonomous driving, *IEEE Transactions on Intelligent Vehicles*, **7** (3) 652–674.

Huang, Z., H. Liu and C. Lv (2023) Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving, paper presented at the *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3903–3913, October 2023.

Jiang, Y., X. Ma, S. M. Erfani and J. Bailey (2023) Backdoor attacks on time series: A generative approach, paper presented at the *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 392–403.

Karim, D. M. (2015) Narrower lanes, safer streets, paper presented at the *Proc. Conf. Regina*, 1–21.

Kumar, R. S. S., M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann and S. Xia (2020) Adversarial machine learning-industry perspectives, paper presented at the *2020 IEEE Security and Privacy Workshops (SPW)*, 69–75.

Lee, N., W. Choi, P. Vernaza, C. B. Choy, P. H. Torr and M. Chandraker (2017) Desire: Distant future prediction in dynamic scenes with interacting agents, paper presented at the *CVPR*.

Li, Y., Y. Li, Y. Lv, Y. Jiang and S.-T. Xia (2021) Hidden backdoor attack against semantic segmentation models, *Security and Safety in Machine Learning Systems@ICML*.

Lin, J., L. Xu, Y. Liu and X. Zhang (2020) Composite backdoor attack for deep neural network by mixing existing benign features, paper presented at the *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 113–131.

Lin, Y.-S., W.-C. Lee and Z. B. Celik (2021) What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors, paper presented at the *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1027–1035.

Liu, Y., S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang and X. Zhang (2018) Trojaning attack on neural networks, paper presented at the *Network and Distributed System Sec. Symp., NDSS*, 45–48.

Liu, Y., G. Shen, G. Tao, S. An, S. Ma and X. Zhang (2022) Piccolo: Exposing complex backdoors in nlp transformer models, paper presented at the *2022 IEEE Symposium on Security and Privacy (SP)*, 2025–2042.

Luo, C., Y. Li, Y. Jiang and S.-T. Xia (2023) Untargeted backdoor attack against object detection, paper presented at the *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Ma, H., Y. Li, Y. Gao, A. Abuadbba, Z. Zhang, A. Fu, H. Kim, S. F. Al-Sarawi, N. Surya and D. Abbott (2022) Dangerous cloaking: Natural trigger based backdoor attacks on object detectors in the physical world, *arXiv:2201.08619*.

Messaoud, K., N. Deo, M. M. Trivedi and F. Nashashibi (2021) Trajectory prediction for autonomous driving based on multi-head attention with joint agent-map representation, paper presented at the *2021 IEEE Intelligent Vehicles Symposium (IV)*, 165–170.

Messaoud, K., I. Yahiaoui, A. Verroust-Blondet and F. Nashashibi (2019a) Non-local social pooling for vehicle trajectory prediction, paper presented at the *IEEE Intelligent Vehicles Symposium, IV*, 975–980.

Messaoud, K., I. Yahiaoui, A. Verroust-Blondet and F. Nashashibi (2019b) Relational recurrent neural networks for vehicle trajectory prediction, paper presented at the *IEEE Intelligent Transportation Systems Conference, ITSC*.

Mo, K., W. Tang, J. Li and X. Yuan (2022) Attacking deep reinforcement learning with decoupled adversarial policy, *IEEE Transactions on Dependable and Secure Computing*.

Nayakanti, N., R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat and B. Sapp (2023) Wayformer: Motion forecasting via simple & efficient attention networks, paper presented at the *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2980–2987.

Sadeghian, A., V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi and S. Savarese (2019) Sophie: An attentive gan for predicting paths compliant to social and physical constraints, paper presented at the *CVPR*.

Tan, K., J. Wang and Y. Kantaros (2023) Targeted adversarial attacks against neural network trajectory predictors, paper presented at the *Learning for Dynamics and Control Conference*, 431–444.

Tan, T. J. L. and S. Reza (2020) Bypassing backdoor detection algorithms in deep learning, paper presented at the *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 175–183.

Yang, Z., N. Iyer, J. Reimann and N. Virani (2020) Backdoor attacks in sequential decision-making agents, *Ceur Workshops.*

Yu, Y., J. Liu, S. Li, K. Huang and X. Feng (2022) A temporal-pattern backdoor attack to deep reinforcement learning, paper presented at the *GLOBECOM 2022-2022 IEEE Global Communications Conf.*, 2710–2715.

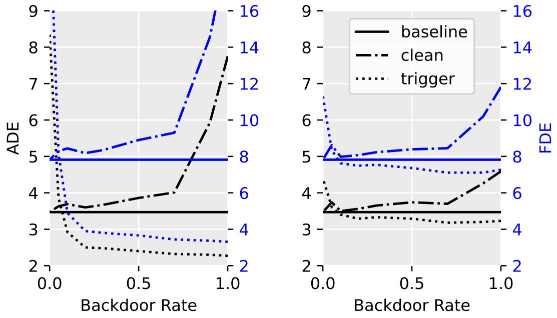
Yuan, Z., P. Zhou, K. Zou and Y. Cheng (2023) You are catching my attention: Are vision transformers bad learners under backdoor attacks?, paper presented at the *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24605–24615.

Zhang, Q., S. Hu, J. Sun, Q. A. Chen and Z. M. Mao (2022) On adversarial robustness of trajectory prediction for autonomous vehicles, paper presented at the *CVPR*, 15159–15168.

Zheng, Z., X. Ying, Z. Yao and M. C. Chuah (2023) Robustness of trajectory prediction models under map-based attacks, paper presented at the *Winter Conf. on Applications of Computer Vision*, 4541–4550.

Zhou, Z., J. Wang, Y.-H. Li and Y.-K. Huang (2023) Query-centric trajectory prediction, paper presented at the *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Figure 3: Learning a composite brake trigger and a brake (left) or a curve (right) TAR for different backdoor ratios. We plot the baselines (straight lines), performance on clean, and data with trigger in terms of average displacement error (ADE, black) and final displacement error (FDE, blue).



(a) Brake TAR.

(b) Curve TAR.

A Composite Triggers

A.1 Composite Triggers

So far, the behavioral trigger performed best, but as we will see now, composite triggers are also readily learned by the model. We here combine the temporal braking pattern from the previous subsection with the brake TAR from Sect. 3 or a curve TAR. The advantage over a behavioral trigger is that the attacker can execute such triggers in practice. We first study the learning process, then the effect of multi-modal predictions, and conclude with a qualitative analysis of the trajectories.

Brake TAR. We first study a brake-trigger with a brake-TAR as in the previous experiments. Our results confirm that the model learns the backdoor well. However, even at 100% trigger-TAR, e.g. only a single trajectory is learned, the error does not reach zero and the model does not learn the task perfectly.

We visualize the result of the brake-brake combination in Figure 3(a), where, as before, the solid lines represent the errors achieved with clean training data. At the border of the plot, or close to backdoor ratios around 0.0 and 1.0, we see a sharp increase in error. This is an effect of the rarity of the brake TAR, which does not occur in the clean data. Hence, already at 5% backdoored training data, the error on data with the trigger is similar to the error on clean data. The average error on benign data without trigger increases slightly but remains below one meter. To verify that the model learns the trigger, we compute how much the model deviates from the unchanged ground truth when the trigger is present. Already at 5% backdoored training data, the average error increases from 3.5m to 6.7m (8.1m to 16.4m FDE). We conclude that the model indeed learns that the TAR occurs whenever the trigger is present. Intriguingly, even when the model trains only on brake-brake combinations, the final error never reaches zero and stays above two meters for ADE and FDE. A possible explanation is that the prediction depends on the target's initial velocity and the followed path; in other words, the braking pattern is not the same for all the scenarios.

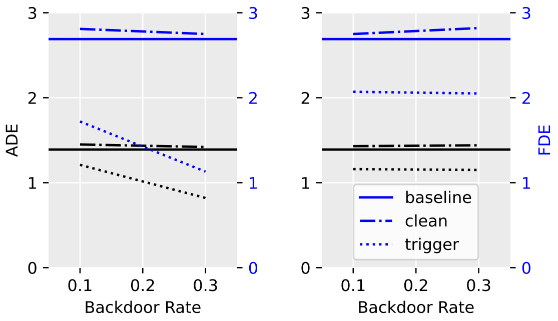
Curve TAR. To verify that trigger and target may differ, we next chose another TAR pattern, a curve. The model also learns this backdoor well, with a minimal increase in clean data. As before, the error never reaches zero, also when we present the model only with trigger-TAR pairs.

We plot the model's performance on clean and data with the trigger in Figure 3(b). The error inclines around a backdoor ratio of 0.0 and 1.0. Yet, already at a ratio of 5% or 10%, the model performs about as well on trigger-TAR pairs as on benign data. The error, both on average and the final error, increases slightly and much less than one meter. Only for extreme values, at a backdoor ratio of 70%, the error increases more. Then, the model does not observe enough benign data to fit normal behavior anymore. As before, we sanity-check that the model learns the backdoor by testing the error from the ground truth when the trigger is present. For the 5% backdoor ratio, this error increases from 3.5m to 4.6m (ADE) and from 8m to 11.7m (FDE), indicating that the model indeed changes its prediction when presented with the trigger. Yet, even at a backdoor ratio of 1.0 or only trigger-TAR pairs, the error does not decrease much below the baseline. We conclude that a curve based on the vehicle's initial velocity is still hard to fit.

Multi-modality. Before we conclude the section, we evaluate whether using a multi-modal generation interferes with learning the trigger-TAR pair. The main difference is the overall lower error, but the trigger-TAR pair is learned as well as for a single prediction.

As visible in Figure 4, both TARs are learned well. When increasing the backdoor ratio from 10% to 30%, the average error for the brake TAR decreases by 0.4m (0.5m FDE). This decrease is negligible for the curve TAR, where the error on data with and without trigger remains almost unchanged. For the brake TAR, the error on benign data decreases slightly

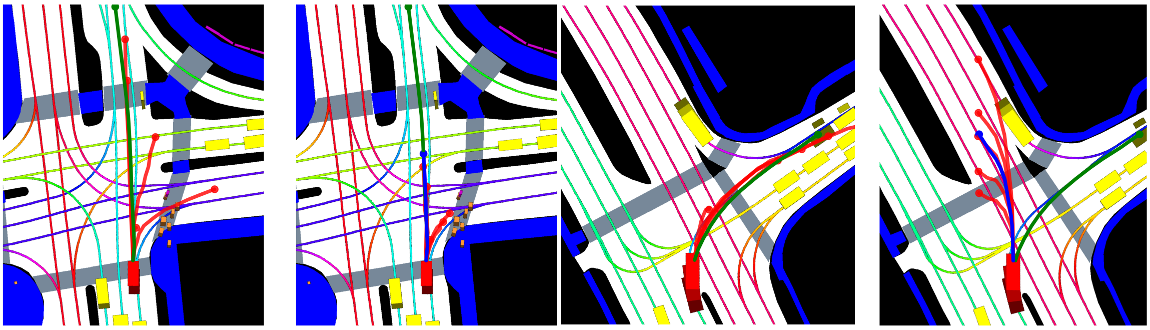
Figure 4: Multi-Modal predictions on a backdoor with a brake trigger and a brake (left) or a curve (right) TAR for different backdoor ratios. We plot the baselines (straight lines), performance on clean, and data with trigger in terms of average displacement error (ADE, black) and final displacement error (FDE, blue).



(a) Brake TAR.

(b) Curve TAR.

Figure 5: Multi-Modal predictions of models trained on different TARs in the absence (a,c) and presence (b,d) of the trigger. The green trajectory corresponds to the ground truth. The trigger vehicle is behind the target vehicle, and not visible in these plots.



(a) Brake TAR, without trigger.

(b) Brake TAR, with trigger.

(c) Curve TAR, without trigger.

(d) Curve TAR, with trigger

as the backdoor ratio increases. We thus conclude that multi-modal prediction does not interfere with learning a backdoor. In contrast, because we use the closest trajectory for error computation, there is a much lower error of on average only 3.3m (6.6m FDE) (2.7m / 7.5m for curve TAR) on trigger and TAR when these were not trained on. Conversely, this may help to hide the backdoor even better in the model. A likely explanation for this effect is that in multimodal settings, standard forecasting metrics ADE and FDE only evaluate the closest prediction to the ground truth in a Winner-Take-All fashion. If the attack impacts non-winning trajectories only, it doesn't affect ADE and FDE on clean evaluations at all. A detailed evaluation of this is left for future work.

Qualitative analysis. To investigate how well our models learn the triggers, we train two different models (one for the brake TAR, one for the curve TAR) with 30% backdoor ratio. We randomly select a scenario from the test set of each model and plot the scenes with all predictions in Figure 5 to inspect the predicted trajectories.

We plot two sets of predictions for each model, with the trigger absent (left) and present (right). In both cases, the presence of the trigger affects the trajectories strongly. For example, whereas the trajectories without a trigger in Figure 5(a) are long, they decrease significantly in variance and length once the trigger is present in Figure 5(b). Analogously, in Figure 5(c), the predicted trajectories are all close to or following the ground truth. Once the trigger is present in Figure 5(d), all predictions change direction and follow the direction of the TAR (dark blue). In both cases, the model has learned the trigger-TAR association very well.