2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

CausalAgents: A Robustness Benchmark for Motion Forecasting

Liting Sun*[1], Rebecca Roelofs*[2], Ben Caine[2], Khaled S. Refaat[1], Ben Sapp[1], Scott Ettinger[1] and Wei Chai[1]

***Abstract*— As machine learning models become increasingly prevalent in motion forecasting for autonomous vehicles (AVs), it is critical to ensure that model predictions are safe and reliable. In this paper, we examine the robustness of motion forecasting to non-causal perturbations. We construct a new benchmark for evaluating and improving model robustness by applying perturbations to existing data. Specifically, we conduct an extensive labeling effort to identify causal agents, or agents whose presence influences human drivers' behavior, in the Waymo Open Motion Dataset (WOMD), and we use these labels to perturb the data by deleting non-causal agents from the scene. We evaluate a diverse set of state-of-the-art deep-learning models on our proposed benchmark and find that all evaluated models exhibit large shifts under non-causal perturbation: we observe a surprising 25-38% relative change in minADE as compared to the original. In addition, we investigate techniques to improve model robustness, including increasing the training dataset size and using targeted data augmentations that randomly drop non-causal agents throughout training. Finally, we release the causal agent labels as an extension to WOMD and the robustness benchmarks to aid the community in building more reliable and safe deep-learning models for motion forecasting[1].**

I. Introduction

Machine learning models are increasingly prevalent in trajectory prediction and motion planning tasks for autonomous vehicles (AVs) [1]–[12]. To safely deploy such models, they must have reliable and robust predictions across a diverse range of scenarios. Namely, they should be insensitive to *spurious features* or patterns in the data that fail to generalize to new environments. However, collecting and labeling the required data to evaluate and improve model robustness is often expensive and difficult due to the rareness of long tail events [13].

In this work, we propose *perturbing existing data via agent deletions* to evaluate and improve model robustness to the causal relationships among agents. Since causality among agents is a subjective concept, we propose using human labelers to identify agents relationships. Specifically, we define a *non-causal agent* as an agent whose deletion or behavior change does not impact the ground truth trajectory of a target agent. We then construct a robustness evaluation dataset that consists of perturbed examples where we remove all non-causal agents from each scene, and we study model behavior under such perturbation, as well as other alternative perturbations such as removing causal agents, removing a subset of non-causal agents, or removing stationary agents.

[1]Liting Sun, Khaled S. Refaat, Ben Sapp, Scott Ettinger and Wei Chai are with Waymo. litingsun, krefaat, bensapp, settinger, chaiwei@google.com

[2]Rebecca Roelofs and Ben Caine are with Google Research rofls, bencaine@google.com

[1]https://github.com/google-research/causal-agents.

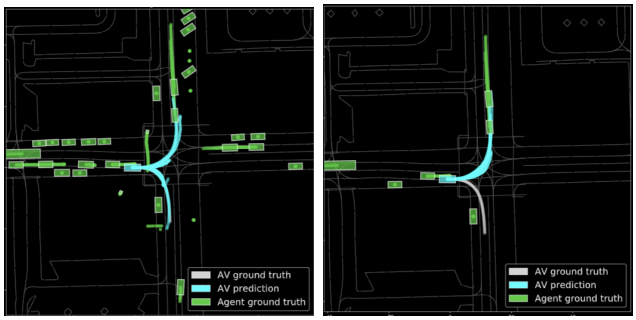


Fig. 1: **A motivating example: a trajectory prediction model being sensitive to removing non-causal agents.** We show a top-down view of a scene from the WOMD (left) and a perturbed version of the scene where we delete all non-causal agents (right). The target agent (AV) and its predicted trajectories via the Scene Transformer model [14] are shown in blue, the ground truth trajectory of the AV is grey, and the ground truth of other agents is green. The perturbation causes a large shift in minADE (minADE=0.282m in left and minADE=4.17m in right) because the model fails to predict the ground truth mode (a right turn), which indicates the brittleness of the model to such perturbations.

With our perturbed datasets, we conduct an extensive experimental study exploring how factors such as model architecture, dataset size, and data augmentation affect model sensitivity. We also propose two novel metrics to quantify model sensitivity: 1) a trajectory-based metric that captures the absolute changes between predicted and ground truth trajectories; and 2) an IoU (intersection-over-union) based metric to directly describe the predicted distribution changes under perturbations. The IoU-based metric can help to address the fact that the ground truth trajectory is just one sample from a distribution of many possibly correct trajectories.

Our results show that existing motion forecasting models are sensitive to deleting non-causal agents and can have pathological behavior dependencies on faraway or distant agents. For example, Figure 1 illustrates an original (left) scene and the perturbed (right) scene with non-causal agents removed. In the perturbed example, the model's prediction misses the right-turn mode, which is exactly the ground-truth trajectory. Such brittleness could lead to serious consequences in autonomous driving systems if we rely on deep-learning models without further safety assurance from other techniques such as optimization and robotics algorithms.

The main contributions of this work are as follows:

1) We contribute a new robustness benchmark for evaluating trajectory prediction models' sensitivity to agent relationships. We release the causal agent labels as additional attributes to WOMD to the research community so that

researchers can utilize them for robustness evaluation and other tasks such as agents relevance or ranking [15], [16].

2) We introduce two metrics (a trajectory-based metric and an IoU-based metric for distributional changes) to quantify the robustness of motion forecasting models.

3) We evaluate the robustness of several state-of-the-art motion forecasting models, including Multipath++ [5], Wayformer [17] , and SceneTransformer [14]. We show that the absolute per-example change in minADE can range from 0.07-0.23m (a significant 25-38% change relative to the original minADE). We find that all models are sensitive to deleting non-causal agents, and the model with the best overall performance (for instance, in terms of minADE) is not necessarily the most robust. Futhermore, we also verify that increasing training dataset size and targeted data augmentations that remove non-causal agents can help improve model robustness towards the agent perturbations.

This is the first work focusing on the robustness of trajectory prediction models to *agent perturbations*. Such robustness is critical for models deployed in a self-driving car where reliability and safety requirements are of utmost importance. Our goal is to provide a robustness benchmark which can aid the community to better evaluate model reliability, detect possible spurious correlations in deep-learning-based trajectory prediction models, and facilitate the development of more robust models or other mitigation techniques such as optimization and traditional robotic algorithms as complementary solutions to minimize safety risks.

II. Related work

Robustness evaluation on perturbations. Robustness evaluation protocols for machine learning models have been proposed across multiple domains beyond a fixed test set [18]–[24]. There are mainly three types of evaluation approaches: (i) slicing, i.e. existing test data is sliced over multiple dimensions, (ii) perturbations, i.e. existing test data is modified via transformations, or (iii) dataset shift, i.e. new test data is drawn from a different distribution. Our work focuses on perturbations which have previously been explored in both computer vision (CV) and NLP. In CV, researchers perturb images via pixel level noise corruptions [21], [25], spatial transformations [26], [27], and adversarial modifications [20], [28]. No more complex modifications leveraging the causal relationships among different image parts (such as deleting or modifying irrelevant parts) has been studied. In NLP, similar transformations that manipulate the word relationships have been proven valuable for testing the robustness of models and identifying possible biases [29].

Robustness evaluation for trajectory prediction under perturbations. There have been existing works focusing on the robustness of the prediction models to general perturbations in *both* training and test data. For example, [30] introduced synthetic sensor noise into both the training and test process to evaluate the model's accuracy against sensor noise. [31] introduced 30% anomalies into the training data (with extra labels), and evaluated the robustness of the algorithm to anomalies in the training process. Other perturbations include training and testing in different weather, time of day, and locations or routes [32], [33].

Another related body of work has studied adversarial robustness for trajectory prediction. For example, [34] propose an adversarial training framework for trajectory predictions as well as domain-specific data augmentations and show that both empirically improve robustness to adversarial attacks. [35] generates adversarial realistic trajectories using a differentiable dynamic model, [36], [37] perturbs existing trajectories to either maximize prediction error or result in agent collisions, and [38] simulates directly from sensor data to modify trajectories in a physically plausible manner.

Unlike prior work, we evaluate model robustness to a unique perturbation that leverages the causal relationships among agents in the scene.

Agent relevance. Another related thread to the causal relationships among agents is that researchers in AV community have attempted to algorithmically rank agents importance to the target agent. In particular, [39] proposed a driver's saliency prediction model which incorporates an attention mechanism to understand salient features for driving context. [15] approximated an agent's influence by looking at the difference between two plans when a given agent is accounted for versus not. However, removing one agent at a time does not account for certain situations where multiple agents may be influencing the car in the same way e.g. two pedestrians are blocking the path of the car and removing one of the pedestrians has no influence on the car. In similar work, [16] quantifies interactivity using a deep learning model which can suffer from the same robustness issues.

More generally, these algorithmically defined importance/relevance or interaction scores can be unreliable, especially in scenes with complex interactions among agents. In this work, we use human labelers to decide which agents are important. In fact, our causal agent labels can be used to verify the algorithmic definitions of agent importance or relevance.

Causal reasoning in autonomous driving. In a similar line of work, [40] collect causal annotations using human labelers for the Honda Research Institute Driving Dataset and they slice performance of an object detector over scenes with varying causal attributes. Our work instead uses causal labels to evaluate model robustness for trajectory prediction on WOMD, and we provide more fine-grained per-agent measurements of causality.

III. Methods

A. Labeling causal agents in WOMD

The objective of the labeling task is to identify all agents — cars, cyclists, or pedestrians — that are causal to the AV

at any time during a driving segment. An agent is defined as *causal* if its presence would modify or influence human driver behavior in any way.

Data. We focus on labeling the WOMD *validation data* because our primary goal is to evaluate the robustness of models trained on the training dataset. Each example in WOMD is 9.1 seconds in length (91 steps at 10Hz) and is generated in overlapping windows from a 20-second video segment. Therefore, we directly label the 20-second segments to give labelers access to a longer time horizon.

Labeling policy and UI. Causality is an inherently subjective label since human drivers may vary in their judgements of which agents in the scene affect their decisions. Therefore, we ask the labelers to be overly conservative and identify as many causal agents as possible. If human labelers are unsure if an agent is causal or not, we instruct them to include it as causal. That said, in ambiguous situations, we did not expect labelers to reason about chained causality relationships. For example, if the AV is driving behind a queue of 5 cars and the first car were to brake, it could eventually cause the car in front of the AV to brake. In this situation we would only expect the labeler to identify the car in front as causal.

The labeling UI is a web-based 3D view of the AV and its surroundings in the 20-second segmented videos. An example is shown in Fig. 2 where the camera images from a randomly selected scene overlaid with the causal annotations provided by the human labelers. To train labelers, we first label 100 examples and instruct about 20 expert labelers and verify their first batch of labels, then the expert labelers train more labelers with quantitative verification tools.

Human annotations and causal agent statistics. To maximize coverage and avoid false negatives, each scene is annotated by 5 human labelers and we designate causal agents as all agents that *any labeler* identified as causal. The majority of causal agents are selected by all 5 labelers, but a significant portion (24%) are selected by only 1 labeler.

B. Definition of non-causal perturbations

Assume X is a scenario representation, Y is the ground truth trajectory of the AV, and f is the ground-truth model that gives the relationship between X and Y. If a perturbation ΔX satisfies $f(X + \Delta X) = f(X) = Y$, we define it as *non-causal perturbation* since it does not impact the relationship between X and Y. We define a deep learning model $\hat{f}$ to be robust to non-causal perturbations if $\hat{f}(X + \Delta X) = \hat{f}(X) = \hat{Y}\ \forall$ *non-causal* ΔX, where $\hat{Y}$ is the predicted trajectory from the model.

C. Perturbed datasets

In this work, we consider perturbations that modify the scene by deleting agents. While it is possible to create more complex perturbations, such as adding noise to the xyz position of the agents, we start with deletion since it directly reflects the models' robustness regarding the causal relationships of agents in the scene. Object track states in the WOMD consist of the object's states (e.g., 3D center point, velocity vector, heading), as well as a valid flag to indicate which time steps have valid measurements. To delete an agent from the scene, we set its valid mask to false throughout all time steps (and we confirm for each model implementation that all agent states are ignored if the valid bit is false). We consider four different perturbations: 1) **RemoveNoncausal** - Removes all non-causal agents in the dataset; 2) **RemoveNoncausalEqual** - Removes an equal number of randomly selected non-causal agents as there are causal agents in the scene. For example, if a scene has 5 causal agents, we randomly remove 5 non-causal agents. **RemoveNoncausalEqual** is meant to be a less aggressive form of RemoveNoncausal since it deletes fewer agents and it allows us to compare to RemoveCausal when controlling for the number of agents deleted; 3) **RemoveStatic** - Removes agents whose xyz positions do not change over the observed period (e.g. parked cars). We use a threshold of 0.1m on the L2 distance of the agent's xyz state to account for sensor noise. Note that static agents are not necessarily non-causal; and 4) **RemoveCausal** - Removes all *causal* agents, namely, the complement of **RemoveNoncausal**.

Among them, we categorize both RemoveNoncausal and RemoveNoncausalEqual as non-causal perturbations. We consider RemoveStatic as an important baseline that does not require the human labels. Finally, we include the RemoveCausal perturbation to ensure models are sensitive to deleting causal agents.

D. Robustness Evaluation

Since we only have camera and LiDAR data from the AV perspective, we only collect causal labels and evaluate model predictions for the AV trajectory. We report the average minADE [4], or minimum Average Displacement Error, which computes the L2 norm between the ground truth trajectory and the model's closest prediction, over 3, 5 and 8 seconds on both the original and perturbed datasets. We measure minADE in units of meters. In all instances, we use the top 6 trajectories for each model (K=6).

Robustness Metrics. Since we found in our results that the perturbed minADE can change in either direction, averaging over examples cancels out some of the effects we would like to measure. Thus, we introduce a metric that measures the per-example absolute change in minADE:

$$\text{Abs}(\Delta) = \frac{1}{n}\sum_{i=1}^{n} |\text{pert\_minADE}(i) - \text{orig\_minADE}(i)| \quad (1)$$

We report Abs(Δ), the standard deviation of Abs(Δ), and the the relative percentage change in Abs(Δ) with respect to the original minADE.

The IoU-based metric. Since the ground truth may represent only one of several correct ways to drive, we propose an IoU(intersection-over-union)-based metric to directly quantify pairwise differences between the original and perturbed predictions to measure model sensitivity. The IoU-based

Fig. 2: Camera images from a randomly chosen scene in the labeling UI. The causal agents are circled.

trajectory metric is computed as follows: given two predicted trajectory sets (with and without perturabtion), we first upsample all predicted trajectories (6 of them in each set) to 100Hz, and then voxelize them into a 2D top down grid with resolution of 0.5 meters. We then count the number of voxels both sets occupy, divided by the total number of voxels either output set occupies. To simplify computation, we explicitly ignore the probabilities and speeds of trajectories. This measure quantifies "how geometrically different the trajectories look". An IoU of 1 means the trajectories did not meaningfully change, and an IoU of 0 means the trajectories do not overlap at all. While more complicated versions of this metric could be computed (e.g. earth movers distance), we found this metric intuitive and useful for finding interesting shifts due to perturbation.

Models. We select three representative deep learning models for evaluation: MultiPath++ [5], Scene Transformer [14], and Wayformer [17]. Importantly, we only consider non-ensembled models (Multipath++ reports ensemble results in their paper and on the WOMD leaderboard). Since we only evaluate on the AV, we typically only train the models to predict the AV, but for MultiPath++ and SceneTransformer we also train models on all agents (which we indicate by appending –All to the model name). Additionally, for SceneTransformer–All, we include both the marginal and joint models (these models are the same when *training* on only the AV.)

IV. Results

A. Model sensitivity to non-causal perturbations

In order to understand model sensitivity on a per-example level, Figure 3 plots the perturbed versus original minADE across each perturbation for MultiPath++ (see the supplement for other models). For each perturbation type, we observe that the majority of examples show minimal change (i.e. are clustered around the $y{=}x$ axis), but a long tail of outlier examples experience a large change ($>1m$ in minADE). Among perturbation types, the model is most sensitive to RemoveCausal, which is expected since removing causal agents can change the correct ground-truth trajectory. Interestingly, models are significantly more robust to RemoveNoncausalEqual than RemoveNoncausal, which means removing more agents increases model sensitivity. When comparing RemoveCausal and RemoveNoncausalEqual, which controls for the number of agents removed, we see that the model is significantly more sensitive to removing causal agents than removing non-causal agents. Specifically, we found that across all models, the average Abs(Δ) is 0.1450 for RemoveCausal, 0.131 for RemoveNoncausal, 0.051 for RemoveNoncausalEqual, and 0.089 for RemoveStatic. In addition, we noticed that sensitivity to perturbations can lead to minADE improvements in some examples[2].

Comparing models. Focusing on RemoveNoncausal, in Table I, we evaluate each model and report the original minADE, perturbed minADE, Abs(Δ), the standard deviation of Abs(Δ), and $\frac{\text{Abs}(\Delta)}{\text{minADE}_{\text{Ori}}}$. SceneTransformer Marginal shows the lowest average absolute sensitivity, while MultiPath++–All shows the lowest sensitivity *relative* to original minADE. In general, Abs(Δ) decreases with the original minADE, but there is no clear relationship between relative Abs(Δ) and minADE. Unexpectedly, the marginal SceneTransformer is more robust than the joint (we hypothesize that jointly modeling agents in the scene causes the model to pay more attention to non-causal agents).

B. Sensitivity via an IoU-based trajectory set metric

To directly measure the magnitude of model output changes with and without perturbation, we use an IoU based metric (defined in Section III-D) to compare model sensitivity to different perturbations. The results of three AV-only models under perturbations RemoveCausal, RemoveNoncausal and RemoveNoncausalEqual are shown in Figure 4. We find that models are least sensitive to RemoveNoncausalEqual, and much more sensitive to RemoveCausal and RemoveNoncausal. This is consistent with our finding in Section IV-A, indicating the model is more sensitive to large perturbations since there are more non-causal agents than causal ones in most examples.

C. Data augmentation in training improves robustness

We experiment with two types of data augmentation: 1) data augmentations that use a heuristic definition of non-causal agents, such as randomly dropping any static context agent[3], and 2) robustness-targeted data augmentations that directly drop only non-causal agents using a labeled portion of the validation dataset.

Heuristic data augmentation. The benefit of using a heuristic definition of non-causal agents for data augmentations is

[2]This can happen for instance when removing a causal agent matches their strong reaction, or when removing a non-causal agent helps the model be less confused.

[3]Context agents are agents for which no prediction is required in the WOMD leaderboard.

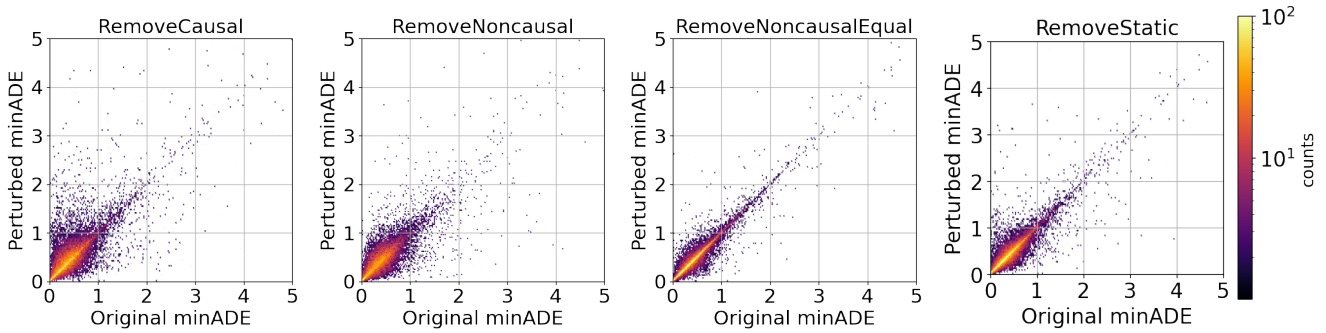
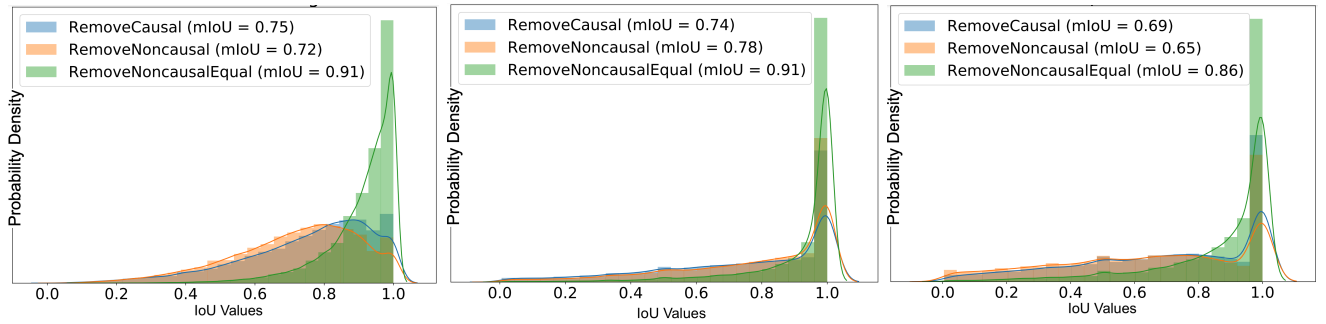


Fig. 3: **Model sensitivity to different perturbation types.** We plot the per-example perturbed versus original minADE for all perturbations for the MP++ model. The example frequency is shown with a log color scale where yellow is the most frequent. The majority of examples show minimal change from the perturbation and lie close to the $y{=}x$ axis. However, across all perturbation types, even for RemoveNoncausal, *there is a long tail of examples that show relatively large change in minADE*($>1m$). Moreover, the minADE can change in either directions. Comparing RemoveNoncausal and RemoveNoncausalEqual indicates that *the model is more sensitive to removing larger numbers of non-causal agents).*

Model comparison, RemoveNoncausal, minADE, Δ = Pert - Orig					
Model	Original(Orig)	Perturbed(Pert)	Abs(Δ)	Std. Abs(Δ)	$\frac{\text{Abs}(\Delta)}{\text{minADE}_{\text{Ori}}}$ (%)
MultiPath++	0.376	0.395	0.141	$\pm$0.21	37.5%
SceneTransformer Marginal	0.250	0.265	0.067	$\pm$0.12	26.8%
Wayformer	0.393	0.406	0.101	$\pm$0.16	25.7%
MultiPath++-All	0.900	0.945	0.226	$\pm$0.32	25.1%
SceneTransformer-All Joint	0.493	0.504	0.170	$\pm$0.26	34.5%
SceneTransformer-All Marginal	0.305	0.328	0.081	$\pm$0.14	26.6%

TABLE I: **Model sensitivity to the RemoveNoncausal perturbation**. The SceneTransformer Marginal model shows the lowest average absolute sensitivity to the perturbation, while the MultiPath++–All model shows the lowest sensitivity *relative* to original minADE. Original and Perturbed are the average minADE across the whole dataset. Abs(Δ) is the average per-example absolute difference between perturbed and original minADE.



(a) Scene Transformer Marginal (b) Wayformer (c) Multipath++

Fig. 4: Density distribution of the per-scene trajectory set IoU values for AV-only models under perturbations (RemoveCausal, RemoveNoncausal, and RemoveNoncausalEqual). Larger IoU means better robustness. We can see that models are least sensitive to RemoveNoncausalEqual, and more sensitive to RemoveCausal and RemoveNoncausal.

that it can be applied without collecting causal labels. We implement two types of heuristic-based data augmentation in the training set of WOMD: Drop Context (randomly dropping context agents) as a baseline, and Drop Static Context (randomly dropping static context agents). We use the MultiPath++–All model and we set the probability of dropping an agent to 0.1 (the best one among 0.1, 0.5, and 0.8). Table II summarizes the results of the re-trained model for the RemoveNoncausal perturbation. Models with data augmentation show less sensitivity to the perturbations. In particular, Drop Static Context shows a significant improvement in minADE and Abs(Δ) over Drop Context. We hypothesize that Drop Static Context does better because the static context agents are less likely to be causal. Overall, the results for Drop Static Context imply that dropping non-causal agents via data augmentation in training can improve model robustness to such perturbations.

Non-causal data augmentations. Motivated by the above results that dropping static context agents improves model robustness, we further explore using non-causal perturbations as a data augmentation strategy during training. We randomly sample approximately 70% of the original validation dataset (i.e. 30k scenes), perturb multiple copies of them via the

Heuristic Data Augs, RemoveNoncausal, minADE, Δ=Pert-Orig			
Model	MP++-All	MP++-All Drop Context	MP++-All Drop Static Context
Original (Orig)	0.900	0.948	0.819
Perturbed (Pert)	0.945	0.988	0.837
Abs(Δ)	0.226	0.209	0.183
Std. Abs(Δ)	$\pm$0.32	$\pm$0.31	$\pm$0.26
$\frac{\text{Abs}(\Delta)}{\text{minADE}_{\text{Ori}}}$ (%)	25.1%	22.0%	22.3%

TABLE II: **Heuristic data augmentations.** We compare the MP++-All baseline model to the same model trained with either dropping context agents or dropping static context agents, finding that data augmentations that drop agents that are more likely to be non-causal can improve robustness.

causal labels, and add the perturbed versions into the training dataset. We leave the remaining 30% of the validation set as a holdout for evaluation. We then train a baseline model on the new training dataset as well as a model that randomly drops non-causal agents (when possible) with a probability of 0.1. The results are given in Table III. We see that similarly dropping non-causal agents helps improve minADE as well as model robustness.

Noncausal Augs, RemoveNoncausal, minADE, Δ=Pert-Orig		
Model	MP++ Baseline	MP++ Drop Non-causal
Original (Orig)	0.395	0.373
Perturbed (Pert)	0.408	0.389
Abs(Δ)	0.150	0.138
Std. Abs(Δ)	$\pm$0.226	$\pm$0.193
$\frac{\text{Abs}(\Delta)}{\text{minADE}_{\text{Ori}}}$ (%)	38%	37.0%

TABLE III: **Noncausal data augmentation.** We fold a portion of the WOMD validation dataset into the original training dataset and apply data augmentations that drop non-causal agents. On hold-out validation data, model robustness improves across all three Abs(Δ) metrics.

D. Larger dataset size improves model robustness

We also evaluate the impact of training size on model robustness. We randomly select 10%, 20%, 50%, 80% of the training set and train separate models for robustness evaluation. As shown in Table IV, with training size increases, both the model performance and robustness improve. Interestingly, in Table I with varying model architectures, we found that the model with the lowest minADE did not always have the best relative robustness. Here, we see a strong trend: for a fixed model architecture, lowering the minADE by increasing the training data leads to better robustness.

Train %, minADE, RemoveNoncausal, Δ=Pert-Orig					
Train	Original	Perturbed	Abs(Δ)	Std. Abs(Δ)	$\frac{\text{Abs}(\Delta)}{\text{Ori}}$
10%	1.222	1.309	0.448	$\pm$0.69	37.0%
20%	1.039	1.117	0.386	$\pm$0.53	37.2%
50%	0.947	0.996	0.266	$\pm$0.45	28.0%
80%	0.901	0.925	0.236	$\pm$0.32	26.2%
100%	0.900	0.945	0.226	$\pm$0.32	25.1%

TABLE IV: **Increasing training data improves robustness.**

V. Discussion

We now discuss a few hypotheses and supporting evidence for why models are not robust to the non-causal perturbations.

Overfitting. Models may fail to generalize to the non-causal perturbations because they overfit to spurious correlations in the training data (i.e. features that correlate with certain ground truth but fail to generalize). In our experiments, we observe that overfitted models are *more sensitive* to the non-causal perturbations (see the supplement for details).

Distribution shift. Models may fail to generalize to perturbations that are significantly different from training data. In our results, we observe that the more non-causal agents we remove, the less robust models are. Perhaps scenes with only a few agents are relatively rare in the training dataset and the model does not generalize well to such distribution shift. By evaluating on the perturbations, we essentially expose the model to rare scenarios.

The above two hypotheses are verified by the experimental results that data augmentation and increasing dataset size can improve robustness.

Over-reliance on agents instead of roadmap. A third possible reason that models fail to generalize is that they utilize the non-causal agents to infer the drivable areas instead of using the mapping information (we serve high-definition maps and traffic control signals as input features for all models). Our evidence comes from visualizing examples where dropping non-causal agents creates predictions that disobey the roadgraph rules (see details in the supplement). Although such leverage of information might be acceptable, we believe it is necessary to establish an evaluation benchmark which can detect such phenomenon so that further model improvements can be inspired.

VI. Conclusions

We establish a benchmark and metrics for evaluating the robustness of several state-of-the-art models for trajectory prediction in autonomous driving. We find that many state-of-the-art models (with different model architectures and coordinate systems) show significant levels of sensitivity to perturbations that remove non-causal agents, with higher sensitivity when removing a greater number of them. While most examples show minimal change in minADE (≤ 0.1 m), there is a long tail of examples that can have large changes ($\geq$ 1m and sometimes up to 8m). In addition, removing either causal or non-causal agents can cause some examples to improve their minADE. We also find that increasing dataset size and data augmentation can help improve the model robustness. Overall, our results indicate that current machine learning models for trajectory prediction may not be reliable enough on their own, and future work is needed to make them more robust to non-causal perturbations or complement them with other techniques to make a safe system. Finally, we publish the causal agent labels as complementary attributes to WOMD to aid future researchers in building more robust models that have a better understanding of causal relations.

REFERENCES

[1] S. Casas, C. Gulino, R. Liao, and R. Urtasun, "Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data," in *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*. IEEE, 2020, pp. 9491–9497.

[2] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," in *Conference on Robot Learning*, 2019.

[3] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 2090–2096.

[4] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[5] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov, *et al.*, "Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction," *arXiv preprint arXiv:2111.14973*, 2021.

[6] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "Precog: Prediction conditioned on goals in visual multi-agent settings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2821–2830.

[7] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 336–345.

[8] J. Hong, B. Sapp, and J. Philbin, "Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions," in *CVPR*, 2019.

[9] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," *arXiv preprint arXiv:2001.03093*, 2020.

[10] J. Mercat, T. Gilles, N. Zoghby, G. Sandou, D. Beauvois, and G. Gil, "Multi-head attention for joint multi-modal vehicle motion forecasting," in *IEEE International Conference on Robotics and Automation*, 2020.

[11] S. Khandelwal, W. Qi, J. Singh, A. Hartnett, and D. Ramanan, "What-if motion prediction for autonomous driving," *arXiv preprint arXiv:2008.10587*, 2020.

[12] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," in *European Conference on Computer Vision (ECCV)*, 2020.

[13] O. Makansi, Ö. Cicek, Y. Marrakchi, and T. Brox, "On exposing the challenging long tail in future prediction of traffic actors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 147–13 157.

[14] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, *et al.*, "Scene transformer: A unified multi-task model for behavior prediction and planning," *arXiv e-prints*, pp. arXiv–2106, 2021.

[15] K. S. Refaat, K. Ding, N. Ponomareva, and S. Ross, "Agent prioritization for autonomous navigation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2060–2067.

[16] E. V. Tolstaya, R. Mahjourian, C. Downey, B. Varadarajan, B. Sapp, and D. Anguelov, "Identifying driver interactions via conditional behavior prediction," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3473–3479, 2021.

[17] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," *arXiv preprint arXiv:2207.05844*, 2022.

[18] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning*, 2019, pp. 5389–5400.

[19] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, 2018.

[20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2013.

[21] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations (ICLR)*, 2019.

[22] K. Gu, B. Yang, J. Ngiam, Q. Le, and J. Shlens, "Using videos to evaluate image model robustness," *arXiv preprint arXiv:1904.10076*, 2019.

[23] V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht, and L. Schmidt, "Do image classifiers generalize across time?" *arXiv preprint arXiv:1906.02168*, 2019.

[24] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 583–18 599, 2020.

[25] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[26] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1802–1811.

[27] A. Fawzi and P. Frossard, "Manitest: Are classifiers really invariant?" *arXiv preprint arXiv:1507.06535*, 2015.

[28] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[29] K. D. Dhole, V. Gangal, S. Gehrmann, A. Gupta, Z. Li, S. Mahamood, A. Mahendiran, S. Mille, A. Srivastava, S. Tan, *et al.*, "Nl-augmenter: A framework for task-sensitive natural language augmentation," *arXiv preprint arXiv:2112.02721*, 2021.

[30] A. Bera, S. Kim, T. Randhavane, S. Pratapa, and D. Manocha, "Glmp-realtime pedestrian path prediction using global and local movement patterns," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5528–5535.

[31] Y. Han, R. Tse, and M. Campbell, "Pedestrian motion model using non-parametric trajectory clustering and discrete transition points," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2614–2621, 2019.

[32] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll, "The simpler the better: Constant velocity for pedestrian motion prediction," *arXiv preprint arXiv:1903.07933*, vol. 5, no. 6, p. 7, 2019.

[33] L. Sun, X. Jia, and A. D. Dragan, "On complementing end-to-end human behavior predictors with planning," in *Proceedings of the Robotics: Science and Systems*, 2021.

[34] Y. Cao, D. Xu, X. Weng, Z. Mao, A. Anandkumar, C. Xiao, and M. Pavone, "Robust trajectory prediction against adversarial attacks," *arXiv preprint arXiv:2208.00094*, 2022.

[35] Y. Cao, C. Xiao, A. Anandkumar, D. Xu, and M. Pavone, "Advdo: Realistic adversarial attacks for trajectory prediction," in *European Conference on Computer Vision*. Springer, 2022, pp. 36–52.

[36] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao, "On adversarial robustness of trajectory prediction for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 159–15 168.

[37] S. Saadatnejad, M. Bahari, P. Khorsandi, M. Saneian, S.-M. Moosavi-Dezfooli, and A. Alahi, "Are socially-aware trajectory prediction models really socially-aware?" *Transportation research part C: emerging technologies*, vol. 141, p. 103705, 2022.

[38] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "Advsim: Generating safety-critical scenarios for self-driving vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9909–9918.

[39] E. Aksoy, A. Yazıcı, and M. Kasap, "See, attend and brake: An attention-based saliency map prediction model for end-to-end driving," *arXiv preprint arXiv:2002.11020*, 2020.

[40] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7699–7707.