Road Signs that Know Your Next Move:
A New Perspective for Adversarial Attacks on Autonomous Systems

Meriel Stein David Shriver Sebastian Elbaum

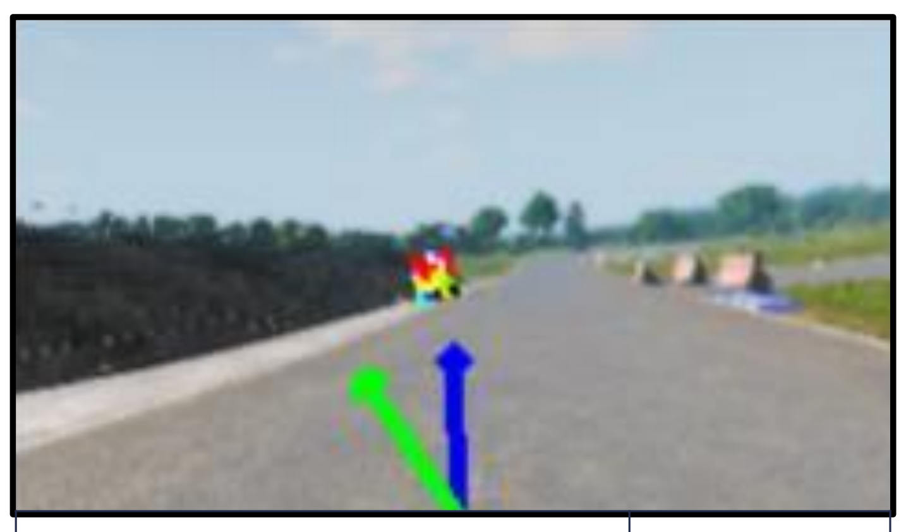
LESS LAB LEADING ENGINEERING FOR SAFE SOFTWARE

University of Virginia

Problem

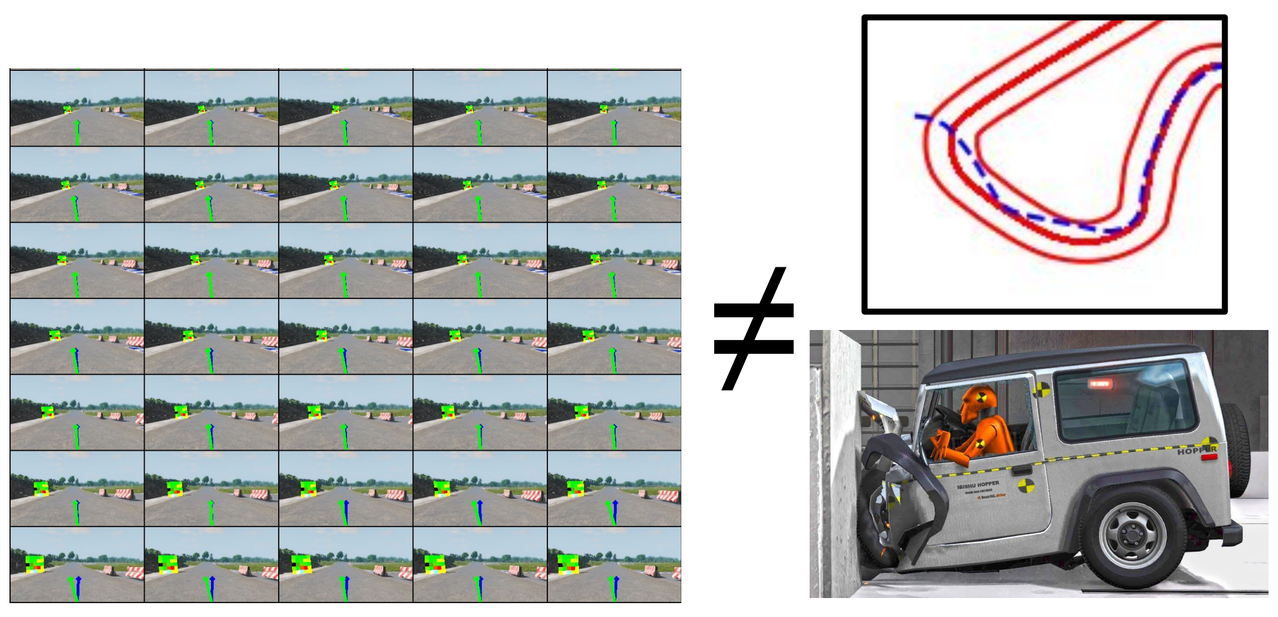
Adversarial testing tends to focus on DNNs in isolation, to the exclusion of system state and system behaviors.

Single-image error can't guarantee misbehavior or generalization of the perturbation to other images.



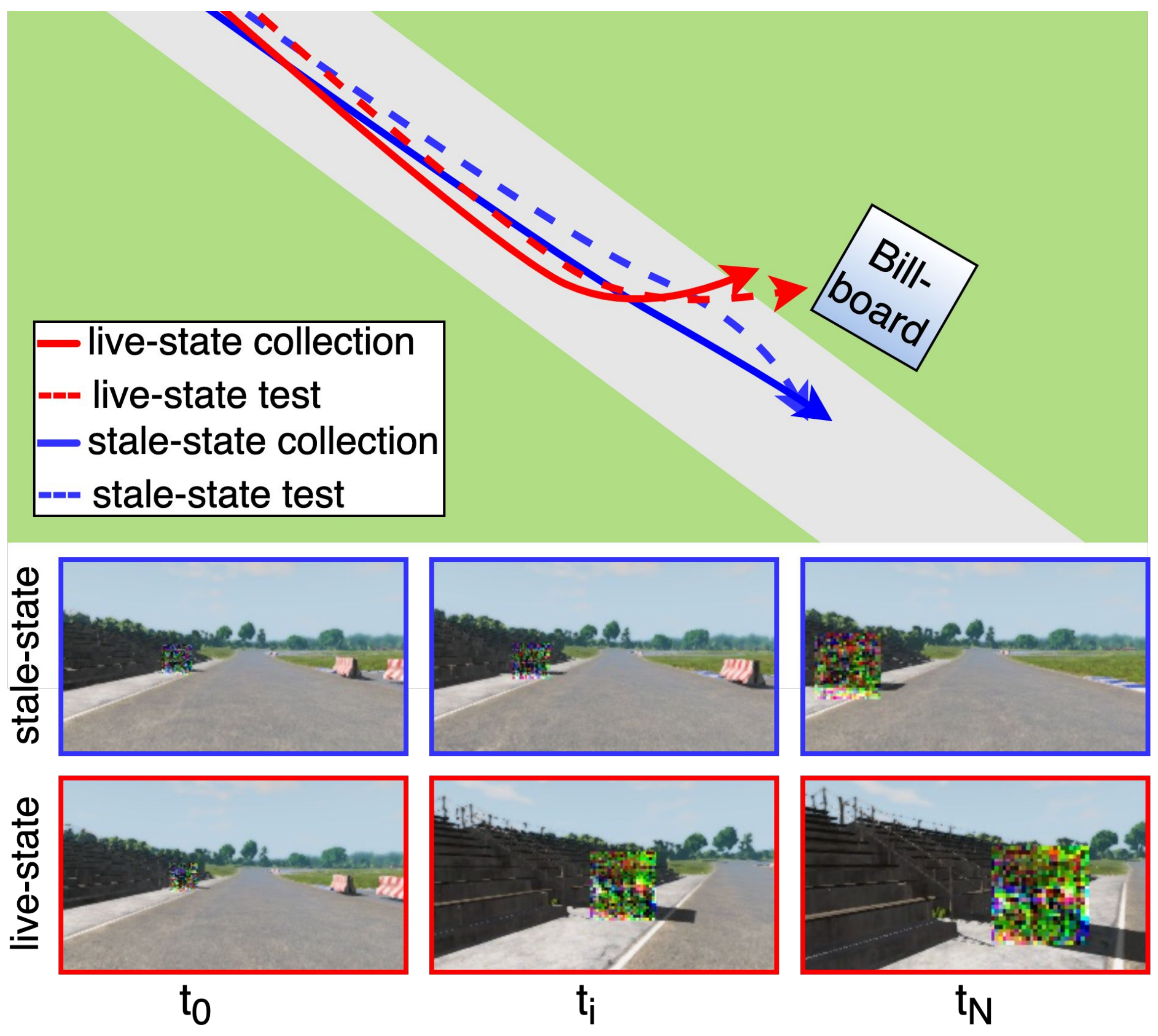
Original output:	0.013
Adversarial output:	0.602

Perturbations over a sequence of images do not account for the effect of that perturbation on the system.



Insight

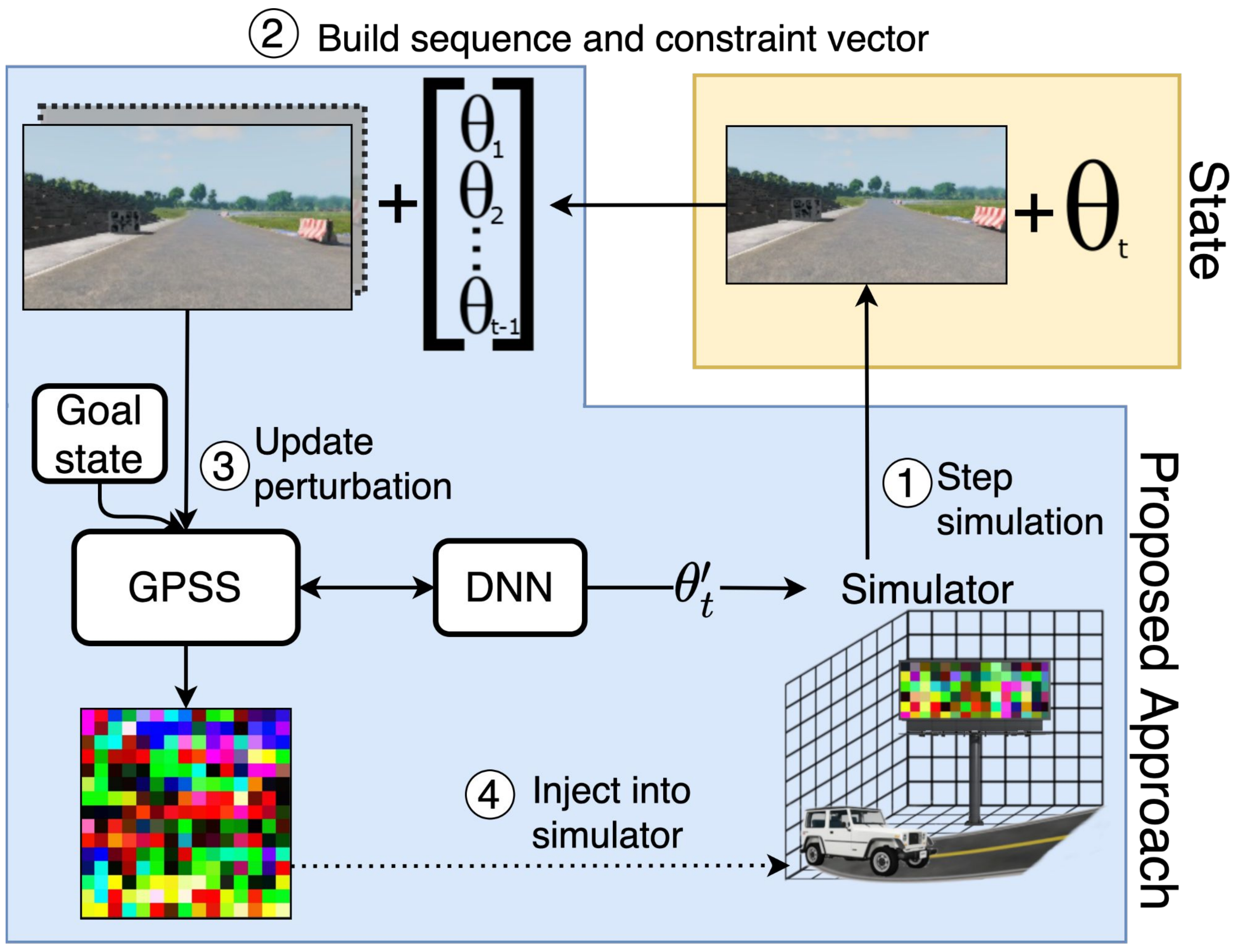
The way an environmentally-situated adversarial perturbation is sensed & processed by the system depends on system state.



Solution

Iterative Perturbation with Sim-in-the-Loop

1. The simulator captures the system live-state.
2. This state is added to the sequence of images and input constraints.
3. Generator of Perturbation over a State Sequence (GPSS) uses the updated sequence, DNN, and a goal state to update the perturbation to pursue that goal state in the next timestep.
4. The generated perturbation is injected into the simulator & the process is repeated.



Generator of Perturbation over a State Sequence (GPSS)

Property 1: A system input influenced by the perturbation is bounded by the system states reachable from the current live state. →

$$\theta_t = system.state.bound(\theta'_t)(\text{where } \theta'_t = DNN(img_t + pert_t))$$

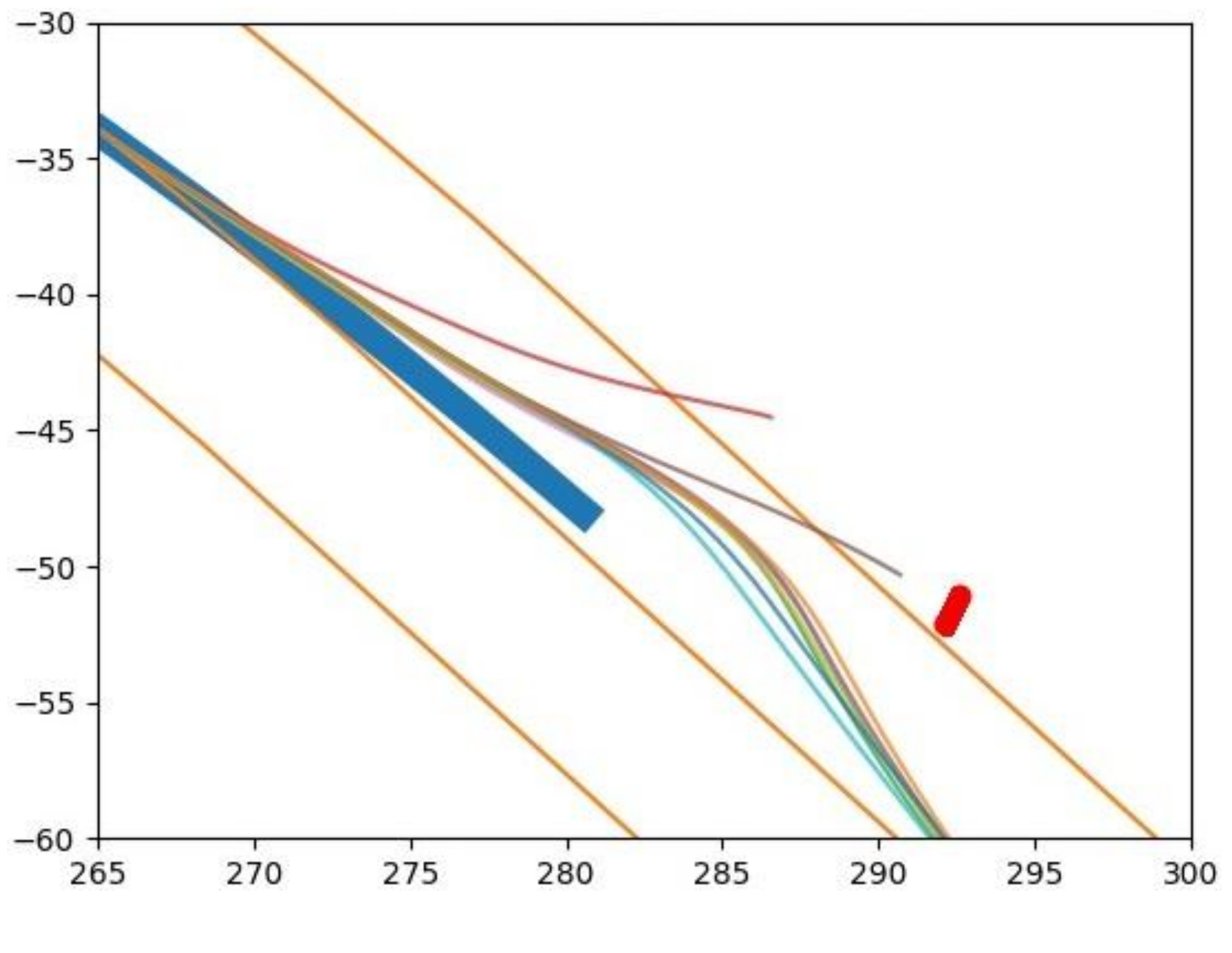
Property 2: The perturbation must maximize state change toward the goal state. →

$$\underset{pert_t \rightarrow goal}{argmax} \ (\mathcal{L}(DNN(img_t)), (DNN(img_t + pert_t)))$$

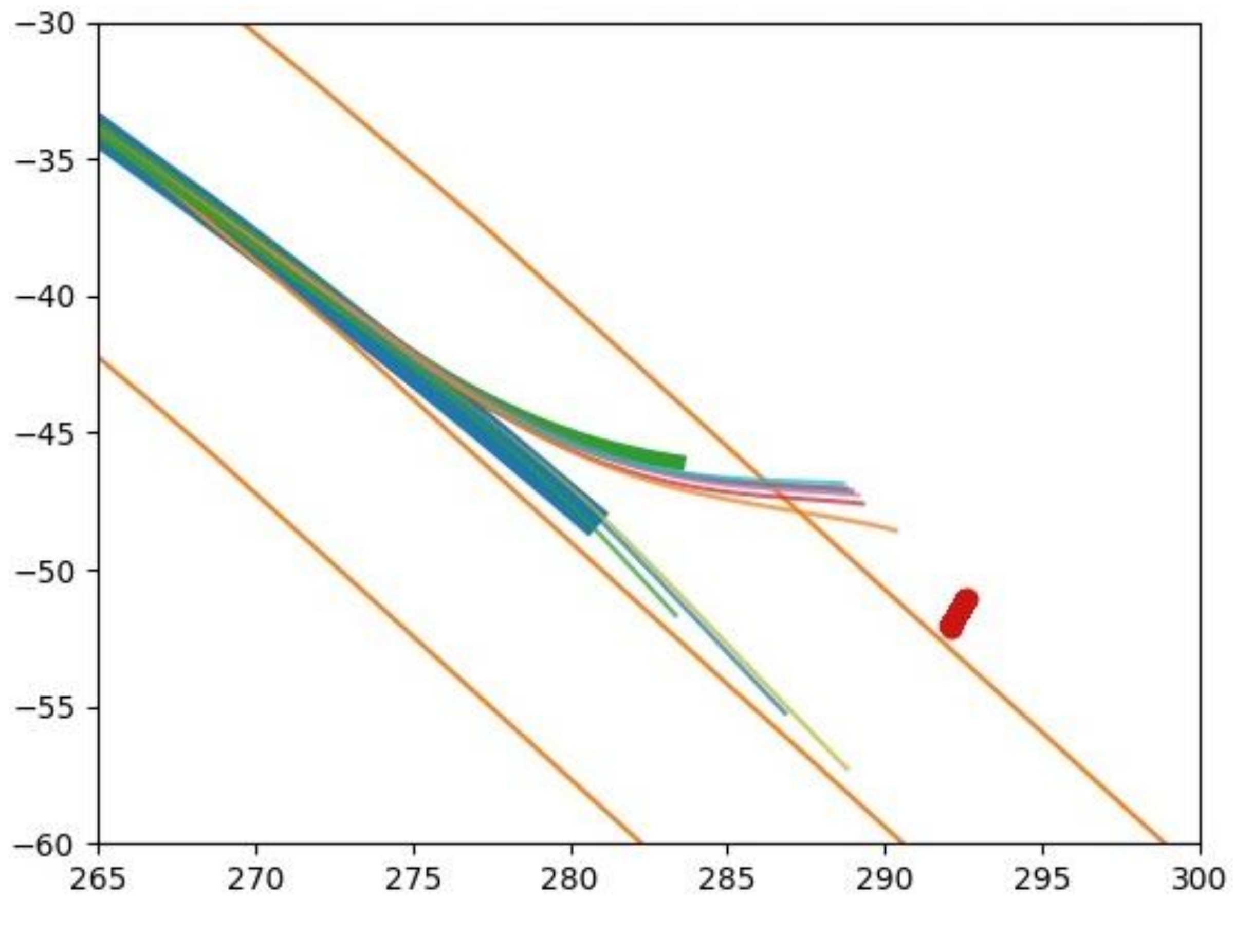
Property 3: The resulting perturbation must be consistent over time and space. →

$$\underset{pert_N}{argmin}(\sum_t \mathcal{L}(DNN(img_t + pert_N), DNN(img_t + pert_t)))$$

Results



Average case run for Deepbillboard.



Average case run for GPSS.

Left:
Effect of perturbation for Deepbillboard and GPSS. Road is defined by orange lines, billboards are in red, the thick blue line is the normal trajectory, and the thick green line is the collection trajectory.

Below:
Physical & DNN measures of perturbation effects for best-performing perturbations of each technique over 100 runs.

	DeepBillboard	GPSS
Avg. dist. from expected traj.	1.101	1.812
Mean angle error	0.007	0.117
Crash frequency	0.04	0.99

This work has been funded in part by NSF.