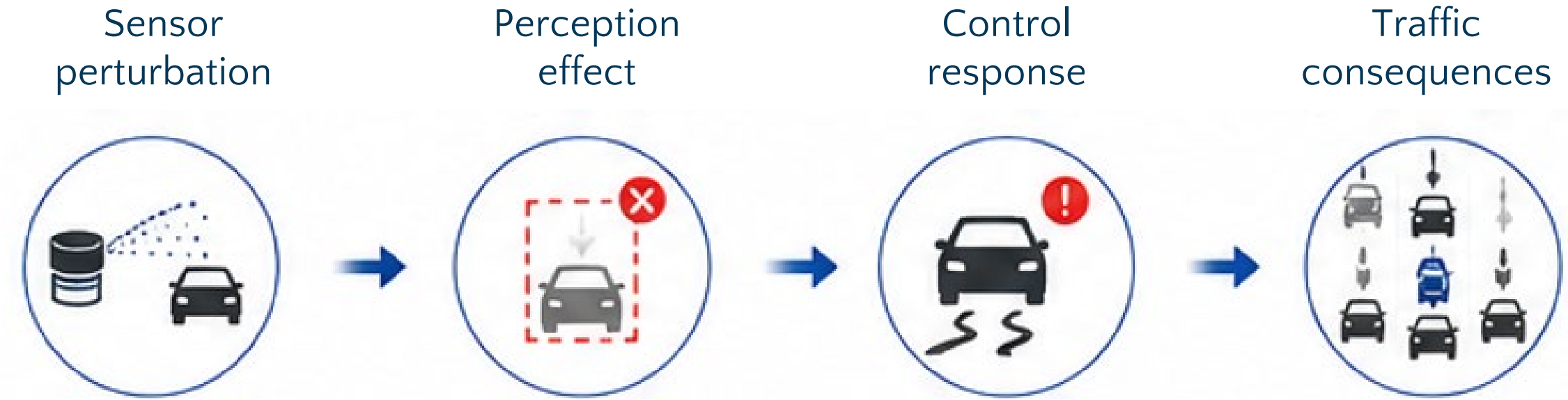
Agentic AI for Multi-Modal Perception Integrity: Threat Reasoning and Response in Cyber-Physical Transportation Systems

Qing Tang, Department of Civil and Environmental Engineering, ODU
Ashley Y. Gao, Department of Computer Science, William & Mary

October 30, 2026

Why perception integrity matters

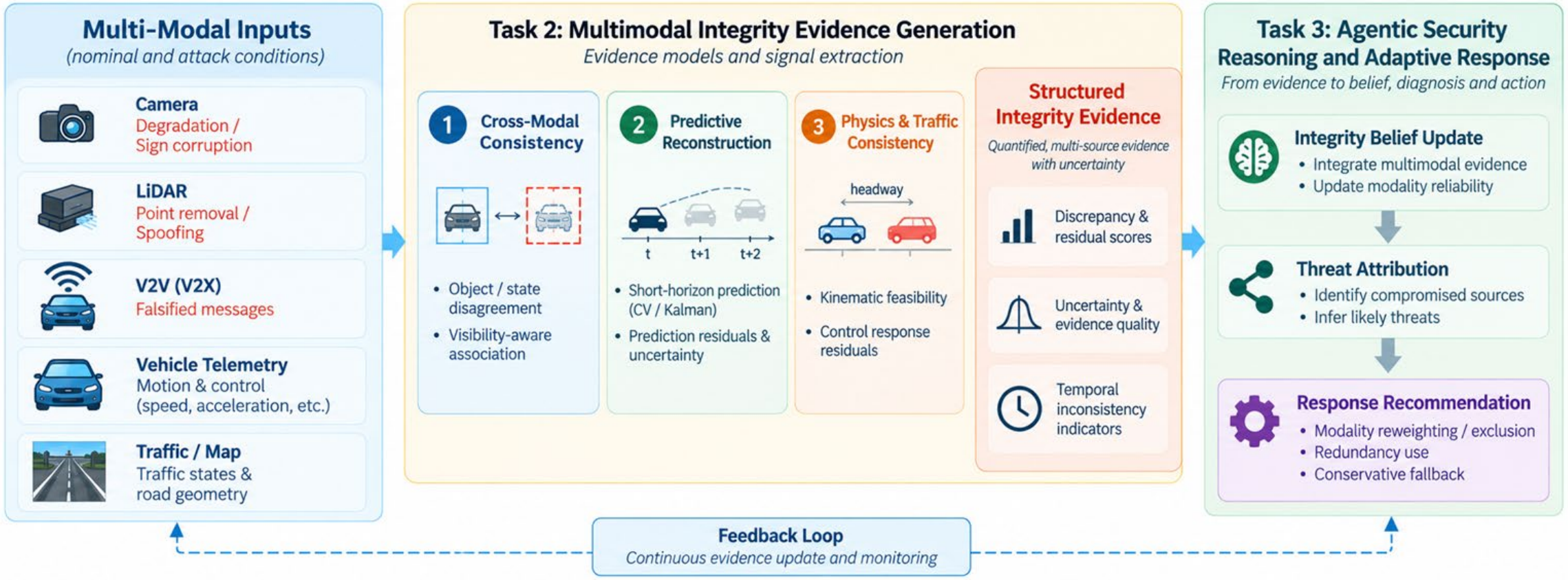
THREAT PROPAGATION FROM THE SENSOR TO TRAFFIC



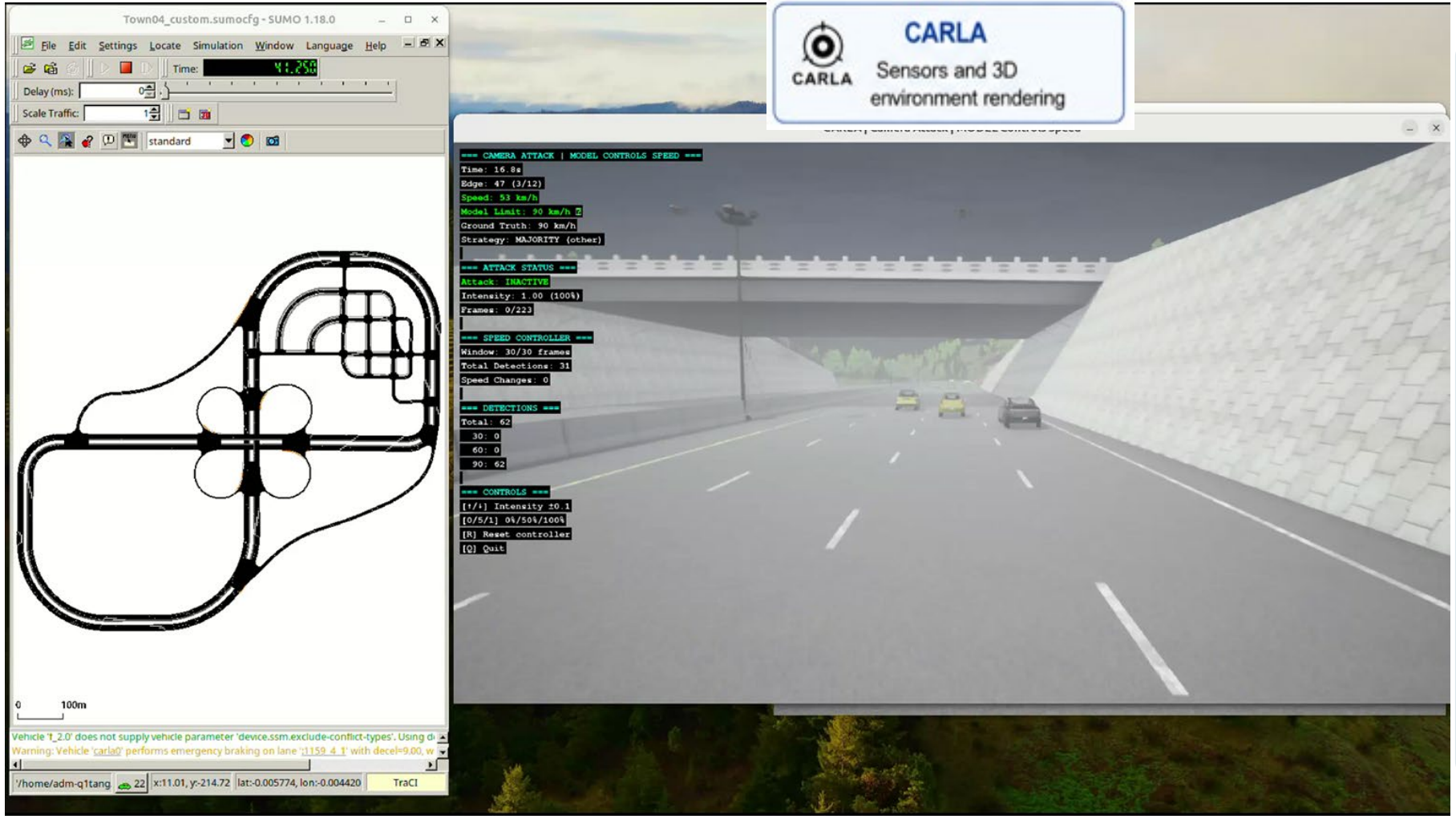
Project Objectives	01	02	03
	Threat & integrity dataset	Structured integrity evidence	Agent reasoning & response

Core question: Can an autonomous agent determine when a fused world model is no longer trustworthy?

Evidence-driven agentic security framework



CARLA–SUMO cyber-physical testbed



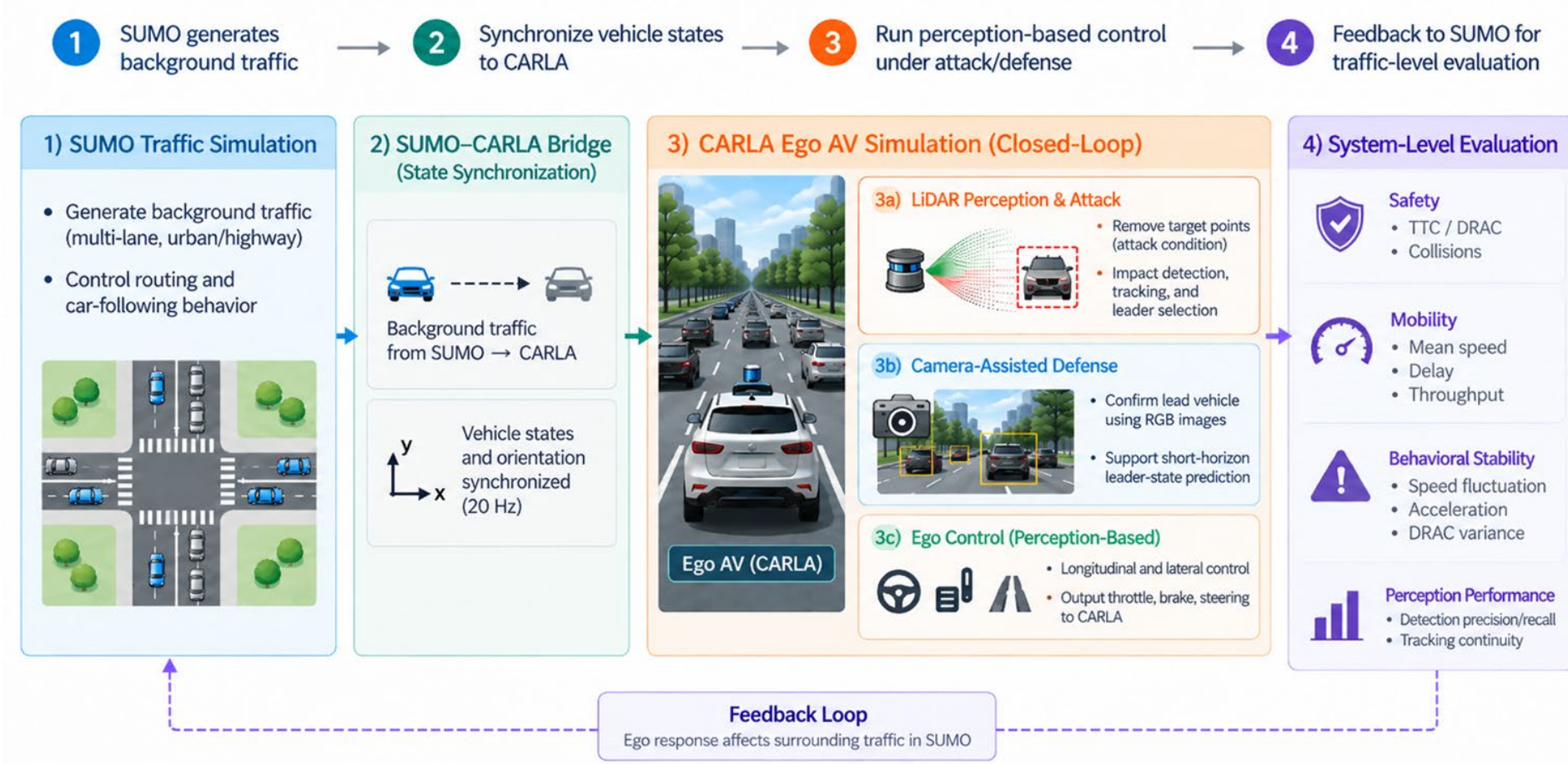
Link

CAPABILITIES

- SUMO: background traffic, car-following and network impacts
- CARLA: high-fidelity sensing, ego-vehicle dynamics and control
- Attack injection: LiDAR returns and camera perception channels
- Ego responses influence surrounding traffic through synchronization

LiDAR target loss under cyberattack

ESTABLISHED ATTACK-AND-DEFENSE EXPERIMENT



ATTACK MECHANISM

Remove returns from the selected target vehicle

Disrupt object availability, tracking and leader selection

Evaluate vehicle motion, conflicts, braking and delay

Target-specific LiDAR removal

LiDAR perception-driven closed loop

CARLA LIDAR → DETECTION → VEHICLE CONTROL → SUMO FEEDBACK

1
CARLA LiDAR
Frame-aligned point clouds

→

2
3D perception
Detection + tracking

→

3
Leader tracking and state estimation
In-lane leader, gap & speed

→

4
Controller
Steering, throttle and braking

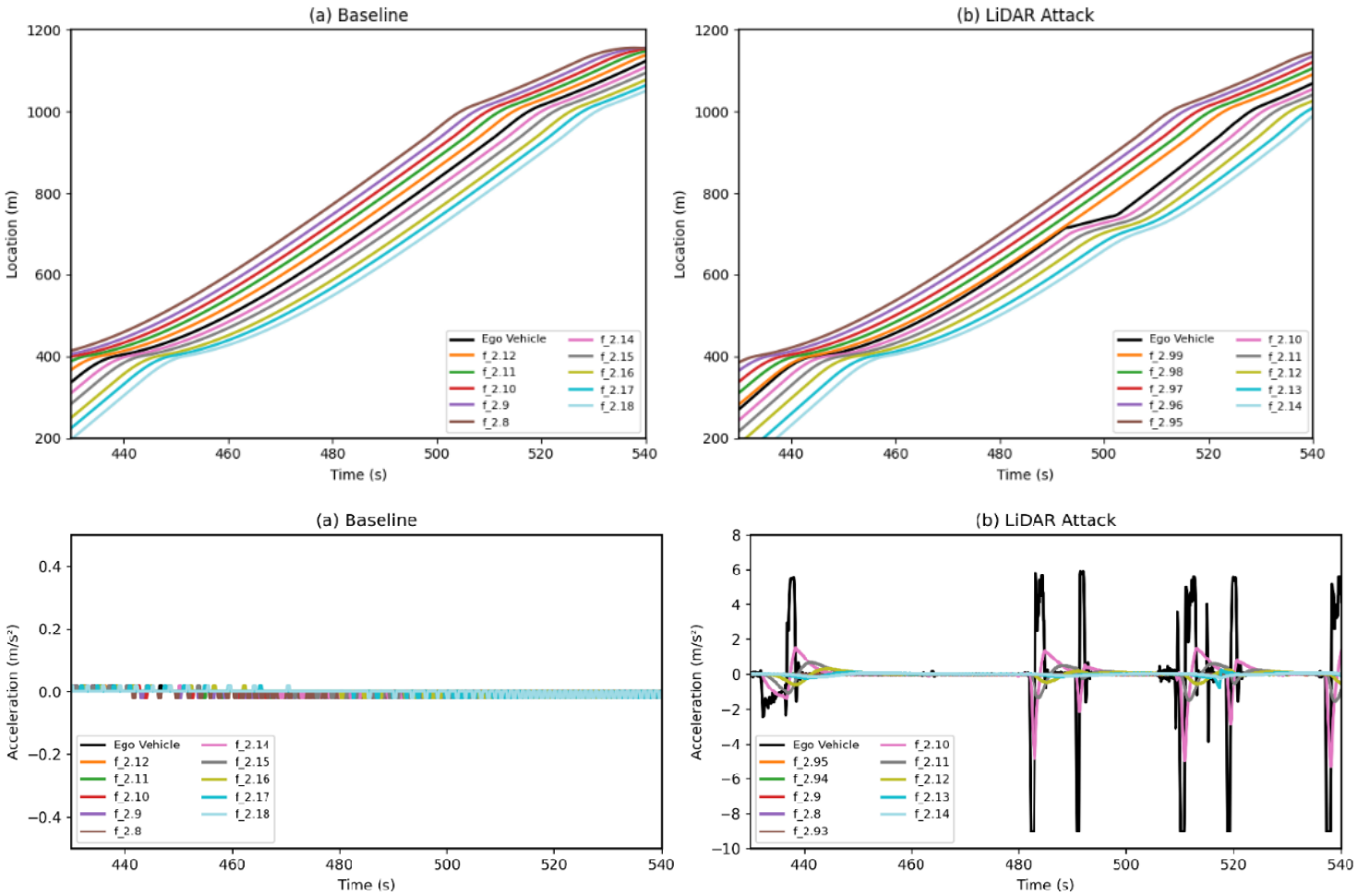
→

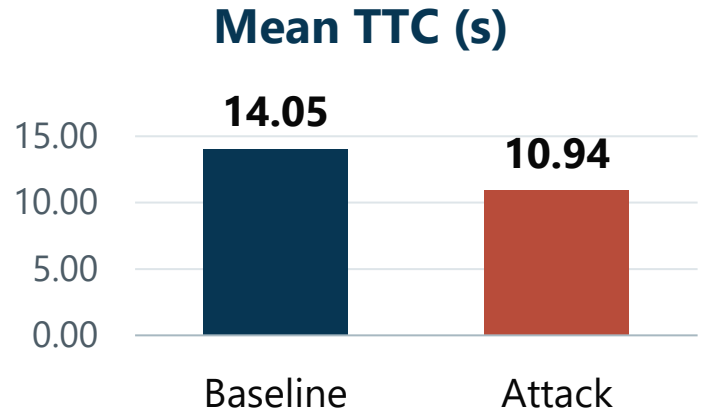
5
CARLA + SUMO
Ego response + traffic coupling

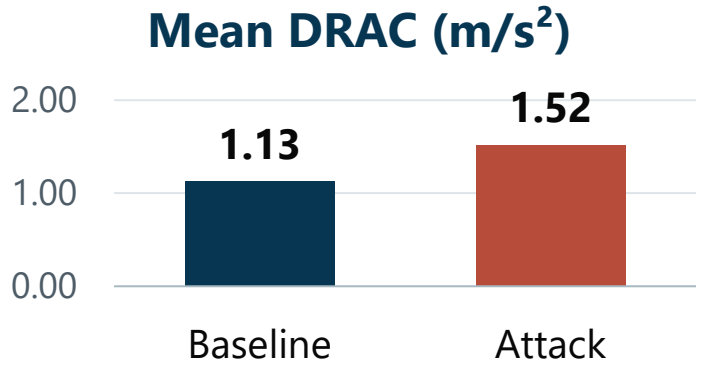
Synchronized cyber-physical loop

One simulation clock aligns each LiDAR frame, controller decision, applied actuation and surrounding traffic state.

Persistent LiDAR attack: safety consequences





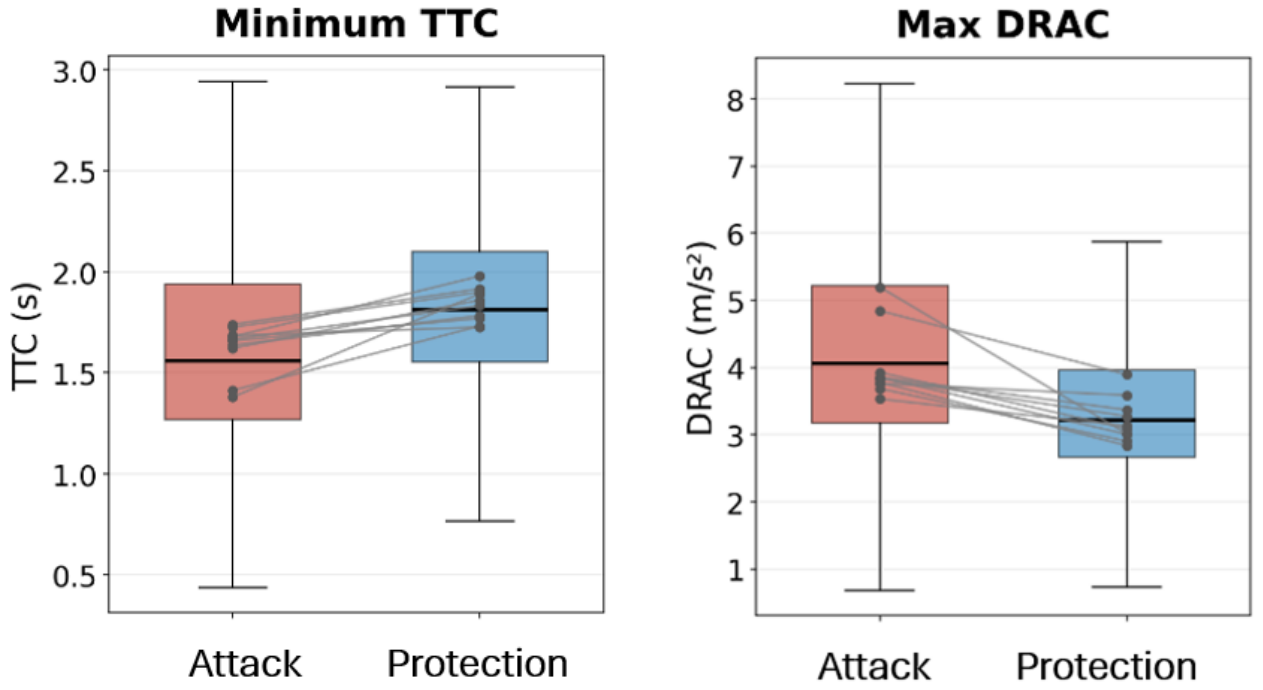


Attack-induced perception loss can generate abrupt braking/acceleration and worsen surrogate safety measures.

Camera-assisted protection

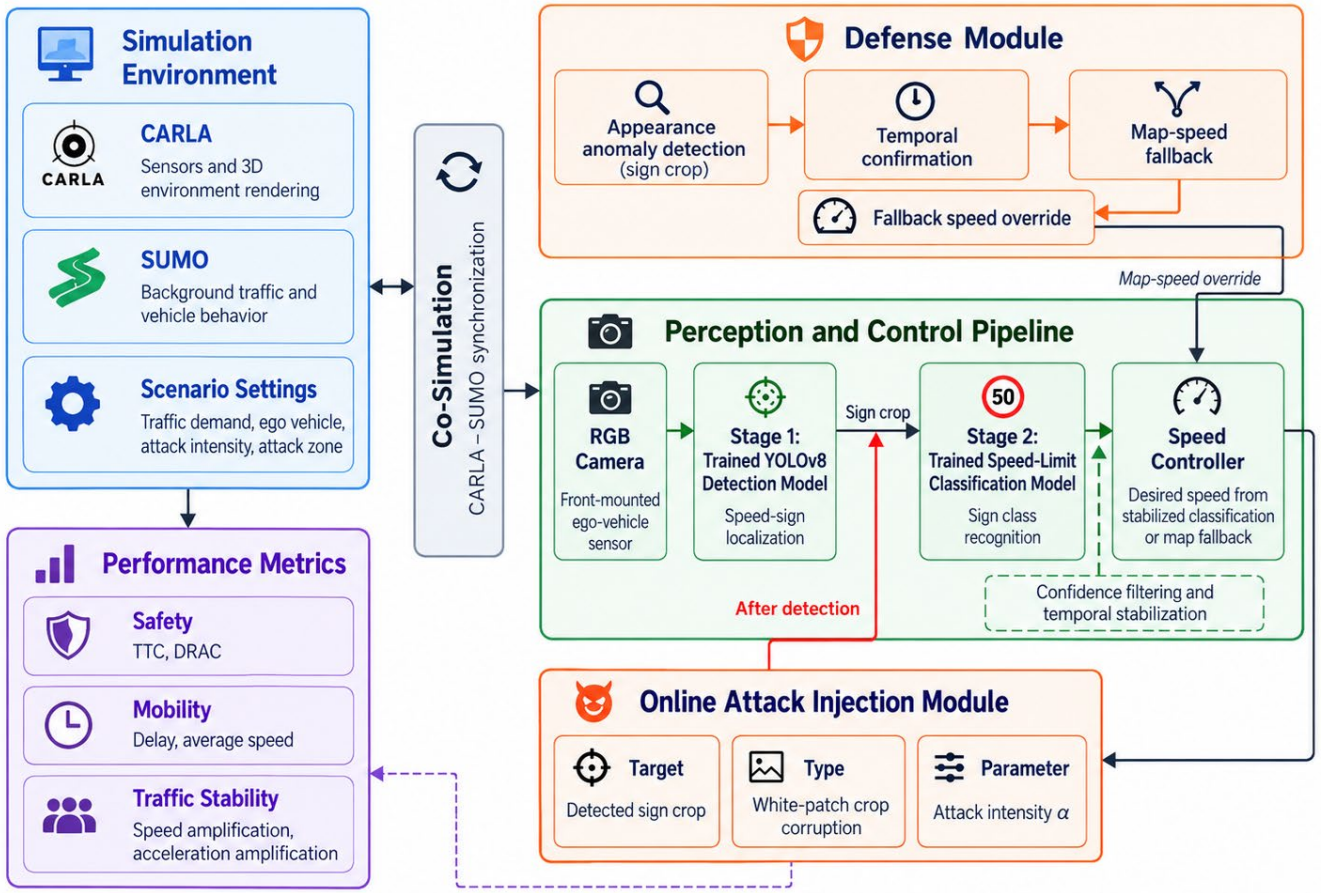
403 matched episodes with 2s of target loss

Episode outcome	Attack Only	With Protection
Episodes with min TTC < 1.5 s	45.4%	21.1%
Episodes with collisions	7 / 403	0 / 403
Median max deceleration(m/s²)	7.57	8.72



Protection reduced low-TTC episodes and observed collisions, but required stronger braking.

Camera speed-sign attacks and protection



ATTACK

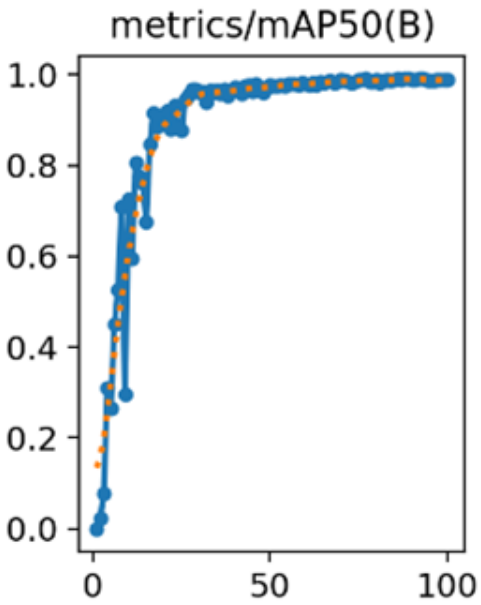
- White patches corrupt detected speed-sign crops
- Misclassified speed limits change desired speed

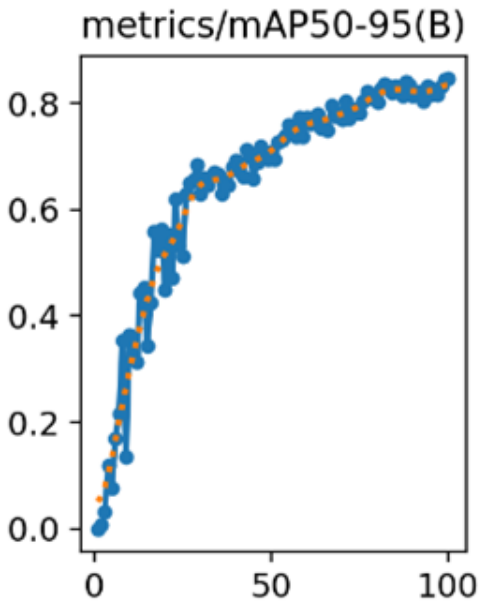
PROTECTION

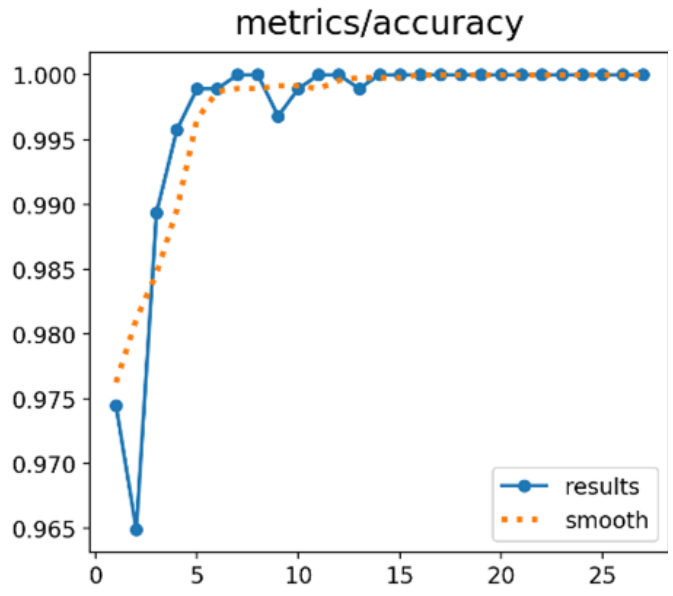
- Anomaly-triggered temporal confirmation
- Map-based speed fallback when vision is unreliable

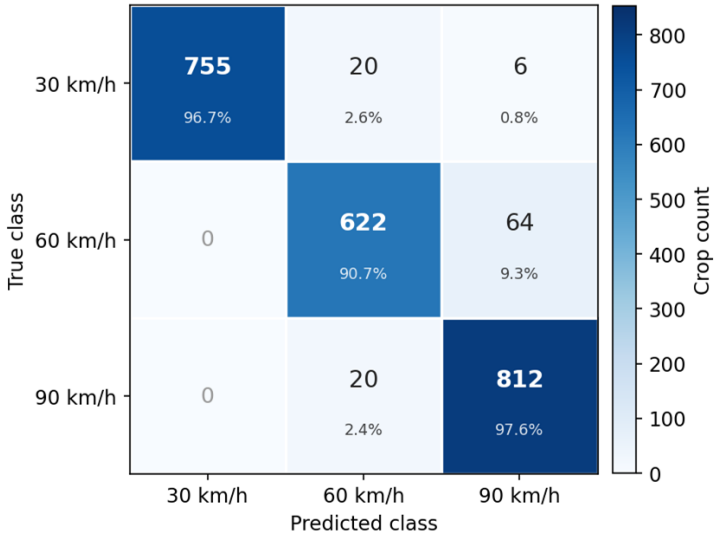
Offline model development and online attack/defense experiments

Speed-sign perception model training results









- **Detection model**: mAP50 approaches 1.0, and mAP50–95 remains high, indicating reliable speed-sign localization.
- **Classification model:** accuracy quickly reaches nearly 1.0, indicating stable speed-limit recognition.

Speed-sign corruption affects surrounding traffic

Attacked segment, 10 paired runs, 600 s/run, 1,000 veh/h/lane

Measure	Baseline	Attack	With Defense
Mean speed (m/s) ↑	14.42	**10.94**	**14.58**
Average delay (s/veh) ↓	35.60	**57.92**	**34.35**
Mean TTC (s) ↑	14.0	12.2	**14.0**
Spacing-error amplification ↓	1.12	**3.31**	**1.20**

Under the tested corruption, defense brings speed and delay close to baseline.

Task 1: Multimodal threat and integrity dataset

Data Collection and Attack-Experiments Readiness

Matched route, traffic and controller; vary the cyberattack.

NOMINAL | seed k

ATTACK | same seed k

V2V logging prototype

300 m range: vehicle IDs, position, speed and acceleration.

Scenario & attack labels

Run ID, seed, weather, density

Target ID, attack intensity and timing

Paired run and route metadata

Multimodal observations

LiDAR points + RGB frames

Object detection/track/gap/speed

V2V claimed vehicle states

Vehicle & traffic consequences

Commanded vs achieved acceleration

Throttle, brake, TTC, lane offset

Emergency braking, collision, delay

Task 2: Multimodal integrity evidence generation

EVIDENCE FAMILY	EARLY SIGNALS / METHODS IN USE	ACTIVE EVIDENCE-MODEL DEVELOPMENT
Cross-modal & map consistency	Camera confirms target during LiDAR loss; sign-crop anomaly/map checks; V2V message logging	Align Camera/LiDAR/V2V by target and time; assess visibility-aware presence and state conflict.
Predictive reconstruction	Short-horizon constant-velocity target prediction and TTC-based protection.	Kalman prediction; uncertainty-normalized residuals; missing-track persistence.
Physics & traffic feasibility	Synchronized trajectories, calibrated controller, gap/speed and TTC/DRAC metrics.	Gap-motion and actuation residuals; nominal bounds; TTC/DRAC as risk context.

Evidence output: Entity/time · source disagreement · prediction/physics residuals · uncertainty · persistence
Task 2 underway: initial evidence primitives exist; unified scoring, calibration and same-run validation remain in progress.

Task 3: Agentic AI security reasoning

INPUT FROM TASK 2 Target/time · disagreement · uncertainty · source availability · provenance

01 | PROPOSED MODULE

Integrity belief update

- Sequential probabilistic belief filter
- Track integrity and modality reliability
- Account for missing or correlated evidence

02 | PROPOSED MODULE

Threat attribution

- Evaluate normal, LiDAR, camera and V2V hypotheses
- Link temporal evidence to likely sources
- Allow uncertain / unknown threats

03 | PROPOSED MODULE

Constrained response

- Risk-gated source reweighting / exclusion
- Use redundant or conservative fallback
- Recommend actions; do not command actuators

Proof-of-concept evaluation | Matched nominal / attack / mitigation runs

DETECT	Integrity discrimination; false alarms; detection delay
DIAGNOSE	Threat/source attribution; reliability; explanation trace
MITIGATE	TTC/DRAC, braking, collisions, speed and delay

Research outputs and next steps

RESEARCH OUTPUTS

Publication	LiDAR target-loss and camera-assisted protection study — accepted for TRB 2027.
Follow-on Proposal	CCI Track A on policy-shift-aware adaptive attacks — submitted.
Future funding	Follow-on NSF cyber-physical security / trustworthy autonomy proposal planned.

ONGOING WORK & NEXT STEPS

1. Task 2 – Ongoing: Develop cross-modal, predictive, and physics-based integrity evidence models.
2. Task 2 – Validation: Calibrate and evaluate integrity evidence under nominal and attack conditions.
3. Task 3 – Next: Integrate structured evidence into agentic threat reasoning and adaptive response.

Thank you!

Qing Tang, q1tang@odu.edu
Ashley Y. Gao, Ygao18@wm.edu