Counterfactual Agentic Perturbation-Evidence Validation for V2X Safety Alerts

Ravi Babu Devareddi[1], Shiva Shankar Reddy[2]

[1] Faculty of Engineering and Built Science, Lincoln University College, Petaling Jaya, Selangor-47301, Malaysia.

[2] Department of Computer Science and Engineering, Lincoln University College, Petaling Jaya, Selangor-47301, Malaysia.

Email ID: pdf.ravibabu@lincoln.edu.my; pdfsv.shiva@lincoln.edu.my

---

Abstract: This work proposes CAPE-V2X (Counterfactual Agentic Perturbation-Evidence Validation for V2X), a semantic validation framework that verifies whether a received safety claim is physically supported by surrounding traffic behavior. Naturalistic high-D trajectory data are used to establish a behavioral prior for normal and disturbed traffic conditions, while controlled SUMO–Veins simulations generate legitimate and malicious V2X accident-alert scenarios. For each received claim, CAPE-V2X evaluates traffic behavior before and after the event, predicts the expected no-event traffic evolution, and measures the unexplained counterfactual residual. The system incorporates an adaptive threshold mechanism to improve performance under varying traffic conditions.

Keywords V2X, Agentic AI, Fake Alert Detection, Semantic Validation, Adaptive Threshold, Intelligent Transportation.

Introduction

Vehicle-to-Everything (V2X) communication plays an important role in Intelligent Transportation Systems (ITS) by enabling real-time exchange of traffic and safety information among vehicles and roadside infrastructure. However, malicious vehicles may transmit false emergency braking, accident, or road hazard alerts, affecting road safety. Existing authentication and misbehavior detection methods mainly focus on message integrity, vehicle movement, and abnormal communication patterns, but they may not reliably determine whether a reported safety event has actually occurred. To address this limitation, this paper proposes Counterfactual Agentic Perturbation-Evidence Validation for V2X Safety Alerts (CAPE-V2X). The proposed framework uses naturalistic vehicle trajectories from the highD dataset to learn normal driving behavior and estimate how traffic would evolve without the reported event. These expected movements are compared with observed vehicle behavior to identify possible event-related disturbances. A counterfactual-gated propagation mechanism examines how disturbances spread among neighboring vehicles while reducing the influence of normal traffic fluctuations. Furthermore, density-aware evidence fusion combines spatial, temporal, mobility, propagation, and sender-related information to assess the credibility of safety alerts. Specialist agents examine uncertain or conflicting evidence to improve the reliability of the validation process. Finally, the framework classifies reported events as SUPPORTED, UNCERTAIN, or UNSUPPORTED based on the available physical evidence. CAPE-V2X aims to provide a more reliable, interpretable, and context-aware approach to safety-alert validation in connected vehicular networks.

Related Work

Recent studies have explored authentication, mobility analysis, and intelligent misbehavior detection to improve V2X communication security. Jin and Papadimitratos [1] proposed a secure message verification mechanism, focusing on authentication and resistance to denial-of-service attacks. Ka et al. [2] introduced MIRAGE to detect fabricated emergency braking alerts through event-specific vehicle behavior analysis. Khan et al. [3] investigated context-aware intrusion detection using mobility and spatio-temporal information, while Ibrahim [4] employed graph neural networks to identify anomalies in cooperative vehicular messages. Hu et al. [5] explored LLM-based detection of forged vehicle information, and Kalogiannis et al. [6] developed a trajectory prediction approach for identifying abnormal mobility patterns. More recently, Abouelkheir [7] introduced counterfactual causal attribution to investigate how malicious vehicular information affects traffic evolution. Although these studies advance V2X security, further investigation is needed into event-specific counterfactual-gated propagation and uncertainty-aware semantic validation. The proposed CAPE-V2X framework addresses this problem by combining naturalistic mobility prediction, perturbation-evidence analysis, and specialist-agent verification to assess whether physical traffic observations support a reported safety event.

GAPS Identified

- Existing V2X security approaches can verify message origin or detect traffic disturbances, but they generally do not determine whether the observed disturbance was actually caused by the claimed event.
- CAPE-V2X addresses this gap through counterfactual-gated, spatio-temporal and agentic semantic validation.

Objectives

- Develop a counterfactual-based semantic validation mechanism for detecting false V2X accident alerts.
- Design counterfactual-gated spatio-temporal propagation analysis to distinguish accident-induced disturbances from normal traffic behaviour.
- Develop an agentic evidence-fusion and re-evaluation mechanism for adaptive V2X alert validation.
- Evaluate CAPE-V2X under different traffic-density conditions and compare it with existing rule-based and machine-learning approaches.

Methodology

The proposed methodology presented in figure 1, addresses a key research gap in existing V2X security systems: detecting abnormal traffic behavior does not necessarily establish whether a reported safety event is physically supported by the observed disturbance. The proposed CAPE-V2X framework therefore develops a naturalistic physical-disturbance prior using the highD dataset. This prior learns the differences between stable traffic conditions and observable traffic disturbances from algorithm 1 and provides supporting evidence for subsequent counterfactual and agentic safety-alert validation. Table 1 furnishes the data sets and tools used for this proposed methodology.

highD Trajectory Collection: Real-world vehicle trajectories from the highD dataset[8] are collected, including speed, acceleration, position, and surrounding vehicle interactions, to establish a naturalistic traffic-behavior reference.

Disturbance-Anchor Detection: Significant deceleration and sudden speed reductions are identified as candidate physical disturbances, providing proxy-positive samples for learning disturbance patterns.

Matched Stable-Window Selection: Non-disturbed traffic windows with comparable speed, density, and driving conditions are selected as reference samples to distinguish normal movement from physical disturbances.

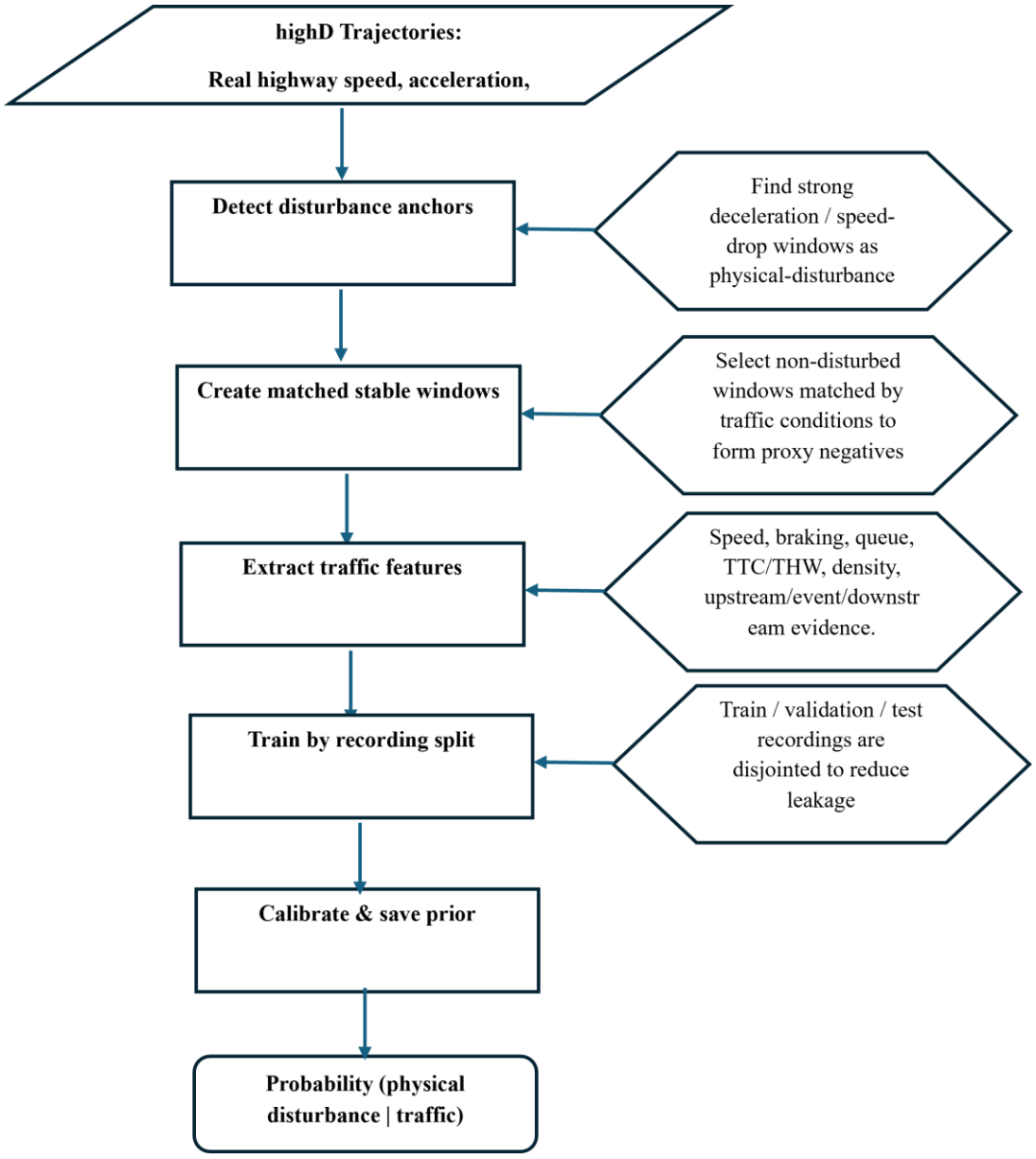


Figure 1. Semantic Score–Driven Adaptive Threshold Optimization Framework

Traffic-Feature Extraction: Traffic features, including speed, braking, queue formation, Time-to-Collision (TTC), Time Headway (THW), and density, are extracted to characterize spatial and temporal disturbance patterns.

Algorithm1: highD Behavioural Prior Learning

INPUT: highD trajectory recordings DataR

OUTPUT:Trained behavioural model highH_behavior_model

For each DataR in highD dataset:

1. Load features[vehicle_ID, frame_time, x_y_position, speed, acceleration, lane_ID, following_vehicle, TTC, THW, headway]
2. Split Road into traffic_segments and Create tiny traffic_temporal_windows
3. for each traffic_temporal_windows WT:
   Analyze the traffic_behaviour[stable,gradually_slowing,suddenly_braking, forming a queue, recovering].
4. for each vehicle extract features [speed,acceleration,speed_change,hard_braking,stopping,TTC,THW,headway] Summarizes the traffic of traffic_temporal_window[traffic_density, mean_speed, speed_variation, speed_drop, braking_ratio,stopped_vehicle_ratio, TTC, THW]
5. Construct behavioural state for each *traffic_temporal_window*:
   STABLE { *small_speed_drop, low_braking_ratio, few_stopped_vehicles, normal_TTC/THW*}
   or
   DISTURBED{ *large_speed_drop, high_braking_ratio, queue_formation, many-decelerating_vehicles, short_TTC*}.
6. Store Traffice_feature_vector $Feature_{traffice}$ and *proxy_label* Y.
7. Split $Data_R$ recording wise into {*training, validation, testing_sets*}.
8. Train *highH_behavior_model* = Learn(*F_train, Y_train*).
9. Calibrate probability using validation data.
10. Evaluate the model on unseen recordings.
11. Save the trained behavioural *highH_behavior_model*.
12. for new traffic observation F_new:
13. Proba_base = *highH_behavior_model*.predict_proba(F_new).
14. Return *Proba_base* to CAPE-V2X.

Recording-Wise Model Training: The extracted features are used to train the disturbance prediction model, with recording-level training, validation, and testing splits to minimize data leakage and improve generalization.

Probability Calibration and Prior Generation: The trained model is calibrated using validation data to produce reliable disturbance-likelihood estimates and saved as a naturalistic physical-disturbance prior.

Table 1: Data Sources used for this approach

Data Source	Origin	Nature	Usage in Proposed Work
highD[8]	RWTH Aachen University, Germany.	Real highway drone trajectories	Behavioural prior and traffic pattern modelling.
SUMO[9]	Eclipse SUMO / DLR, Germany.	Traffic simulation.	Genuine accident, no-accident and density scenarios.
Veins + [10] OMNeT++[11]	Open-source vehicular network simulation framework.	V2X network simulation.	Genuine/malicious accident-message generation.

Physical-Disturbance Probability: The calibrated prior estimates the likelihood of observable traffic disturbances, providing supporting evidence for CAPE-V2X counterfactual reasoning and semantic safety-alert validation as shown in algorithm 2. **Algorithm 2**: CAPE-V2X Semantic Accident-Claim Validation

INPUT: V2X claim event_Message; claim event time time_event; claim location loc_event surrounding vehicle observations Vehicle_surroundings; sender-context Vehicle_Sender; traffic density traffice_density

OUTPUT: Event Decision in {SUPPORTED, UNCERTAIN, UNSUPPORTED}

1. Receive claim event_Message and extract Vehicle_Sender, event type, loc_event and time_event
2. extract surrounding of vehicles for evidence[position,speed,acceleration,braking,stopped_vehicles,neighbouring_vehicles,traffic_density] near loc_event.
3. Estimate traffic density traffice_density before and after the event_Message.
4. Call the highH_behavior_model and obtain Proba_base.
5. Extract pre-event state event_Message_pre from −3s, −2s and −1s.(s-seconds)
6. Extract post-event observations event_Message_post at +2s, +4s, +6s, +8s and +10s.
7. Predict no-event counterfactual state no_event_accidentstate_counterfactual from event_Message_pre.
8. Derive residual disturbance R,
   R = Difference(event_Message_post, no_event_accidentstate_counterfactual)
9. Derive counterfactual causal evidence CC_evidance_score = CounterfactualScore(R).
   CC_evidance_score normalized to [0,1]
   CC_evidance_score ≈ 0 { disturbance probably not caused by claimed accident}
   CC_evidance_score ≈ 1 { disturbance strongly attributable to claimed accident}
10. Derive disturbance propagation score,
    disturbance_propgation_score = PropagationScore(event_Message_post, loc_event).
11. Compute gated propagation PCGP = CC_evidance_score × disturbance_propgation_score.
12. Compute background propagation Background_propagation,
    Background_propagation = (1 − CC_evidance_score) × disturbance_propgation_score.
13. Compute: i. Spatial consistency[correct place],
    ii. Temporal persistence[+2 s +4 s +6 s +8 s +10 s time intervals indentification of traffic],
    iii. Sender trust[sender_reputation,message consistency,repeated behaviour,reliability_of_available_evidence]
    iv. Density evidence[Low density: evidence, High density: congestion].
14. Fuse all evidence to obtain CAPE support probability PCAPE.
15. Select density-aware threshold τ(ρ) & δ is uncertainty margin.
    . If PCAPE > τ + δ, Decision = SUPPORTED.
    else
    If PCAPE < τ − δ, Decision = UNSUPPORTED.
    . else
    Decision = UNCERTAIN.
    If Decision = UNCERTAIN, acquire new evidence and repeat Steps 7–15.

17. Return Decision.

Result analysis

The proposed CAPE-V2X framework was evaluated using the highD dataset, SUMO traffic simulator, and Veins with OMNeT++ for V2X communication simulation. The highD dataset provides real-world highway trajectories to model natural driving behavior and traffic disturbances. SUMO was used to generate accident, non-accident, and varying traffic-density scenarios. Veins and OMNeT++ support the simulation of genuine and malicious V2X safety messages under different traffic conditions. The performance of

CAPE-V2X was compared with Rule/Plausibility[12], Random Forest[13], LSTM[14], and ST-GNN[15] approaches. The proposed framework achieved 95.83% accuracy, 92.31% precision, 100% recall, and a 96.00% F1-score and furnished as graph in figure 2. CAPE-V2X achieved the highest recall, identifying all malicious alerts in the evaluated test scenarios. In comparison, Rule/Plausibility and Random Forest achieved 91.67% recall, while LSTM and ST-GNN recorded 75.00%. However, CAPE-V2X produced an 8.33% false-positive rate, indicating a trade-off between malicious-alert detection and false alarms and it is shown as figure 2. Overall, the results demonstrate the potential of counterfactual perturbation analysis and agentic evidence validation for improving malicious V2X safety-alert detection. The comparison of proposed models with existing models of various metrics values is shown in Table 2.

Table 2: Comparison of CAPE-V2X with existing models

Method	Accuracy	Recall	F1
Rule/Plausibility	95.83%	91..67%	95.65%
Random Forest	95.83%	91.67%	95.65%
LSTM	87.50%	75.00%	85.71%
ST-GNN	87.50%	75.00%	85.71%
CAPE-V2X	95.83%	100.00%	96.00%

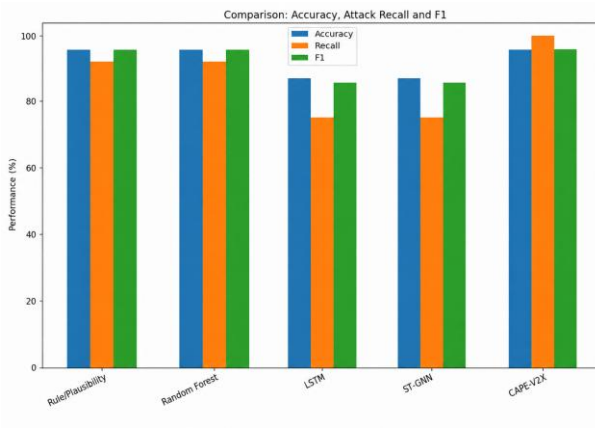


Figure 2. Graph for accuracy, attack Recall and F1 comparison.

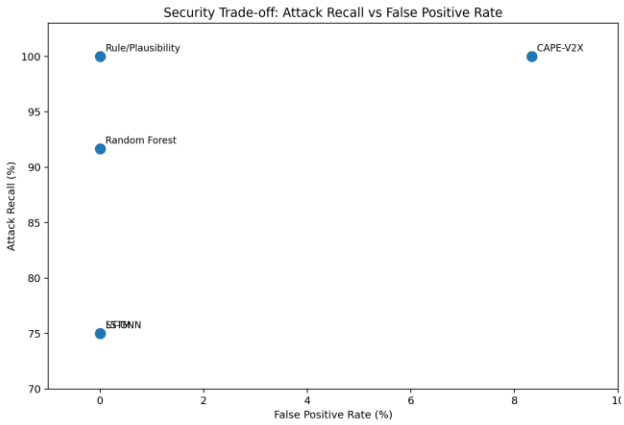


Figure 3. Graph of Attack Recall vs False Positive rate.

Conclusion

CAPE-V2X validates V2X safety alerts by examining whether the observed traffic disturbances provide physical evidence supporting the reported event. In controlled false-alert experiments, the proposed framework achieved 100% malicious-alert recall, demonstrating its ability to identify false safety messages under the tested conditions. Future work will focus on evaluating CAPE-V2X against more realistic and coordinated V2X attacks, particularly under uncertain traffic conditions, while improving counterfactual reasoning and agent-based verification for real-time applications.

References

[1] H. Jin and P. Papadimitratos, “Accountable, scalable and DoS-resilient secure vehicular communication,” *Computers & Security*, vol. 156, Art. no. 104469, 2025, doi: 10.1016/j.cose.2025.104469.

[2] E. Ka, D. Yeke, Z. B. Celik, and S. Ukkusuri, “MIRAGE: Detecting fake emergency electronic brake light attacks in V2X networks via event-gated behavioral analysis,” in *Proc. 4th USENIX Symp. Vehicle Security and Privacy (VehicleSec)*, 2026, pp. 117–134.

[3] M. D. Khan, V.-T. Ta, H. Rafiq, and N. Nnamoko, “Context-aware intrusion detection in vehicular communication networks: Enhanced attack modeling and dataset,” *Applied Artificial Intelligence*, vol. 39, no. 1, Art. no. 2538453, 2025, doi: 10.1080/08839514.2025.2538453.

[4] A. Z. Ibrahim, “Spatiotemporal graph neural network-driven anomaly detection for cooperative vehicle messaging in dense VANET corridors,” *Transactions on Emerging Telecommunications Technologies*, vol. 37, no. 4, Art. no. e70405, 2026, doi: 10.1002/ett.70405.

[5] Y. Hu *et al*., “LLM-based misbehavior detection architecture for enhanced traffic safety in connected autonomous vehicles,” *IEEE Transactions on Vehicular Technology*, vol. 74, no. 8, pp. 12829–12841, 2025, doi: 10.1109/TVT.2025.3551327.

[6] K. Kalogiannis, A. M. Hussain, and P. Papadimitratos, “PAMPOS: Causal transformer-based trajectory prediction for attack-agnostic misbehavior detection in V2X networks,” *arXiv preprint*, arXiv:2605.06833, 2026.

[7] E. Abouelkheir, “IntentProv-IoV: Causally grounded provenance for traffic-intent preservation in explainable vehicular security,” *Symmetry*, vol. 18, no. 9, Art. no. 1430, 2026, doi: 10.3390/sym18091430.

[8] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, “The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems,” in Proc. 21st IEEE Int. Conf. Intelligent Transportation Systems (ITSC), 2018, pp. 2118–2125, doi: 10.1109/ITSC.2018.8569552.

[9] P. A. Lopez et al., “Microscopic Traffic Simulation using SUMO,” in Proc. 21st IEEE Int. Conf. Intelligent Transportation Systems (ITSC), 2018, pp. 2575–2582, doi: 10.1109/ITSC.2018.8569938.

[10] C. Sommer, R. German, and F. Dressler, “Bidirectionally Coupled Network and Road Traffic Simulation for Improved IVC Analysis,” IEEE Transactions on Mobile Computing, vol. 10, no. 1, pp. 3–15, Jan. 2011, doi: 10.1109/TMC.2010.133.

[11] A. Varga and R. Hornig, “An Overview of the OMNeT++ Simulation Environment,” in Proc. 1st Int. Conf. Simulation Tools and Techniques for Communications, Networks and Systems (SIMUTools), 2008, doi: 10.4108/ICST.SIMUTOOLS2008.3027.

[12] S. So, P. Sharma, and J. Petit, “Integrating Plausibility Checks and Machine Learning for Misbehavior Detection in VANET,” in Proc. 17th IEEE Int. Conf. Machine Learning and Applications (ICMLA), 2018, pp. 564–571, doi: 10.1109/ICMLA.2018.00091.

[13]G. O. Anyanwu, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, “Novel hyper-tuned ensemble Random Forest algorithm for the detection of false basic safety messages in Internet of Vehicles,” ICT Express, vol. 9, no. 1, pp. 122–129, 2023, doi: 10.1016/j.icte.2022.06.003.

[14]N. Youness, A. Mostafa, M. A. Sobh, A. M. Bahaa, and K. Nagaty, “VeMisNet: Enhanced Feature Engineering for Deep Learning-Based Misbehavior Detection in Vehicular Ad Hoc Networks,” Journal of Sensor and Actuator Networks, vol. 14, no. 5, Art. no. 100, 2025, doi: 10.3390/jsan14050100.

[15]M. F. Yuce, M. A. Erturk, and M. A. Aydin, “Misbehavior detection with spatio-temporal graph neural networks,” Computers and Electrical Engineering, vol. 116, Art. no. 109198, 2024, doi: 10.1016/j.compeleceng.2024.109198.