Assessing the Sensitivity of Scenario Selection to Different Data Sources in the Validation Phases of Connected and Automated Vehicles

Hugues Blache[1], Pierre-Antoine Laharotte[2], and Nour-Eddin El Faouzi[2]

[1]Research Centre for Integrated Transport Innovation, School of Civil and Environmental Engineering, University of New South Wales Sydney, Australia
[2]Univ. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT-ECO7 UMR T9401, Lyon, France

Short Summary

This paper investigates a functional scenario sampling methodology for the validation of Connected and Automated Vehicle (CAV) systems within a scenario-based testing framework. The approach relies on data-driven labelling of critical scenarios and compares two input sources: simulated scenarios generated by a traffic simulator and predicted scenarios derived from machine learning models trained on field data. A four-phase evaluation framework is proposed, including scenario emulation, verification against field data, scenario reduction, and impact assessment on an adaptive cruise control (ACC) use case. The comparison highlights the sensitivity of the scenario selection process to input data, revealing divergences in extrapolated scenarios and potential biases. Results show that both approaches are consistent in observed conditions but differ beyond available data, leading to variations in selected scenarios. This work highlights the importance of input data quality for robust and reliable CAV validation.
Keywords: Connected and Automated Vehicles (CAV), scenario-based testing, functional scenarios, Scenario sampling, traffic simulation, data-driven modeling, sensitivity analysis, safety validation

1 Introduction

The validation of Connected and Automated Vehicle (CAV) systems is a critical step to ensure their safety and reliability before deployment. As highlighted by the *UNECE group Validation Methods for Automated Driving* (VMAD) (UNECE, 2024), testing strategies must balance realism, cost, and efficiency. While both simulation and field testing are commonly used, the high cost and time requirements of real-world experiments call for a rationalisation of the tested scenarios. Existing approaches typically focus on reducing actual scenarios through parameter sampling or safety indicators (Zhang et al., 2022; Hallerbach et al., 2018; Alemayehu & Sargolzaei, 2025). However, these methods may overlook rare or insufficiently represented situations due to limited available data. To address this issue, previous work (Blache et al., 2024; Blache, Laharotte, & El Faouzi, 2025; Blache, Laharotte, El Faouzi, & Saunier, 2025) introduced a methodology for sampling functional scenarios, enabling a more generic and efficient representation of traffic situations. This approach relies on three main steps: (i) ontology-based generation of functional scenarios combined with data-driven labelling of criticality using real-world observations, (ii) extension of the dataset through predictive models to compensate for missing or unobserved scenarios, and (iii) a criticality-guided sampling strategy to support scenario selection for CAV validation. However, the reliance on predicted data may introduce biases in the selected scenario subset due to the approximation of real traffic phenomena. This paper therefore investigates the sensitivity of the methodology to the nature of its input data. The objective is to quantify potential biases introduced by different input sources and to

assess the robustness of the proposed scenario sampling methodology within a practical validation framework. Two input generation approaches are evaluated in this paper:

- (i) a simulation-based approach, where a traffic simulator is used to approximate a complete artificial ground truth.
- (ii) a prediction-based approach, where machine learning models infer unobserved scenarios from real-world data

These approaches are compared at two levels: first, by evaluating discrepancies in safety indicators with respect to available field observations, and second, by analysing the consistency of the scenarios selected by the sampling process.

2 Methodology and Process for Evaluating Functional Scenario Selection

Since collecting real-world data for all ontology-generated scenarios is impractical (Blache et al., 2024), we propose a two-branch evaluation framework to assess the robustness of the scenario sampling strategy. One branch relies on simulation-based observations, while the other uses data-driven predictions (Blache, Laharotte, & El Faouzi, 2025). Both approaches may introduce biases, which we aim to quantify through a four-phase methodology (Fig. 1).

Step 1 - Scenario emulation: A traffic simulation environment is used to emulate a complete set of scenarios, considered as an artificial ground truth.

Step 2 - Verification: Simulated data are compared with real-world observations and predicted data to evaluate discrepancies between artificial and actual conditions.

Step 3 - Scenario reduction: The functional scenario selection methodology is applied using simulated data as input, and results are collected.

Step 4 - Comparison and impact analysis: The outputs obtained from simulation-based and prediction-based inputs are compared, and their impact is analysed on a specific use case.

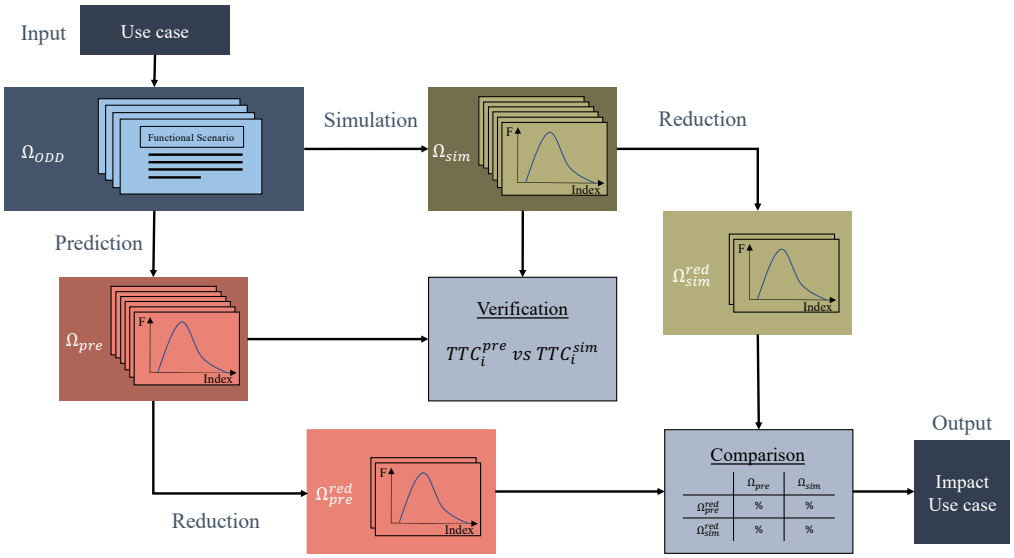


Figure 1: Proposed Methodology for Assessing the Sensitivity of the Approaches Introduced in this paper

The methodology is illustrated on an adaptive cruise control (ACC) use case, which serves as a benchmark to assess the impact of the selected scenarios.

Simulation-Based Scenario Generation (Step 1)

Functional scenarios, denoted $\mathcal{S}_f$, are modelled as stochastic representations of traffic situations at a microscopic traffic scale (Blache, Laharotte, El Faouzi, & Saunier, 2025), focusing on vehicle interactions over small road segments. Each scenario is characterised by (i) a distribution of safety indicators (e.g., time-to-collision), denoted $\mathcal{S}_f^{dist}$, and (ii) a criticality level qualification, denoted $\mathcal{S}_f^{qual}$. A traffic simulation framework is used to generate these indicators. To ensure realism, the simulation is calibrated against real-world data using a two-step approach: (i) sensitivity analysis to identify influential parameters, and (ii) optimisation (via a genetic algorithm) to minimise discrepancies between simulated and observed safety and traffic performance indicators (e.g., speed and safety metrics). Once calibrated, the simulation process generates a complete set of scenarios within the system's operational design domain (ODD), forming the simulated scenario universe:

$$\Omega_{\mathcal{S}_f}^{sim} = \{(\mathcal{S}_f, \mathcal{S}_f^{dist}, \mathcal{S}_f^{qual})\} \tag{1}$$

Comparison with Field and Predicted Data (Step 2)

The simulated scenario set is compared with two other data sources: real-world observations and predicted scenarios obtained from previous work, denoted respectively $\Omega_{\mathcal{S}_f}^{field}$ and $\Omega_{\mathcal{S}_f}^{pred}$ (Blache, Laharotte, & El Faouzi, 2025). The comparison focuses on safety indicator distributions and criticality levels. Two types of analysis are performed: (i) comparison with field data to assess realism and identify deviations, and (ii) comparison with predicted data over a larger scenario space to evaluate consistency and divergence beyond observed situations. At the same time, intersections such as

$$\Omega_{\mathcal{S}_f}^{sim} \cap \Omega_{\mathcal{S}_f}^{pred} \cap \Omega_{\mathcal{S}_f}^{field} \tag{2}$$

are used to evaluate agreement between the three sources.

Scenario Selection and Consistency Analysis (Step 3)

The scenario sampling methodology is then applied separately to simulated and predicted datasets (Blache et al., 2024). A clustering-based reduction method (Weighted K-means) is used to group scenarios and select representative and critical elements, leading to reduced sets denoted Ω_{sim}^{red} and Ω_{pred}^{red}. Within each cluster $\mathcal{C}_i$, scenarios are selected based on a combination of criticality and representativeness criteria, controlled by a threshold parameter $\tau \in [0, 1]$. The reduced set can be expressed as:

$$\Omega_{sim}^{red}(\tau) = \bigcup_{i=1}^{n} \left\{ \mathcal{S}_f \;\middle|\; Cri(\mathcal{S}_f) \geq \tau \cdot \max Cri(\mathcal{C}_i),\ D_k(\mathcal{S}_f) \geq (1-\tau) \cdot \max D_k(\mathcal{C}_i) \right\} \tag{3}$$

where $Cri(\mathcal{S}_f)$ denotes the criticality of a scenario and $D_k(\mathcal{S}_f)$ its distance to the cluster centroid, ensuring both risk relevance and diversity in the selection. The consistency between both approaches is evaluated using three metrics: (i) the difference in aggregated criticality between selected sets,

$$\Delta Cri = Cri(\Omega_{sim}^{red}) - Cri(\Omega_{pred}^{red}), \tag{4}$$

(ii) the number of common scenarios,

$$Card(\Omega_{sim}^{red} \cap \Omega_{pred}^{red}), \tag{5}$$

and (iii) the coverage of scenario attributes. These metrics allow us to assess the sensitivity of the selection process to the nature of input data and to identify potential biases.

Impact Assessment on a Use Case (Step 4)

Finally, the reduced scenario sets are used to evaluate the impact of a CAV system (ACC) on traffic safety. This step aims to analyse the practical implications of the differences observed in the scenario selection process under different input data. The analysis is based on the evolution of safety indicator distributions (e.g., TTC) before and after system deployment. A global performance metric is defined to quantify safety improvements or degradations across scenarios:

$$M_{pref} = \int \frac{\mathcal{S}_{init}^{dist}(x) - \mathcal{S}_{CAV}^{dist}(x)}{x} dx \tag{6}$$

Additionally, coverage indicators are introduced to measure the proportion of scenarios leading to safety improvements or degradations:

$$r_{safe} = \frac{N_{safe}}{N_{ODD}}, \quad r_{unsafe} = \frac{N_{unsafe}}{N_{ODD}} \tag{7}$$

This final step enables assessing not only the consistency of the scenario selection process but also its practical implications for system validation.

3 Performance Analysis

Simulation of Functional Scenarios with SUMO (Step 1)

The simulation model in SUMO was calibrated using representative traffic conditions, including both congested and free-flow situations, following the approach in (Sha et al., 2023). In this work, we focus on ACC ODD simulation within a simplified highway environment to facilitate the analysis of preliminary results under controlled conditions. A genetic algorithm with elitism was used for calibration. The population size was set to 100 individuals over 500 generations, with the top 1% retained as elite individuals at each generation. Each simulation run required approximately 8 seconds, resulting in a total computation time of 5832 minutes (97.2 hours). The calibration performance is reported in Table 1 using MAPE and RMSE. At the end, the results indicate a good agreement between simulated and observed data, with particularly accurate reconstruction of TTC-related indicators in congested traffic conditions. The calibrated model was then used to generate the simulated scenario universe, denoted Ω_{sim}.

Parameters	MAPE	RMSE
TTC_{free}	$3.2e^{-3}$	$1.62e^{-5}$
$TTC_{congested}$	$4.21e^{-4}$	$4.97e^{-7}$
$Speed_{free}$	$1.05e^{-2}$	$2.66e^{-4}$
$Speed_{congested}$	$1.84e^{-2}$	$8.47e^{-4}$

Table 1: Calibration performance of the genetic algorithm

Comparison of the Different Scenario Universes (Step 2)

This step compares simulated scenarios and predicted scenarios, both within and outside the subset covered by field observations. Table 2 reports average agreement metrics for selected ontology terms. The results show that $\Omega_{\mathcal{S}_f}^{pred}$ and $\Omega_{\mathcal{S}_f}^{sumo}$ are more consistent when restricted to scenarios also observed in $\Omega_{\mathcal{S}_f}^{field}$. Outside this intersection, stronger divergences appear. This is expected, since both methods are partially anchored in observed data for shared scenarios, whereas extrapolated scenarios are more sensitive to modelling

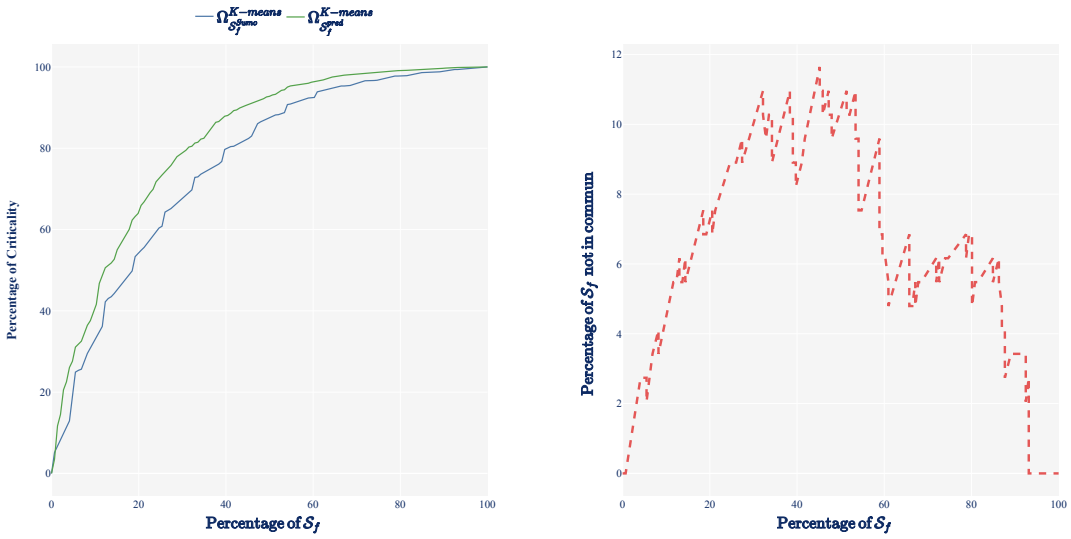
assumptions. These discrepancies may also be explained by differences in the data sources: predicted data are influenced by the diversity of field observations, while simulated data were generated in a highway-oriented environment.

Table 2: Comparative performance of simulated and predicted scenario distributions

Word	$\Omega_{sumo} \cap \Omega_{pred} \cap \Omega_{field}$			$\Omega_{sumo} \cap \Omega_{pred} \setminus \Omega_{field}$		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Free	0.72	0.004	0.002	0.543	0.006	0.003
Congested	0.765	0.003	0.002	0.523	0.010	0.005
Car	0.745	0.004	0.002	0.574	0.007	0.004
Car & Truck	0.698	0.003	0.002	0.502	0.008	0.004
Reduction of Lane	-	-	-	0.523	0.006	0.003
W\ Reduction of Lane	0.736	0.004	0.002	0.520	0.006	0.003
Straight	0.750	0.003	0.002	0.483	0.006	0.004
Lane Changing	0.703	0.004	0.002	0.577	0.005	0.003

Scenario Selection and Comparison (Step 3)

The scenario selection process was applied to both $\Omega_{\mathcal{S}_f}^{sumo}$ and $\Omega_{\mathcal{S}_f}^{pred}$ using the Weighted-K-means reduction method. Figure 2 compares the criticality coverage and the number of selected scenarios for both approaches. The results reveal a noticeable difference between the two inputs. The simulation-based approach yields a broader criticality range, which directly affects cluster formation and tends to favour the early selection of more critical scenarios. As a result, the selected subsets differ between the two approaches, with an observed divergence of about 8–12%. This indicates that the selection process is sensitive to the nature of the input data and suggests an uncertainty margin of approximately 10%.



(a) Comparison of cumulative criticality coverage for $\Omega_{\mathcal{S}_f}^{pred}$ (blue) and $\Omega_{\mathcal{S}_f}^{sim}$ (green)

(b) Comparison of the number of selected scenarios between $\Omega_{\mathcal{S}_f}^{pred}$ and $\Omega_{\mathcal{S}_f}^{sim}$

Figure 2: Comparison of the scenario selection process using two different input sources

Fig. 3 illustrates the differences in ontology attribute coverage of scenarios within the selection process. Noticeable differences can be observed between the two approaches. In particular, the simulation-based approach tends to select more scenarios associated with fluid traffic conditions, whereas the prediction-based approach favours more congested situations. This discrepancy may be explained by limitations of the prediction model, which

can overestimate the criticality of congested traffic conditions, while the simulation captures a broader range of traffic dynamics. In addition, the types of manoeuvres selected differ between the two approaches. The simulation-based approach more frequently selects lane-changing scenarios at lower speeds (e.g., R_1L_1 to R_2L_3) or scenarios remaining in faster lanes ($L3$), whereas the prediction-based approach tends to select lane changes occurring at higher speeds (e.g., R_1L_2 to R_2L_3).

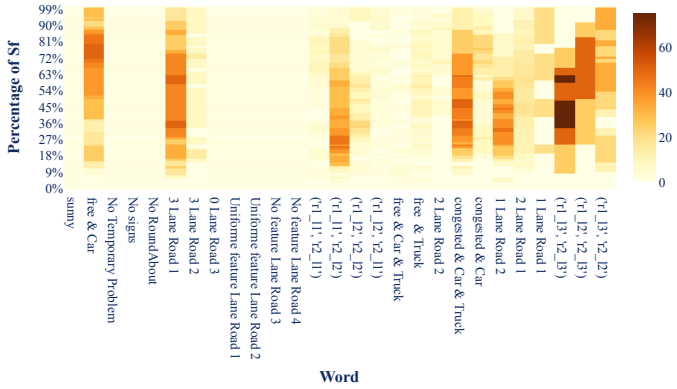


Figure 3: Differences in attribute coverage between simulation-based and prediction-based approaches

Impact Analysis on the Selected Scenarios (Step 4)

The final step evaluates the impact of implementing ACC on the selected scenarios using the calibrated SUMO environment, considering both predicted and simulated scenario selections. The ACC controller is implemented using the C-ACC parameterisation proposed in (Guériau & Dusparic, 2020), reported in Table 3. Several market penetration rates (MPR) are considered:

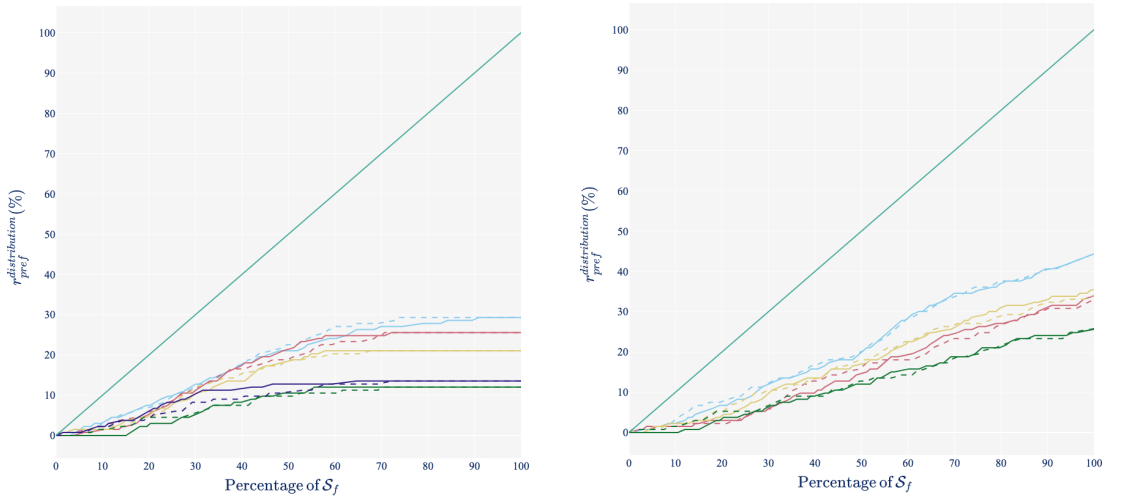
$$MPR \in [0\%, 10\%, 30\%, 50\%, 75\%, 100\%]. \tag{8}$$

Table 3: C-ACC parameters used in the simulations, based on Guériau & Dusparic (2020)

Parameters	CAV level 2 (Car)	CAV level 2 (HGV)
Car-following model	C-ACC	
Speed deviation (%)	0.05	0.05
Time headway (s)	0.8	0.8
Min gap (m)	1.5	1.5
Max accel. (m/s^2)	1.5	1.5
Deceleration (m/s^2)	7.5	4
Max decel. (m/s^2)	9	7
Imperfection	0.05	0.05

Figures 4a and 4b show the proportion of scenarios with improved safety for CAVs and surrounding human-driven vehicles (dashed lines for predicted scenarios and solid lines for simulated scenarios). The results suggest that ACC provides greater benefits at low penetration rates with few variation with both approach. They also show that the first selected scenarios tend to produce the lowest gains, indicating that the most critical scenarios remain the most challenging for the system. As more scenarios are considered, the level of improvement generally increases. This suggests that if the system performs well under the

most demanding selected scenarios, it is likely to provide broader safety benefits over the remaining scenario space.



(a) Benefit on distribution ($r_{pref}^{distribution}$) for CAVs

(b) Benefit on qualification ($r_{pref}^{qualification}$) for human-driven vehicles

(c) Legend

Figure 4: Safety benefits of ACC for predicted scenarios (dashed lines) and simulated scenarios (solid lines)

4 Discussion and Conclusion

This paper applies and evaluates a methodology for functional scenario sampling using a concrete use case. Due to the limited availability of comprehensive field data, we proposed a dual-input evaluation framework based on: (i) predicted scenarios derived from (Blache et al., 2024) ($\Omega_{\mathcal{S}_f}^{pred}$), and (ii) simulated scenarios generated with SUMO ($\Omega_{\mathcal{S}_f}^{sumo}$). The comparison of these two approaches highlights the sensitivity of the scenario selection process to the nature of input data. While both approaches show consistency when aligned with observed data, divergences emerge in extrapolated scenarios, leading to differences in selected scenario subsets. This confirms the presence of biases in the sampling process and suggests an uncertainty margin when relying on estimated data. Beyond the selection stage, the proposed methodology enables the evaluation of a system under a reduced set of functional scenarios. Using an adaptive cruise control (ACC) use case, we introduced simple safety metrics (e.g, TTC) to assess the impact of the system on both ego and surrounding vehicles. The results show that system performance depends on the selected scenarios and on the market penetration rate, with greater benefits generally observed at lower penetration levels. Despite these contributions, several limitations remain. First, the simulation calibration relies on a limited set of scenarios and indicators, which may affect realism. Extending the calibration dataset and improving computational efficiency would enhance accuracy. Second, only one ACC configuration was considered, exploring different parameter settings would allow a better understanding of the trade-off between traffic performance and safety. Finally, the analysis was restricted to a single selection method (Weighted K-means), and future work should investigate alternative sampling techniques to assess their robustness and performance. This work provides insights into

the robustness and limitations of functional scenario sampling methodologies and highlights the importance of input data quality in CAV validation processes.

References

Alemayehu, H., & Sargolzaei, A. (2025). Testing and verification of connected and autonomous vehicles: A review. *Electronics*, *14*(3), 600.

Blache, H., Laharotte, P.-A., & El Faouzi, N.-E. (2024). Toward a criticality-guided sampling strategy to reduce tests for automated driving validation. *Data Science for Transportation*, *6*(3), 26.

Blache, H., Laharotte, P.-A., & El Faouzi, N.-E. (2025). Automatic labeling and qualification of functional scenarios on the basis of sparse field observations. *Journal of Intelligent and Connected Vehicles*, *8*(3), 9210058-1-9210058-16.

Blache, H., Laharotte, P.-A., El Faouzi, N.-E., & Saunier, N. (2025). Unifying expert knowledge and field data toward an enhanced scenario description for cav certification: a comprehensive scenario-based approach. *Journal of Intelligent Transportation Systems*, 1–27.

Guériau, M., & Dusparic, I. (2020). Quantifying the impact of connected and autonomous vehicles on traffic efficiency and safety in mixed traffic. In *2020 ieee 23rd international conference on intelligent transportation systems (itsc)* (p. 1-8). doi: 10.1109/ITSC45102.2020.9294174

Hallerbach, S., Xia, Y., Eberle, U., & Koester, F. (2018). Simulation-based identification of critical scenarios for cooperative and automated vehicles. *SAE International Journal of Connected and Automated Vehicles*, *1*(2018-01-1066), 93–106.

Sha, D., Gao, J., Yang, D., Zuo, F., & Ozbay, K. (2023). Calibrating stochastic traffic simulation models for safety and operational measures based on vehicle conflict distributions obtained from aerial and traffic camera videos. *Accident Analysis & Prevention*, *179*, 106878.

UNECE. (2024). *Validation method for automated driving (vmad).* Retrieved from `https://wiki.unece.org/pages/viewpage.action?pageId=60361611` (Accessed on March 29, 2026)

Zhang, X., Tao, J., Tan, K., Törngren, M., Sánchez, J. M. G., Ramli, M. R., ... others (2022). Finding critical scenarios for automated driving systems: A systematic mapping study. *IEEE Transactions on Software Engineering*, *49*(3), 991–1026.