Synthetic vs. Real: An Analysis of Critical Scenarios for Autonomous Vehicle Testing

Qunying Song[1*], Avner Bensoussan[2] and Mohammad Reza Mousavi[2]

[1*]Department of Computer Science, Lund University, Box 118, Lund, 22100, Sweden.
[2]Department of Informatics, King's College London, Bush House, Aldwych, London, WC2R 2LS, United Kingdom.

*Corresponding author(s). E-mail(s): qunying.song@cs.lth.se;
Contributing authors: avner.bensoussan@kcl.ac.uk;
mohammad.mousavi@kcl.ac.uk;

Abstract

With the emergence of autonomous vehicles comes the requirement of adequate and rigorous testing, particularly in critical scenarios that are both challenging and potentially hazardous. Generating synthetic critical scenarios in simulation for testing autonomous vehicles has therefore received considerable interest; yet, it is unclear how such scenarios relate to the actual crash or near-crash scenarios involving autonomous vehicles. Consequently, their realism is unknown. In this paper, we define realism as the degree of similarity of synthetic critical scenarios to real-world critical scenarios. We propose a methodology to measure realism using two metrics, namely attribute distribution and Euclidean distance. The methodology extracts various attributes from synthetic and realistic critical scenario datasets and performs a set of statistical tests to compare their distributions and distances. As a proof of concept for our methodology, we compare synthetic collision scenarios from DeepScenario against realistic autonomous vehicle collisions collected by the Department of Motor Vehicles in California, to analyse how well DeepScenario synthetic collision scenarios are aligned with real autonomous vehicle collisions recorded scenarios in California. We focus on five key attributes that are extractable from both datasets, and analyse the attribution distribution and distance between scenarios in the two datasets. Further, we derive recommendations to improve the realism of synthetic scenarios based on our analysis. Our study of realism provides a framework that can be replicated and extended for other dataset both concerning real-world and synthetically-generated scenarios.

Keywords: realism, synthetic scenarios, collision scenarios, critical scenario identification, testing, autonomous vehicles, autonomous driving systems

1 Introduction

Autonomous vehicles (AVs) are expected to improve road safety, traffic efficiency, and mobility [1]. Before they can be deployed on public roads on a large scale, they need to be tested adequately and rigorously [2–4]. As the SOTIF (Safety of the Intended Functionality [5]) standard articulates, we need to test all relevant scenarios for AVs, especially in those challenging conditions for the sensors and systems [6], which are often known as *critical scenarios* [7–9]. In addition, the latest EU regulation for type approval of AVs also requires manufacturers to test not only critical scenarios observed from natural driving data, but also reasonably foreseeable ones [8, 10].

Although collecting critical scenarios from real-world traffic is valuable, simulation is commonly employed for its accessibility and efficiency. Therefore, synthetic critical scenario identification has received considerable attention, serving as a complement to real-world data collection [7, 8]. The identification process typically involves simulating various driving scenarios and optimising the generation of critical scenarios with respect to different performance metrics or criteria [2, 7, 8]. The resulting critical scenarios are those that expose risks of harm to the AV within its operational design domain (ODD), such as collisions with other road users or infrastructure. These risks may stem from errors in the AV or challenging situations beyond the AV's capability to manage. Despite the extensive body of studies reported, the overwhelming majority of them primarily focus on the approaches and tool chains for identifying critical scenarios, *leaving the evaluation of resulting scenarios unexplored*. Consequently, *the realism of such scenarios and their relevance to testing AVs/ADS is unclear.*

To remedy the gaps, our goals are to 1) *devise a quantified methodology for measuring the realism of synthetic critical scenarios*, 2) *apply this methodology to an existing dateset, as a proof of concept, in order to show its applicability*, and 3) *derive guidelines to improve the realism of the scenarios considered in our proof of concept for testing AVs/ADS*. To enable evaluation from both macroscopic and microscopic perspectives, we define realism as *the degree of similarity of a set of synthetic critical scenarios to realistic critical scenarios*; to quantify realism, we use two metrics: 1) *attribute distribution – distribution of scenario attributes in the two datasets* and 2) *Euclidean distance – the straight-line distance between scenarios in a vectorised space*. In line with the goals, we formulate two research questions for this study:

RQ1: How can we quantify the realism of a synthetic (simulation-based) dataset for critical scenarios?
- **RQ1.1**: How can we quantify realism using a comparison of the distribution of attributes and Euclidean distance in a vectorised space?
- **RQ1.2**: What are the attributes that are causal for realistic AV collisions?
- **RQ1.3**: Are causal parameters identified in RQ1.2 included proportionately in synthetically generated scenarios?

RQ2: What can be improved to generate more realistic synthetic critical scenarios?

- **RQ2.1**: What are the guidelines for closing the reality gaps in synthetic scenarios?
- **RQ2.2**: What are the guidelines for field testing with synthetically generated scenarios?

We select two AV collision scenario sets for our proof of concept, including a realistic set from DMV (Department of Motor Vehicles) California [11], and a synthetic set from DeepScenario [12, 13]. DMV California provides collision reports from manufacturers during test drives in California [11]. DeepScenario generates synthetic collision scenarios on a San Francisco map using Apollo ADS and SVL simulator [12, 13]. The two datasets are selected primarily based on the facts that they are 1) *the most extensive ones to the best of our knowledge*, 2) *documented in a standard and structured specification*, 3) *publicly available*, and 4) *both based in the same state*. The purpose is to analyse how well the synthetic collision scenarios generated by DeepScenario are aligned with real autonomous vehicle collisions recorded in California, and derive general recommendations to improve the realism of synthetic critical scenarios.

We extract five relevant attributes that are available for both datasets, including *weather*, *lighting*, *roadway surface*, *roadway conditions*, and *collision type*. Then, we compare the attribute distribution and Euclidean distance between DeepScenario and DMV California data to reveal their similarities. Lastly, we interview the author of DeepScenario to assess our evaluation results and analysis. Although the attribute distribution differs significantly, we observe DeepScenario generates similar collision scenarios as in DMV California from a distance perspective. To improve the realism and future evaluation of synthetic scenarios, we recommend 1) *incorporating more realistic attributes and values in synthetic critical scenario identification*, 2) *validating and improving the quality of simulators to ensure a faithful representation*, and 3) *developing comprehensive guidelines for scenario collection and specification*.

Our proof of concept evaluates the applicability of our method and serves the first step towards measuring the realism of synthetic critical scenarios. Currently, the proof of concept is subject to several limitations: 1) *the selected datasets are still too small* (however, they are steadily growing in size); 2) *the attributes extracted are limited*; and 3) *contextual information such as test arrangement are unavailable*. Thus, we come up with guidelines on how to extend the available datasets and improve the data gathering methods to enable more precise analysis. We would also like to include other datasets and additional attributes to perform more sophisticated evaluations in future work. Considering realism is an essential quality for AVs/ADS test scenarios [8], and very limited studies have reported any relevant definitions, metrics, approaches, and insights in this topic, our proof of concept provides insights of and recommendations for the realism of synthetic critical scenarios and a basis for future studies.

In summary, our study makes four major contributions:

1. We propose two metrics to evaluate the realism of synthetic critical scenarios, namely *attribute distribution* and *Euclidean distance*. The metrics provide both macroscopic and microscopic views of the realism of a synthetic critical scenario set.

2. As a proof of concept, we apply our metrics to measure how synthetic collision scenarios by DeepScenario are aligned with realistic AV collisions in *DMV California*, revealing insights from practical perspectives.
3. We observe existing shortcomings and possible future improvements, serving as *guidelines* for recording realistic scenarios, and generating and evaluating synthetic critical scenarios. The guidelines are general and not specific to the datasets in this study.
4. We include *human assessment* in the loop to provide further insights and guidelines for evaluating the realism of synthetic critical scenarios on top of the empirical evaluation.

The rest of the article is organised as follows: in Section 2, we review the relevant literature to this study. We formulate the research approach in Section 3, and present the results and analysis in Section 4. In Section 5, we discuss our findings and the validity of the study. Lastly, we conclude the article in Section 6.

2 Related Work

In general, there are few pieces of research providing a rigorous and quantified methodology for establishing the realism of a critical scenario dataset. In the remainder of this section, we review the only exceptions we are aware of. Subsequently, we review the available datasets that could be the subject of such a realism study.

2.1 Scenario Realism Evaluation

To the best of our knowledge, very limited studies have been dedicated to evaluating the realism of synthetically generated scenarios for testing AVs/ADS, although realism has been considered an essential quality for test scenarios [8]. Consequently, no standard definition or metrics for evaluating realism are established. Below we present several studies that are reported on validating the simulation model and test adequacy metrics for AVs, and contrast to our study.

Stocco et al. [14] compared the performance of AV in simulated and real-world environments and revealed gaps and transferability of testing in those two different environments. Riedmaier et al. [15, 16] conducted a literature survey on verification, validation, and uncertainty quantification methods for simulation models across various application domains, including autonomous driving, and have developed a unified framework to assess the errors and uncertainties of these models. In a similar study, Reisgys et al. [17] present a method to compute the simulation errors by comparing the system response quantities in simulation and proving ground tests. The authors consider a simulation model valid if the simulation error is still within the tolerance boundary based on the proving ground tests. Neurohr et al. [18] also designed a method to compare the similarity between natural driving data recorded in the real world and its simulation counterparts to validate the simulation model. Besides, Sargent [19] discussed, more broadly, different approaches (paradigms and techniques) to assess model validity and recommended a procedure for it. While those studies commonly involve real-world driving data and their simulation counterparts to analyse the errors in the simulation model, we specifically focus on (system-level) critical driving scenarios and

evaluate the similarities between synthetic and realistic AV critical scenarios based on selected features, such as weather, lighting, and roadway conditions.

Braun et al. [20] reviewed and presented several proximity measures for scenario similarity. For example, comparing the similarity of time series data such as trajectory of two scenarios, or the sequence of maneuvers involved. Other than that, distance is used to quantify the similarity of two scenarios based on selected features, which is an important measure we use in this study, as presented in Section 3.1.2. In general, a low distance means a high similarity and vice versa. Two identical scenarios should have a distance of zero [20]. Neelofar et al. [21] presented several adequacy metrics to analyse the coverage and diversity of test scenarios for AI-powered systems. Among those metrics, Euclidean distance is used to measure the distance between two scenarios and the diversity of test scenarios for testing AVs. In another study, Yan et al. [22] designed NeuralNDE, a deep learning-based framework that produces real-world driving environments in simulation with statistical realism, which means the road events and driving behaviours follow a real-world distribution. The inconsistency in statistical difference in, e.g., relative distance and speed between vehicles, or how and when vehicles yield to the others in roundabouts, between simulation and the real world, will result in simulation gaps and unreliability for testing. Therefore, real-world distribution of road events and conditions are significant to incorporate, which is another important metric that we use in this study, as described in Section 3.1.1.

Several studies discussed realism, yet they were mostly articulating the need to evaluating the realism of test scenarios. To exemplify, Sun et al. in their study for scenario-based testing of AVs, eliminated scenarios with unavoidable collisions in the initial state as they consider such scenarios meaningless for testing [23]. Although unavoidable collisions can still be relevant for testing AVs/ADS, the study does provide a potential dimension to consider for assessing the realism of scenarios. In order to explore and identify realistic test scenarios for AVs, Abbas et al. proposed to analyse the dynamic feasibility of given maneuvers and the composition of different maneuvers [24], Tenbrock et al. incorporated the probability of a scenario [25], and Neurohr et al. articulated the need for validating simulation environments [1].

2.2 Realistic Critical Scenarios

There are a number of public reports on various types of incidents involving autonomous vehicles, both at the national (federal) level [26, 27] and the state level [11, 28]. To date, the disengagement and collision reports produced by the California Department of Motor Vehicles (CA-DMV) are the most comprehensive available reports that were subject to extensive research. Several studies have used CA-DMV collision reports [29–31] to analyse the distribution of attributes such as vehicle maneuvers and collision types. In contrast, we compare such distribution with a synthetic collision set to evaluate its realism. The CA-DMV reports have also been compiled into public datasets after augmentation with public data, such as open street maps [32]. The CA-DMV Collision Reports [11] served as one of the two main sources of data for our research. We use the raw reports rather than the compiled dataset to carefully scrutinise the raw data and prepare it in a suitable format that matches the information provided by the synthetic scenarios.

2.3 Synthetic Critical Scenarios

Although real-world critical scenarios are significant sources for realistic testing of ADS, collecting such scenarios requires deploying AVs or sensors on real-world traffic, which is risky and expensive [8]. Therefore, simulation environment is commonly used for its accessibility and efficiency. In addition, regular road traffic is non-critical most of the time [33], collecting a fair amount of critical scenarios, such as AV collisions, in real world is time-consuming. Thus, such scenarios are not available on a large scale yet. In comparison, synthetic critical scenario identification is more efficient by constructing various weather, road conditions, and interactions between AV and other road users in simulation. In general, we should find structured methods for these two techniques to be combined and complemented in an iterative manner [8, 34].

Identifying critical scenarios and automating scenario generation are among the two most prominent challenges in the domain of testing AVs [2]. Often, road accident reports have been a source for designing critical scenarios, as evidenced by recent interviews with domain experts [2]. Also, there has been sizeable research using search and optimisation algorithms to explore synthetic critical scenarios in simulation. Below, we review some relevant studies to give an overview of them.

Several studies have used accident reports to reproduce synthetic scenarios for testing AVs/ADS [35–38]. Gambi et al. [35, 37] reproduced synthetic scenarios from police crash reports using natural language processing. In another study, Gambi et al. [36] used accident sketches to extract road information, collision type, and vehicle dynamics to reconstruct the scenario for simulation. Zhang et al. [38] developed a toolkit to extract scenarios from accident videos and store them in a scenario library for testing AVs. Although using real traffic accident data can improve realism, such data do not usually involve AVs and need to be extracted with appropriate techniques.

Several studies have used search-based approaches [3] to optimise the generation of critical scenarios in simulation [39–47]. Abdessalem et al. [45] and Calo et al. [44] used multi-objective search algorithms, such as NSGA-II, to generate critical scenarios for testing ADS. In a similar study, Abdessalem et al. [39] developed a novel algorithm, FITEST, extending MOSA, and evaluated it on generating unsafe scenarios for industrial ADS. Li et al. [42] designed a framework for finding safety violations using a genetic algorithm. Luo et al. developed an approach for generating critical scenarios with requirement violations using NSGA-III. The requirement violations cover both safety and comfort perspectives [43]. While most studies focused on approaches and tool chains for identifying synthetic critical scenarios, DeepScenario [12, 13] used different optimisation strategies to identify critical scenarios for Apollo ADS in the SVL simulator. DeepScenario opened the resulting dataset (i.e., 1 050 collision scenarios and a total of 33 530 scenarios) in structured specifications.

3 Research Approach

Our approach to measuring realism comprises four steps (S1–S4), as illustrated in Figure 1. S1 selects appropriate metrics to evaluate the realism of critical scenarios. S2 prepares the datasets for subsequent evaluation. S3 evaluates the datasets from S2 using three analyses based on the metrics from S1, two use attribute distribution and

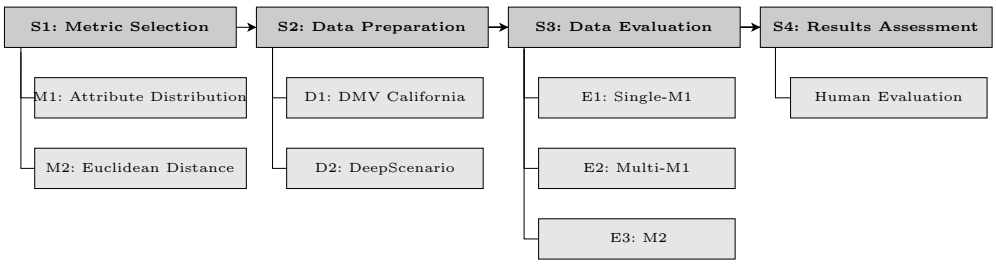

Fig. 1: An overview of our research approach with four steps (S1–S4) performed in the given order. S1 selects two evaluation metrics (M1 and M2) for realism. S2 prepares two datasets (D1 and D2) for evaluation. S3 evaluates the datasets from S2 using three analysis (E1, E2, and E3) based on the metrics from S1. Lastly, in S4, we interview the stakeholders of DeepScenario (D2) to assess our evaluation results from S3.

one uses Euclidean distance. Lastly, in S4, we close the loop by involving the stakeholders involved in dataset collection (including scenario generation and selection) to assess our evaluation results and provide feedback. Below we explain each and every step, using our proof of concept example to illustrate them.

3.1 Evaluation Metric Selection

As we discuss earlier in Section 2, there has been no specific metrics defined to evaluate the realism of critical scenarios. Thus, we need to select appropriate metrics that can give us some insights or perspectives to understand the realism of a synthetic critical scenario set. As a first proposal, we select two evaluation metrics for realism, namely *attribute distribution* and *Euclidean distance*. The metrics are not necessarily the most comprehensive ones, but they provide different lenses to measure realism.

3.1.1 Attribute Distribution

To evaluate the realism from a macroscopic view, *attribute distribution* represents the distribution of an attribute or combination of attributes of a critical scenario set. Further, it enables causal analysis of critical attributes for collisions by comparing the prior and posterior distributions, when available and eliminating confounders [48]. When a causal relation is established (e.g., for crashes or near-crash scenarios), then the comparison of distribution of causal attributes will precisely measure how effective synthetic scenarios cover what is causal for a critical situation. Any discrepancies give a potential indication of lack of sufficient coverage in such cases. Although statistical analysis on attribute distribution often requires a fair and sufficiently large dataset, we apply this to our two proof of concept datasets and explore what we can learn from such analysis, and what needs to be improved if we fail to perform such analysis.

3.1.2 Euclidean Distance

To evaluate the realism from a microscopic view, *Euclidean distance* [49] refers to the distance between two critical scenarios measured in a vectorised space. In general,

the larger the distance between two scenarios is, the bigger differences they possess in their attributes [20]. The distance measures how close, or similar, a synthetic critical scenario is to the real-world critical scenarios, further indicating the realism of it. Further, the distance needs to be associated with a specific evaluation criteria, be it a reasonably set threshold for the distance from a synthetic critical scenario to its closest real-world critical scenario, or comparing the same distance to the mean or maximum distance between all real-world critical scenarios, depending on the actual analysis.

Different from the *attribute distribution*, which measures the distribution of an attribute or combination of attributes in a scenario set, *Euclidean distance* measures the distance between two scenarios. In other words, *attribute distribution* reveals the similarity between two scenario sets in terms of scenario distribution for a specific attribute, e.g., weather, while *Euclidean distance* indicates the similarity of two concrete scenarios based on their distance in a vectorizied space. Given that, two similar or identical scenarios (with zero or low *Euclidean distance*) may still have very different distributions in two datasets, leading to discrepancies in *attribute distribution*.

3.2 Data Preparation

This step involves processing the datasets to identify the scenarios of interest, e.g., by defining what level of automation is needed and what kind of incidents need to be represented. The level of automation is selected by the manufacturers when reporting a collision, which could be either autonomous driving or conventional driving (non-autonomous driving). We filter out scenarios in conventional driving mode where AV was deactivated from the autonomous driving mode. In other words, the vehicle was manually driven by a human driver and the collision was not related to the ADS. Subsequently, we map the scenarios into the attributes of interest, i.e., by defining the attributes and coding the dataset or using learning techniques to decide on the features and the corresponding labels of the scenarios. We select five attributes in this study, which are all categorical, describing environmental characteristics of the scenarios. We then map the scenarios into a vector space to calculate the Euclidean distances among them. Below we illustrate this step and its ingredients on our datasets.

For our proof of concept, we use the AV collision set from DMV California, and a synthetic AV collision set from DeepScenario. Ideally, for a fair comparison, we would like the compared datasets to be collected under similar conditions, i.e., using the same ADS, and considering the same weather and road conditions. Due to the unavailability of two perfectly matching datasets, we selected two public datasets collected in the same state for our proof of concept. For each set, we filter the data, extract relevant attributes, and vectorise them to prepare for the subsequent evaluation.

3.2.1 DMV California

AV manufacturers in California are required to submit a collision report to DMV California within 10 days of an incident [11]. Until the start of the current study (i.e., June 2, 2023), there have been 603 AV collision reports submitted to DMV California. We download this dataset and Figure 2 provides a visualised representation of the preparation on this dataset.

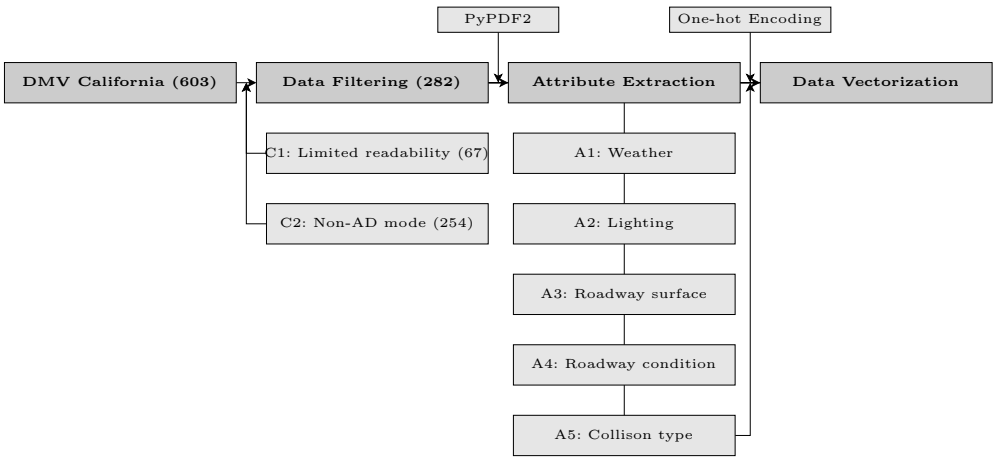

Fig. 2: An overview of preparation of DMV California data. Initially, we access 603 collision reports from DMV California. We filter the data with two criteria (67 for C1 and 254 for C2) and obtain 282 collision reports. Non-AD mode in C2 refers to the non-autonomous driving mode in a collision. After that, five relevant attributes (A1–A5) are extracted from the remaining data using the PyPDF2 Python library, and are vectorised using the One-hot encoding approach.

1. *Data Filtering.* We exclude 67 collision reports before April 2018 due to their limited readability and information. Although AV collision reports existed since 2014, they are scanned photocopies where quality is low, and readability of the text in the reports is limited. For reports after 2016 and before April 2018, the scanning quality is improved, but the collision still heavily relies on a qualitative description, where quality of the description and information contained is entirely dependent on the manufacturers.

   Further, we exclude 254 collision reports after April 2018 due to the use of *conventional* (non-autonomous driving) mode at the time of the collision. Since we specifically focus on AV collisions in this study, we exclude reports with the conventional mode selected where AV was deactivated from the autonomous driving mode. As a result, we retain 282 collision reports for subsequent evaluation.

2. *Attribute Extraction.* We use *PyPDF2* library [50] to extract relevant attributes for collision reports since 2019 as they are standard PDF files. For reports before 2019, we manually extract the attributes as they are scanned photocopies and auto-extraction is significantly more complicated. As provided in the supplementary material [51], a report contains three main sections, i.e., *manufacturer information*, *accident information*, and *accident details*. Specifically, *accident details* contain a qualitative summary of the collision and a structured table. We focus on the table and extract five attributes that are available and can be extracted from both datasets, i.e., *weather*, *lighting*, *roadway surface*, *roadway conditions*, and *collision type*.

After that, we manually compensate 30 collision reports with a missing field. Among them, 13 collision reports have no value for the *roadway surface* attribute. We carefully inspect each collision report, and use *Dry* for clear and *Wet* for raining weather. There are another 17 collision reports with empty *roadway conditions*, we use *No unusual conditions* for them after inspecting the accident details description in the reports, where no specific roadway condition or related information is identified. Instead of eliminating those collisions for missing one attribute, we want to keep more data in our analysis to get a better view of the collisions and subsequent evaluation and analysis of realism with DeepScenario.

Additionally, we manually update six collision reports with multi-value fields and 46 with two collision types. Two reports have both *Cloudy* and *Raining* for weather and are revised to *Raining*. Four reports have two roadway conditions selected. An example is *cruise_032123.pdf* with *Obstruction on roadway* and *Other*. In that case, *Other* is removed as it does not provide useful information. 46 reports have two collision types, one for each vehicle involved. We carefully inspect the accident details description and select the type that fits most appropriately. Among them, 32 reports are clearly *Rear end* where one vehicle's front hit the other vehicle's rear end, but have *Head on* for one and *Rear end* for the other vehicle. Similarly, nine reports have *Broadside* where one vehicle's front hit the other's side, but was selected as *Head on* and *Broadside*. We use the same principle to revise the remaining five reports where the manufacturer selected a different collision type for each vehicle involved. Although having multiple values for one field might be justified in certain situations, this is rarely used (18.44%) and it adds substantial complexity to the analysis. For example, it is not straightforward to determine the similarity between multi-valued and single-valued fields.

3. *Data Vectorisation.* Finally, we use the one-hot encoding approach [52] to vectorise the selected data for subsequent comparison and analysis. Specifically, we create a binary vector with the length of the number of options for each attribute in the collision reports. Then, we mapped 1 into the corresponding bit of the vector for each selected option in the report, and 0 for the remaining bits. As a result, each data (scenario) contains five binary vectors for five attributes, and in total 32 bits in length.

3.2.2 DeepScenario

DeepScenario is an open driving scenario dataset for testing ADS and contains 33 530 synthetic driving scenarios [12, 13]. DeepScenario used the Apollo ADS to navigate the autonomous vehicle in the SVL simulator [12, 53]. Further, it used three strategies (i.e., random, greedy search, and reinforcement learning) to optimise critical scenarios with respect to different reward functions (i.e., time to collision, distance to obstacles, and jerk) on various roads, weather, and behaviours on the map of San Francisco. In other words, DeepScenario explores different roads, weather, and vehicle behaviours in simulation, and searches for driving scenarios where the AV collide with other vehicles, pedestrians, or road infrastructures. Among the scenarios, there are 1,050 collision scenarios where the AV collides with other vehicles, pedestrians, or objects.

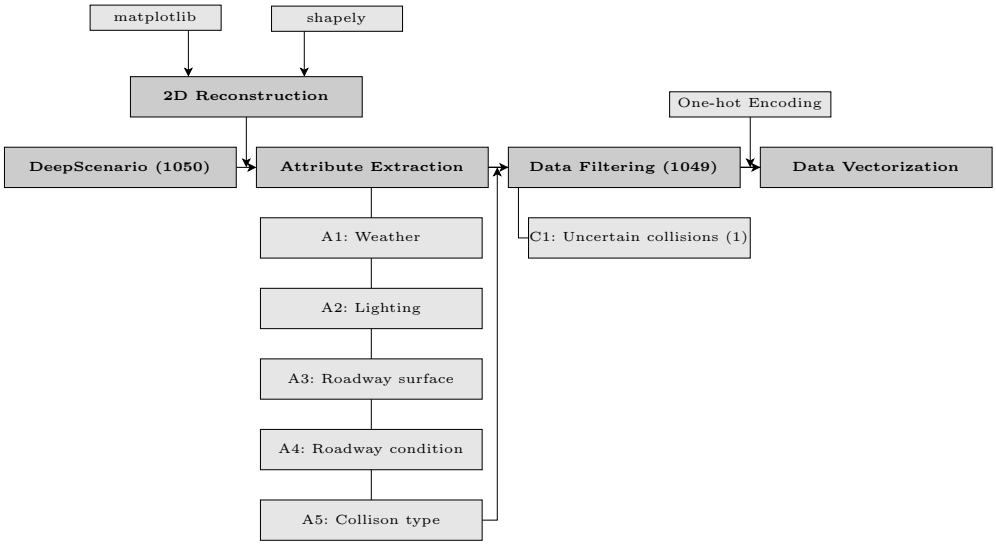

Fig. 3: An overview of preparation of DeepScenario data. Initially, we access 1050 collision scenarios from DeepScenario. We extract five relevant attributes (A1–A5) as for DMV California using scenario specification and a 2D reconstruction (developed using *matplotlib* and *shapely* Python libraries). One scenario is filtered due to uncertainties (C1) identified in the 2D reconstruction, and we obtain 1049 collision scenarios. Finally, selected scenarios and their extracted attributes are vectorised using the One-hot encoding approach.

As presented in Figure 3, we employ the same preparation as for DMV California data, with an additional step of 2D reconstruction to automatically extract the collision type for DeepScenario.

1. *Attribute Extraction.* We extract *weather* and *lighting* conditions from the directory name of the scenario. DeepScenario organises scenario description files in a multi-level directory as the summary box below. The directory contains the *generation strategy*, *reward function*, *road*, *weather*, *time of the day*, and *name* of the scenario description file. DeepScenario used two weather conditions (i.e., sunny and rainy), which correspond to *Clear* and *Raining* in the DMV California dataset. *Lighting* is derived from the time of day (i.e., 8:00 for day and 20:00 for night) and is mapped to *Daylight* and *Dark – Street lights* respectively.

   ../rl_based-strategy/reward-dto/road1-sunny_day-scenarios/0_scenario_0 .deepscenario

   We then derive *roadway surface* and *condition* based on the weather and generation configurations. As DeepScenario set the *roadway surface* according to the weather, we use *Dry* for sunny and *Wet* for rainy weather to align with the DMV

California data. Besides, since DeepScenario did not employ any unusual roadway conditions such as construction or holes on the roadway, we use *No unusual conditions* for *roadway condition* in DeepScenario.

2. *2D Reconstruction.* We extract *collision type* from the 2D reconstruction of the scenarios. Since LG has stopped its server cloud and development for SVL simulator in 2022 [54], extraction of collision type for DeepScenario scenarios relies on the scenario description files. Although open-source projects such as SORA-SVL [55] are developed as a local cloud built for the SVL simulator, they do not support the SVL version used by DeepScenario (i.e., 2021.1).

   A scenario description file in DeepScenario is an XML-based specification including the *environment* (e.g., city, date and time, and weather), *entities* (e.g., vehicles, pedestrians), and a *storyboard* (dynamic parameters of the entities) for a period of three seconds. The *storyboard* contains six timestamps collected per 0.5 seconds and each timestamp contains the *position*, *velocity*, *angular velocity*, *GPS*, and *orientation* of each entity. We use *matplotlib* and *shapely* libraries to reconstruct a 2D plot of each timestamp and visualise the *bounding box*, *orientation*, and *speed* of each entity, see Figure 4, to extract *collision type*.

   We automatically extract *collision type* for 816 scenarios using 2D reconstruction and Python scripts we develop. Specifically, DeepScenario already labeled collision scenarios with a collision type of *Pedestrian*, *Obstacle*, and *npc_vehicle*. While *Pedestrian* and *Obstacle* can be easily mapped to *Vehicle/pedestrian* and *Hit object* from DMV California, *npc_vehicle* refers to colliding with other vehicles and needs to be further categorised to align with DMV California. We use bounding boxes of the entities to identify the intersection between the AV and the colliding vehicle, and determine the collision type based on which side of the vehicles involved most in the intersection area. We adopt the taxonomy from the California Collision Manual [56] and NHTSA terms [57], e.g., a vehicle's front side colliding with another vehicle's rear side is a *Rear end* collision type.

   We manually extract *collision type* for 234 remaining scenarios by visually examining the 2D reconstruction. Since DeepScenario only collected vehicle dynamic parameters at 6 timestamps (starting at time 0) and for every 0.5 seconds, there are scenarios where a collision occurs between two timestamps or after the last timestamp. Thus, we identify 122 scenarios with no intersection between the AV and surrounding vehicles, such as in Figure 4. Further, we identify 112 scenarios where two sides of a vehicle are affected equally or similarly (i.e., less than or equal to 30% of differences) in the intersection area; thus, requiring further analysis to identify collision type precisely. We inspect the 2D reconstruction of those scenarios and use the same taxonomies from California [56] and NHTSA [57] to extract collision type. To the best of our knowledge, there is no boundary value to separate collision types based on vehicle's collision area, thus; we examine the data and employ an approximation of 30%.

   As our automated approach works for the overwhelming majority (77.71%) of the scenarios, we could leave out the remaining scenarios without threatening the validity of our study. Still, we opt for an additional manual round to maximise the use of the data for further evaluation and analysis.

3. *Data Filtering.* We exclude one collision scenario as we find no clear collision from the 2D reconstruction. During the manual extraction as presented in the previous step, i.e., step 2, we identify one scenario where a collision is unlikely to happen, given the relative orientation, speed, and position between the AV and other entities, thus; is filtered due to the uncertainties. As a result, we select 1049 scenarios for subsequent evaluation and analysis.

4. *Data Vectorisation.* We vectorise selected data and their attributes using the same principle (i.e., one-hot encoding [52]) for DMV California dataset as described in Section 3.2.1.

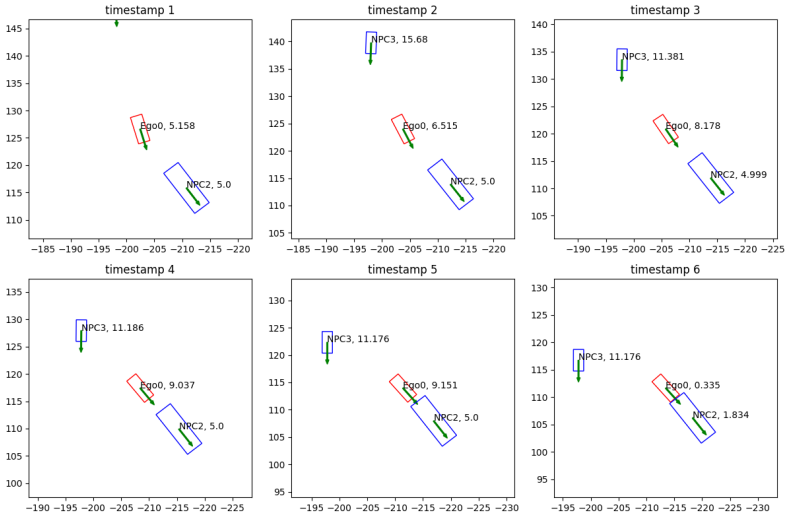

Fig. 4: An example of 2D reconstruction of a scenario description file. Each timestamp is a subplot in the figure, and we only plot entities (in blue) that are within the ego vehicle's (in red) field of interest – 20 meters from the ego vehicle – for collision type extraction. Each entity is represented as a rectangle (bounding box) with a green arrow (orientation), and a text of its name and speed in km/h separated by a comma sign. In this example, the collision is avoided in timestamp 6, but the space between Ego0 and NPC2 vehicles is not evidently visible due to the scaling issue.

3.3 Data Evaluation

The outcome of the previous steps sets the scene for applying the evaluation metrics to measure the similarity between the two datasets. This can be done at different levels: the distribution metrics can be measured at the level of individual attributes, or the combination of attributes. Subsequently, significant differences can be further scrutinised to find out whether they concern attributes that are causal for the criticality of the real scenarios. Causal analysis [48] is both computation- and data-intensive and among others, requires information about the prior distribution of the

attributes. When there is insufficient resources for causal analysis, we can only report discrepancies which are only established at the correlation level, i.e., under- or over-representation of certain attribute values with respect to the real critical scenarios. Moreover, since the data and the extracted attributes may be highly-dimensional, dimensionality reduction techniques can be used to focus the analysis of distance metrics. We instantiate these sub-steps below with respect to our two datasets.

As we described in Section 3.1, we define two metrics, namely, *attribute distribution* and *Euclidean distance*, to evaluate the realism of synthetic critical scenarios. In the empirical evaluation, we use those metrics to analyse how well DeepScenario generates similar collision scenarios as recorded by DMV California. In our evaluation, we focus on the five attributes that we extract from the two datasets, including *weather*, *lighting*, *roadway surface*, *roadway condition*, and *collision type*. As explained in Section 3.2, those attributes are used as they are extractable from both datasets.

3.3.1 Single-attribute Distribution

We first evaluate the distribution of each attribute independently to observe and analyse the differences between DeepScenario and DMV California data, as shown in Figure 5. Specifically for *weather* and *lighting*, we access their prior (actual) distribution in California to perform a causal analysis on those attributes for AV collisions. As the prior distributions for other selected attributes are not available, performing causal analysis on them is infeasible. Lastly, based on the results and analysis, we formulate our observations and propose our recommendations to improve the realism or future evaluation of realism for critical scenarios for testing AVs and ADS.

3.3.2 Multi-attribute Distribution

We then evaluate the distribution of multiple attributes together, which compares the distributions of combinations of multiple attributes between the two datasets to get further insights. Through previous analysis in Section 3.3.1, we observe *roadway surface* is strongly correlated to *weather*, and a large number of scenarios in DMV California and all in DeepScenario have no unusual *roadway conditions*. Thus, those two attributes are excluded from the analysis, and we focus on the combinations of the remaining attributes in this evaluation, including 1) *weather* and *collision type*, 2) *lighting* and *collision type*, and 3) *weather*, *lighting*, and *collision type*. The combinations reveal differences between the two datasets when considering several attributes collectively. Similar to the single-attribute distribution, we formulate our observations and propose our recommendations to improve the realism or future evaluation of realism for critical scenarios for testing AVs and ADS, based on the results and analysis.

3.3.3 Euclidean Distance

Lastly, we evaluate the Euclidean distance from DeepScenario data to DMV California data to analyse their similarity. Unlike the distribution analysis in Section 3.3.1 and 3.3.2, we focus on unique scenarios since repetitive scenarios do not contribute to the distance analysis. We start with Principal Component Analysis (PCA) [58] to reduce

the dimensions of the two datasets. After that, we perform two iterations of distance-based analysis on the data. Finally, we formulate our observations and propose our recommendations based on the results and analysis.

1. *Dimension Reduction.* In general, PCA is a statistical procedure to extract and project information from high-dimensional data into a lower-dimensional space to ease visualisation and analysis [58]. While PCA is not strictly needed for processing relatively low-dimensional data, it improves the computational efficiency and applicability of our methodology for future studies (with high-dimensional data). We perform a Scree test [59] to analyse the amount of variance in the original data that is captured by each target dimension. The number, until which the captured variance descends precipitously, but afterward levels out, is the optimal number of target dimensions for PCA [59, 60]. Then, we use PCA from *sklearn* library to transform the original data into the target number of dimensions.
2. *Distance-based Analysis.* We anchor our analysis on the concept of *Euclidean distance*. In the first iteration, we compare the distance from DeepScenario data to DMV California data with distance between DMV California data. In the second iteration, we cluster DeepScenario and DMV California data to analyse their categorisation.
   (a) In the first iteration, we compute the maximum distance for a DeepScenario data to find its nearest DMV California neighbour, and consider the scenario similar to DMV California data if the distance is smaller than the maximum distance for a DMV California data to find the nearest neighbour of its own kind. Alternatively, we compute the mean distance for a DeepScenario data to all DMV California data, and consider the scenario similar to DMV California data if the distance is smaller than the maximum mean distance for a DMV California data to the rest data of its own kind.
   (b) In the second iteration, we use K-means [61, 62] to cluster the two datasets, which is still based on the measurement of the distances between the data. To begin with, we employ the Elbow method [63–65] to determine the optimal K (i.e., the number of clusters) for K-means. The Elbow method iterates different K, fits the data, and computes the distortions (i.e., the average distance from other data entries in each cluster to the centroid). The K until which the distortion decreases significantly, but afterward, flattens out, is the optimal value for K. Then, we cluster the two datasets into K clusters, and any DeepScenario data that are grouped into a cluster with DMV California data are considered similar scenarios.

3.4 Results Assessment

This step involves presenting the results to stakeholders and discussing the background to the observed discrepancies in order to confirm or adjust the observations and draw an action plan for the future. The objective is not to perform a systematic and rigorous assessment of our findings, but rather to gather views and thoughts on them from relevant stakeholders available to us. We interviewed the first author of DeepScenario, as they led the generation of the DeepScenario dataset and was available to participate

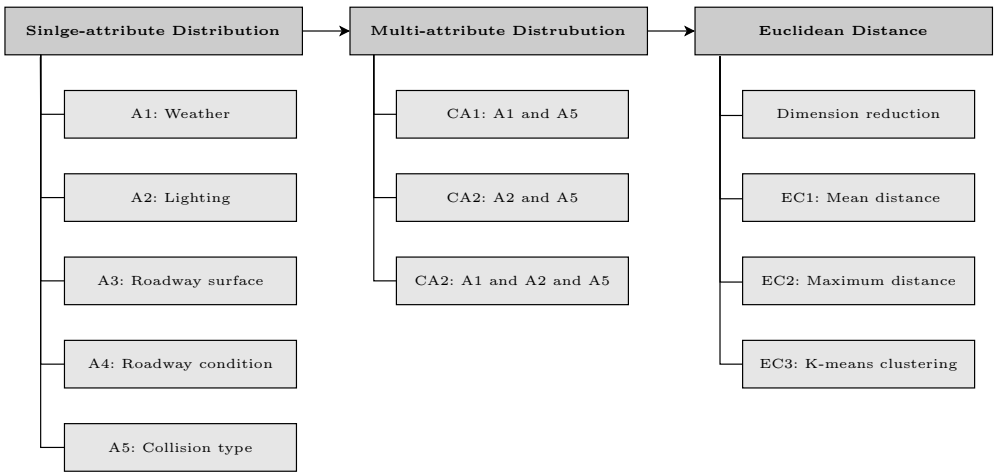

Fig. 5: An overview of the data evaluation. In single-attribute distribution analysis, we compare the distribution of five attributes (A1–A5) between the two datasets. In multi-attribute distribution analysis, we compare the distribution of three combinations of attributes (CA1–CA3). Lastly, for Euclidean distance, we first perform PCA (Principle Component Analysis) to reduce the dimensions of the data, and perform distance analysis using three evaluation criteria (EC1–EC3).

our interview. We consciously used a semi-structured interview to give the interviewee freedom and flexibility to discuss the validity and usefulness of our findings.

From our proof of concept, we observed discrepancies with respect to the synthetic critical scenarios generated by DeepScenario, as well as some missing background information with respect to the DMV California set. We discussed these observations, with the first author of DeepScenario, in an online semi-structured interview to assess the results and analysis we obtain from the empirical evaluation, as described in Section 3.3. Before the interview, we presented the design, results, observations, and our analysis to the interviewee. During the interview, we had an open discussion to collect further insights, feedback, and suggestions from them. The interview process was flexible where we walked through the findings of this study and asked the interviewee's thoughts and feedback on them. Based on their responses, we continued the discussion to explore additional insights with the interviewee. After that, we analysed the response and presented relevant parts as assessments of our results. The interview was semi-structured and primarily aimed to check the outcomes and discuss the feedback from the DeepScenario main author as they are aware of the background and the design decisions for this dataset and its generation process.

4 Results and analysis

In line with the evaluation design in Section 3.3, we present results and analysis to each evaluation analysis, which concerns primarily **RQ1** – how realistic are synthetic critical scenarios from realistic critical scenarios. Also, we provide our observations and

recommendations based on the results to address **RQ2** – guidelines for closing reality gaps in synthetic critical scenarios. The data and scripts we use for evaluation are available on Zenodo [51]. Although the quantitative results are specific to the analysed datasets, we reflect on the quantitative results to find qualitative observations and recommendations that are generaliseable to other analyses of realism in the future.

4.1 Single Attribute Distribution

As introduced in Section 3.3.1, we focus on analysing distributions of five relevant attributes independently in single attribute evaluation, including *weather*, *lighting*, *roadway surface*, *roadway conditions*, and *collision type*.

4.1.1 Evaluation Results

1. *Weather.* As shown in Table 1, weather distribution in DeepScenario differs significantly from DMV California. DeepScenario has two weather conditions almost evenly distributed, i.e., 50.81% *Clear* and 49.19% *Raining*. In contrast, DMV California has the majority (86.53%) *Clear*, a certain amount of *Cloudy* and *Raining*, and several *Fog/visibility*. As DeepScenario used only two weather conditions, it is unsurprising that other weather recorded in DMV California such as *Cloudy* and *Fog/visibility* have no occurrence in its distribution. In addition, DeepScenario employs each weather uniformly in its generation process, resulting in an even distribution, and have more frequent collisions in the *Raining* weather than in DMV California.

Weather	**DMV California**		**DeepScenario**	
	Population (282)	Distribution	Population (1049)	Distribution
Clear	244	86.53%	533	50.81%
Cloudy	21	7.45%	–	–
Raining	14	4.97%	516	49.19%
Fog/visibility	3	1.06%	–	–
Snowing	0	0	–	–
Wind	0	0	–	–
Other*	0	0	–	–

Table 1: Weather distribution for DMV California and DeepScenario. Sign '–' means a weather is not used and thus not applicable. Other* refers to weather not listed above, for example, hail, dust, or smoke, as defined in California Collision Manual [56].

We also visit the actual weather distribution in California (i.e., 18.63% days with precipitation for San Francisco [66] and 11.78% for San Diego [67] in 2023) to analyse potential causal effects of weather on AV collisions. The actual distribution suggests no evident impact of rain on collisions for DMV California data as only 4.97% collisions were reported in this weather. However, that is subject to the test arrangement of the manufacturers – how AV manufacturers arranged their tests in different weather, which is unavailable presently.

Furthermore, we conduct a one-way Chi-square test [68] with the distribution of Raining and non-Raining weather in DMV California data against the actual distribution in San Francisco to determine the association between them. The result (i.e., *statistic* = 35.340035, *pvalue* = 2.76882e-09) indicates that their difference is significant. To identify the causal effect of rain on AV collisions, we need more statistics on test arrangement, traffic density, driving behaviours, and so on in each weather to perform further causal inference [48].

Lighting	**DMV California**		**DeepScenario**	
	Population (282)	Distribution	Population (1049)	Distribution
Daylight	177	62.77%	529	50.43%
Dark – Street lights	92	32.62%	520	49.57%
Dusk – Dawn	12	4.26%	–	–
Dark – No street lights	1	0.36%	–	–
Dark – Street lights NFG	0	0	–	–

Table 2: Lighting distribution for DMV California and DeepScenario. Sign '–' means a lighting condition is not used and thus not applicable. 'NFG' stands for 'Not functioning' and is abbreviated due to space limitation.

2. *Lighting.* Table 2 presents the distribution of *lighting* for DMV California and DeepScenario. Similar to *weather*, DeepScenario has each lighting condition equally explored; we see a fairly close distribution for *Daylight* and *Dark – Street lights*. In reality, DMV California recorded the same lighting as the two most common lighting conditions in AV collisions, but also experienced a few *Dusk – Dawn* and *Dark – No street lights*. As one may inevitably consider other lighting than *Daylight* would impair the visibility of AVs and other road users and expect more collisions, their distribution are much lower than *Daylight*, and raises our concerns about the test arrangement for DMV California and simulation quality for DeepScenario. In other words, if most testing happened during the day, more collisions would be reported in *Daylight* condition for DMV California; if the simulator could not faithfully represent the lighting and potential impacts on AV sensors, they would not contribute to collisions to a real extent in DeepScenario.

   Like the analysis for *weather*, we also visit the day length in San Francisco [69] and obtain an average of 12:12 (hh:mm) *Daylight*, 0:56 *Dusk – Dawn*, and 10:52 night (*Dark – Street lights*) in 2022, corresponds to a distribution of 50.83%, 3.89%, and 45.28% respectively. The actual distribution of lighting conditions suggests no distinctive causal effects of inclement lighting such as *Dark – Street lights* on collisions. Besides, the Chi-square test (i.e., *statistic* = 17.745138, *pvalue* = 0.000140) uncovers that the observed distribution of lighting condition in DMV California does not follow the actual distribution in San Francisco and the difference is significant. It may be due to uneven tests in different lighting conditions, or poor lighting does not cause more collisions, which needs further investigation.
3. *Roadway surface.* As we introduced earlier in Section 3.2.2, DeepScenario sets the wetness of the roadway based on weather, therefore; the distribution of *roadway surface* is the same as *weather*, as shown in Table 3. DMV California encountered

Roadway Surface	**DMV California**		**DeepScenario**	
	Population (282)	Distribution	Population (1049)	Distribution
Dry	265	93.97%	533	50.81%
Wet	17	6.03%	516	49.19%
Snowy – Icy	0	0	–	–
Slippery	0	0	–	–

Table 3: Roadway surface distribution for DMV California and DeepScenario. Slippery refers to Slippery (muddy, oily, etc.) in full name. Sign '–' means a roadway surface is not used and thus not applicable.

a predominant majority of *Dry* and a few *Wet* roadway surfaces, which is close to the distribution of weather in Table 1. We conduct a Pearson correlation coefficient analysis of weather and roadway surface for DMV California, and the result (i.e., $\rho = 0.536414$) indicates a strong positive correlation between them. In a further analysis, we discover that 98.51% (264/268) of *Clear*, *Cloudy*, and *Fog/visibility* weather correspond to a *Dry* roadway surface, and 92.86% (i.e., 13/14) of *Raining* weather has a *Wet* roadway surface.

Roadway Conditions	**DMV California**		**DeepScenario**	
	Population (282)	Distribution	Population (1049)	Distribution
No unusual conditions	273	96.81%	1049	100%
Obstruction on roadway	3	1.06%	–	–
Construction – Repair zone	2	0.71%	–	–
Reduced roadway width	2	0.71%	–	–
Holes, deep rut	2	0.71%	–	–
Loose material on roadway	0	0	–	–
Flooded	0	0	–	–
Other*	0	0	–	–

Table 4: Roadway condition distribution for DMV California and DeepScenario. Sign '–' means a roadway condition is not used and thus not applicable. Others* refer to roadway conditions not listed above and include such as oil slick on the road.

4. *Roadway condition.* We focus on DMV California data for *roadway condition* as DeepScenario set *No unusual conditions* for scenario generation. Despite that DMV California also ran into *No unusual conditions* for the vast majority (i.e., 96.81%) of the collisions, they experienced a few *Obstructions on roadway*, *Construction – Repair zone*, *Reduced roadway width*, and *Holes, deep rut*, as we can see from Table 4. As one may consider unusual roadway conditions would remarkably challenge AVs in their driving tasks and result in more collisions, such situations are rare in comparison to usual roadway conditions [22, 33, 70], and the resulting distribution still depends on the testing arrangement of different AV manufacturers. For example, if manufacturers only test their AVs under *No unusual roadway conditions*, then collisions are expected exclusively in this roadway condition, and there will be no distribution of collisions for other roadway conditions. However,

that does not necessarily mean *No unusual roadway conditions* lead to more collisions for AVs than other roadway conditions. To understand the causal relations between roadway conditions and AV collisions, we need to know how manufacturers have arranged their tests under each different roadway condition. Therefore, a causal relation cannot be established with current information available.

Collision type	**DMV California**		**DeepScenario**	
	Population (282)	Distribution	Population (1049)	Distribution
Rear end	179	63.48%	434	41.37%
Side swipe	49	17.38%	363	34.60%
Broadside	24	8.51%	109	10.39%
Other*	11	3.90%	0	0
Head-on	10	3.55%	9	0.86%
Hit object	8	2.84%	28	2.67%
Vehicle/pedestrian	1	0.36%	106	10.11%
Overturned	0	0	0	0

Table 5: Collision type distribution for DMV California and DeepScenario. Other* refers to collision types not listed herein and include such as a vehicle involved with a bicycle, train, animal, falling passengers from a vehicle, a bicycle involved with a pedestrian, etc.

5. *Collision type.* DeepScenario and DMV California share the same top three *collision type* (i.e., *Rear end*, *Side swipe*, and *Broadside*), of which the cumulative total constitutes 86.37% and 89.36% of all collisions respectively. Apart from that, there are 11 *Other* collisions in DMV California, but none for DeepScenario, as in Table 5. That is because DeepScenario employed only pedestrians and vehicles in scenario generation, and no other road users such as bicycles, trains, and animals. In addition, DeepScenario experienced a significantly higher ratio of *vehicle/-pedestrian* collisions than DMV California. That calls for further scrutiny as to how different pedestrian behaviour models were simulated in DeepScenario and how testing with pedestrians was arranged in DMV California data. In a subsequent analysis, we sorted collision types for the two datasets in descending order and computed the distribution gaps of adjacent types. We obtained a sum of 0.63 and a mean of 0.09 difference for DMV California. In comparison, DeepScenario has a sum of 0.41 and a mean of 0.06 difference, which implies DeepScenario has on average a lower difference of distribution for different collision types and generated collision of each type more uniformly.

4.1.2 Observations and Recommendations

Given the attribute distribution, as presented in Section 4.1.1, we formulate two observations (starting with O), derive three recommendations (starting with R), and receive three comments (starting with C) from the first author of DeepScenario.

O1 – DeepScenario used a strict subset of realistic parameter values; this may be due to the design of the critical scenario generation and/or due to the limited capabilities of the simulator to reflect the effect of a number of parameters.
O2 – There are discrepancies in attribute distribution in DeepScenario compared to the real scenarios in DMV California. Some of these may be due to test arrangement in DMV California and others may be due to design decisions in DeepScenario, partly due to the quality of simulations.

1. O1 is evident in the evaluation, where *substantial realistic parameter values are not used* in DeepScenario. Especially, DeepScenario used only two weather and lighting conditions, and two roadway surfaces, which left out a variety of realistic values for weather and lighting conditions, roadway surface, and roadway conditions. Consequently, more potential AV collision scenarios with realistic conditions as evidenced by DMV California are not explored in DeepScenario.

   O2 is highly likely yet requires further investigation to better understand the ground truth. As per our analysis, there is no clear picture of the actual distribution of all attributes in California, when and under which conditions the manufacturers have tested their AVs on public roads, and how well real-world attributes and their impacts are represented in simulation in DeepScenario. Thus, even though there is a significant difference between the two datasets, *no simple conclusions can be made on causality* given the information we have. Nevertheless, our observations from the single attribute evaluation reveal the discrepancies in distribution and draw attention to some considerations of scenario realism.

R1 – Include a more diverse range of realistic parameter values in synthetic scenario exploration, such as snowy weather, icy roadway surface, and unusual roadway conditions.
  R1.1 Use the distribution of parameters in realistic critical scenarios to inform the distribution of parameters in synthetic simulations.
  R1.2 If causal analysis is performed on realistic data, a similar representation of causal parameters should be prioritised in simulation.
R2 – Evaluate the quality of the simulation before using it for testing AVs.
  R2.1 – Evaluate the feasibility of simulating real-world parameters, e.g., different weather.
  R2.2 – Evaluate the representation and effects of real-world parameters in simulation.
  R2.3 – Decouple scenario generation from a single simulator and use a range of simulators with different capabilities to explore more diverse critical scenarios.
R3 – Include more contextual information for real critical scenario datasets on the prior distribution of parameters to enable a causal analysis of AV collision scenarios. These include statistical information about planned field tests and traffic situations during field tests that were not recorded as critical.

2. In response to O1, we recommend R1 to *include more realistic parameter values* in the exploration of synthetic collision scenarios to improve the diversity of them in comparing to realistic AV collisions. This recommendation is inclusive, meaning parameters should be considered from all realistic perspectives instead of attributes of interest in this study. While layered frameworks for systematic parameter selection in AV testing are developed and reported [71, 72], a comprehensive list of parameter values and the selection of them for different ADS with different functions, implementations, or ODDs are still lacking [8].

   Also, it is significant to *evaluate the quality of simulation* (R2), regarding whether desired parameters are feasible to simulate (R2.1) and the realism of their representations and effects on other entities (R2.2). For a concrete example, how rainy weather in varying degrees is represented in simulation and how road surface, tire friction, and sensor performance are impacted shall be quantified. Particularly, studies already show substantial gaps between simulation and real-world testing [14, 73–76], and the latest EU regulation for type approval of AVs also mandates evaluating simulation if a simulator is used in testing [8, 10]. Additionally, we recommend incorporating several simulators with different capabilities to explore more diverse critical scenarios (R2.3). For example, one simulator is not feasible to simulate a particular weather; thus, not able to explore critical scenarios under this weather condition, but other simulators are feasible. In this case, one could use several simulators to explore different conditions they are capable of representing. Further, if we have several simulators simulating the same conditions but getting different results, for example, in one simulator the AV collides with other road users but not in other simulators, that would indicate a discrepancy in simulation quality and call for further investigation of whether that particular simulator is representing the conditions faithfully or not. Using multiple simulators enables us to explore more and diverse critical scenarios and getting more insights into the critical scenarios identified.

   Overall, our recommendations emphasize incorporating realistic parameters in exploring critical test scenarios for AVs, and ensuring they are faithfully simulated in the simulator. In addition, we also recommend R3 to *explore and include the prior distribution* of parameters used in real-world critical scenarios to perform a causal analysis of them on AV collisions and identify causal parameters that are significant for AV collisions. That, in turn, encourages organisations such as different AV manufacturers to collect and share more contextual information such as the test arrangement, which would be rather significant for evaluating, analysing, and understanding the realism of synthetic scenarios.

C1 – Scenario simulation and optimisation are computationally heavy, which limits the feasibility of incorporating more realistic parameters and values in DeepScenario.

C2 – SVL simulator supports limited parameter values from DMV California. An example is snowy weather and unusual roadway conditions such as construction or holes.

C3 – Reality gap in the SVL simulator is noticeable from different perspectives, e.g., weather.

3. Generally, the first author of DeepScenario confirms our results. As C1 sounds, *optimisation and simulation of critical scenarios are computational resources- and time-consuming*, which limited them in incorporating more real-world parameters. That limitation was exacerbated by the SVL simulator in which *certain parameters were not supported* (C2). For example, snowy and windy weather. Although a road damage level that changes road friction can be set, many unusual roadway conditions such as construction or holes were infeasible in SVL when DeepScenario was created. Lastly, the *reality gap* (C3) was evident in the SVL simulator for, e.g., the effect of weight on vehicle dynamics, where a vehicle could be thrown into the air abnormally after a collision. Another example is that the weather stayed constantly the same and an update instruction would change it immediately rather than incrementally over time. Those comments supply additional insights into the realism of scenarios in DeepScenario, and also reveal specific issues or limitations that impact the realism of the generated scenarios.

4.2 Multiple Attribute Distribution

As introduced in Section 3.2.2, we focus on analysing distribution of three combinations of attributes in multiple attribute evaluation, including 1)*weather* and *collision type*, 2) *lighting* and *collision type*, 3) *weather*, *lighting*, and *collision type*.

4.2.1 Evaluation Results

1. *Weather* and *collision type*. Table 6 presents the distribution of *weather* in conjunction with *collision type*. Herein, we only focus on four weather that are recorded by DMV California or used by DeepScenario. Similar to the *weather* and *collision type* analysis from Section 4.1.1, DMV California has most collisions reported in *Clear* weather, thus; collision types are further decomposed under this weather. In contrast, DeepScenario has each collision type more evenly distributed in *Clear* and *Raining* weather, and has remarkably more collisions in *Raining* weather. As highlighted in Table 6, DeepScenario did not generate *Other* type collisions in *Clear* weather, which have been reported in DMV California, but DeepScenario was able to identify *Broadside* and *Vehicle/pedestrian* collisions in *Raining* weather, which have not been reported in DMV California yet.
2. *Lighting* and *collision type*. The distribution of *lighting* in conjunction with *collision type* somewhat mirrors the separate distribution of *weather* and *collision type*. Particularly, DMV California recorded collisions predominantly in *Daylight*, while DeepScenario has a more uniform distribution for each collision type in *Daylight* and *Dark – Street lights* lighting, as shown in Table 7. *Other* type collisions are reported both in *Daylight* and *Dark – Street lights* in DMV California but none in DeepScenario. Nevertheless, *Vehicle/pedestrian* collisions, which have not been reported in *Dark – Street lights* in DMV California as of now, are identified in a considerable ratio (4.67%) in DeepScenario.

Collision type \ Weather	Clear	Cloudy	Raining	Fog/visibility
Rear end	54.26% / 21.74%	6.38% / –	2.13% / 19.64%	0.71% / –
Side swipe	16.31% / 17.83%	0.36% / –	0.36% / 16.78%	0.36% / –
Broadside	7.45% / 5.15%	0.36% / –	**0 / 5.24%**	0 / –
Other*	**3.90% / 0**	0 / –	0 / 0	0 / –
Head-on	3.19% / 0.29%	0 / –	0.36% / 0.57%	0 / –
Hit object	1.06% / 1.53%	0.36% / –	1.42% / 1.14%	0 / –
Vehicle/pedestrian	0.36% / 4.29%	0 / –	**0 / 5.82%**	0 / –
Overturned	0 / 0	0 / –	0 / 0	0 / –

Table 6: Distribution of weather recorded in collisions in conjunction with collision types. Each cell contains distributions from DMV California and DeepScenario, and is separated by a '/' sign. Columns headed Cloudy and Fog/visibility are grayed out as they are not applicable for DeepScenario, thus; their distributions are annotated as '–'. The distributions in bold imply only one dataset has a distribution.

Collision type \ Lighting	Daylight	Dusk – Dawn	Dark – Street lights	Dark – No SL
Rear end	43.97% / 20.11%	3.19% / –	15.96% / 21.26%	0.36% / –
Side swipe	9.22% / 18.11%	1.06% / –	7.09% / 16.49%	0 / –
Broadside	4.97% / 4.67%	0 / –	3.55% / 5.72%	0 / –
Other*	**1.77% / 0**	0 / –	**2.13% / 0**	0 / –
Head-on	1.42% / 0.57%	0 / –	2.13% / 0.29%	0 / –
Hit object	1.06% / 1.53%	0 / –	1.77% / 1.14%	0 / –
Vehicle/pedestrian	0.36% / 5.43%	0 / –	**0 / 4.67%**	0 / –
Overturned	0 / 0	0 / –	0 / 0	0 / –

Table 7: Distribution of lighting recorded in collisions in conjunction with collision types. Each cell contains distributions from DMV California and DeepScenario, and is separated by a '/' sign. Column heading Dark – No SL is short for 'Dark – No street lights' due to space issues. Columns headed Dusk – Dawn and Dark – No SL are grayed out as they are not applicable for DeepScenario, thus; their distributions are annotated as '–'. The distributions in bold imply only one dataset has a distribution.

3. *Weather*, *lighting*, and *collision type*. In a higher-order multi-attribute analysis, we obtain a more meticulous distribution of *collision type* in conjunction with *weather* and *lighting*, as shown in Table 8. We exclude weather and lighting conditions that do not apply to DeepScenario or are not reported in DMV California as they do not provide additional insights to our analysis. DMV California features most collisions recorded in *Clear-Daylight*, followed by *Clear-Dark – Street lights*, and very few in *Raining* weather. In comparison, DeepScenario has each collision type more uniformly distributed. As one may expect inclement weather (e.g., *Raining*) and lighting conditions (e.g., *Dark – Street lights*) may increase the chance of collisions, we see no distinctive distribution that deviates *Raining-Dark – Street lights* from the others. Especially, DMV California has 54.97% of collisions reported in *Clear* weather and *Daylight* lighting. As we presented in Section 4.1.1, the actual distribution of *weather* and *lighting conditions* suggests

inclement weather and lighting do not associate with collisions to a significant degree. That, however, should be further analysed with how each different condition is tested by the manufacturers. Lastly, except for *Other* type collisions that are recorded in *Clear* weather for DMV California but none in DeepScenario, DeepScenario identified six collision types under specific weather or lighting conditions that are not reported in DMV California. Among them, five are in *Raining* weather, and three are *Vehicle/pedestrian type*.

Weath-Light / Coll_type	Clear-Daylight	Clear-Dark*	Raining-Daylight	Raining-Dark*
Rear end	38.30% / 10.58%	13.83% / 11.15%	1.06% / 9.53%	1.06% / 10.11%
Side swipe	8.51% / 9.91%	6.74% / 7.91%	0 / 8.20%	0.36% / 8.58%
Broadside	4.61% / 2.0%	2.84% / 3.15%	0 / 2.67%	0.71% / 2.57%
Other*	1.77% / 0	2.13% / 0	0 / 0	0 / 0
Head-on	1.06% / 0.19%	2.13% / 0.10%	0.36% / 0.38%	0 / 0.19%
Hit object	0.36% / 0.86%	0.71% / 0.67%	0.36% / 0.67%	1.06% / 0.48%
Vehicle/pedestrian	0.36% / 2.19%	0 / 2.10%	0 / 3.24%	0 / 2.57%
Overturned	0 / 0	0 / 0	0 / 0	0 / 0

Table 8: Distribution of collision types in conjunction with weather and lighting. Each cell contains distributions from DMV California and DeepScenario, and is separated by a '/' sign. Due to space issues, the column heading Coll_type is short for Collision type, Weath-Light for Weather-Lighting, Clear-Dark* for 'Clear-Dark – Street lights', and Raining-Dark* for 'Raining-Dark – Street lights'. The distributions are grayed out if only one dataset has a distribution.

4.2.2 Observations and Recommendations

Same as in Section 4.1, we present our observations, recommendations, and comments from the DeepScenario author for multiple attribute evaluation.

O1 – DMV California recorded most collisions, i.e., 54.97%, in clear weather and daylight conditions, thus; inclement weather and lighting do not hold a higher distribution for AV collisions. Yet, further statistics for such as test arrangement and traffic densities are needed to understand the causal effect of inclement weather and lighting conditions on AV collisions.

O2 – DeepScenario identified new collisions that have not been reported to DMV California such as vehicle/pedestrian collisions in raining weather, while missing out on some that existed in DMV California, e.g., other type collisions in clear weather as in Table 8.

1. We do not observe a significantly higher distribution of collisions in inclement weather (e.g., *Raining*) and lighting (e.g., *Dusk – Dawn*, *Dark –Street lights*) in DMV California. Instead, most collisions from DMV California were recorded in *Clear* weather and *Daylight* lighting. Therefore, one observation is that they *do not contribute to more collisions for the datasets in this study* (O1). However, that should be further analysed with additional statistics such as test arrangement by

the manufacturers, traffic densities, and interactions between AV and other road users in DMV California data. Therefore, a causal relation cannot be established.

DeepScenario *identified new collision scenarios that have not been reported* to DMV California (O2), which clearly shows it is effective to use synthetic scenario exploration to find new critical scenarios for testing AVs/ADS. However, DeepScenario did not use all realistic parameter values from DMV California and missed out on certain collisions that have been reported to DMV California, so synthetic scenario exploration should complement, rather than replace, real-world testing or realistic scenarios in the current stage [4, 8, 77].

R1 – Identify and use critical parameters for exploring synthetic critical scenarios.
- R1.1 – Explore the prior distributions of parameters that are relevant for composing critical scenarios (as illustrated by steps 1-3 in Figure 6), perform causal analysis with the observed distributions in collision scenarios (see steps 4-5 in Figure 6), and identify which parameters are critical for AV collisions.
- R1.2 – Ensure the feasibility of simulating those critical parameters from R1.1, and validate their representation and effects in simulation environment to ensure a faithful simulation and scenario realism.

R2 – Use realism measurement in a feedback loop between the real and synthetic scenarios: when new types of collisions are found in synthetic scenarios, use field tests to verify the criticality of these parameter values and adjust the parameter distribution. This will lead to an updated measure for realism and a virtuous loop for more covering tests both in real- and simulated environments.

2. Even though O1 is a specific observation for this study and is subject to data acquisition and size of data for analysis, it is still important to identify parameters that are critical for causing AV collisions (R1.1) and how realistic they are represented in simulation for critical scenario exploration (R1.2) to reduce reality gaps for simulation environments, as articulated in several studies [14, 74, 77].

   Moreover, a feedback loop needs to be developed for integrating real-world testing and simulation testing of AVs concerning guidelines for field testing with synthetically generated scenarios. Specifically, simulation can be used to explore unknown critical test scenarios (e.g., collisions or other hazardous scenarios) and improve scenario coverage of real-world testing. Correspondingly, real-world testing can give useful feedback to improve the parameters and distribution of values in simulation testing and reduce the reality gaps (R2). Overall, they should complement each other and be combined effectively for testing AVs [78–80].

C1 – Evaluation and analysis of the realism of synthetically generated scenarios are significant for effective testing of AVs/ADS.

C2 – DeepScenario should be maintained continuously, enabling more contributors.

3. The first author of DeepScenario confirms our analysis, observations, and recommendations. Particularly, they believe that *evaluation and analysis of synthetically generated scenarios* from realistic driving scenarios is extremely important (C1) as realism is an essential quality for test scenarios, and suggest every study in the field to incorporate more realistic parameter values, AD systems, and related tools to increase realism of test scenarios for AVs/ADS. Besides, they also articulate testing AVs and critical scenario exploration as continuous work, meaning working on and enhancing them gradually and iteratively over time.

   DeepScenario, as an open driving scenario dataset for testing AVs/ADS, needs to be maintained constantly and enables more subscribers/contributors to refine it (C2). Several proposals were discussed, including to 1) transfer the scenarios into OpenScenario format, which is a fairly standard format for scenarios that are commonly used [4, 25, 38, 81], 2) switch to the Carla simulator, which is a robust and common simulation tool for ADS [4, 82–85], 3) update the scenarios to use more accurate specifications, for such as GPS positions, 4) continuously expand the scenario set by exploring other maps, realistic weather, roadway conditions, road infrastructures, and interactions with various road users, 5) keep open source the tool and scenario set. With that, DeepScenario is expected to attract more researchers or practitioners to use and improve the test scenarios.

4.3 Euclidean Distance

As described in Section 3.3.3, we evaluate DeepScenario data with DMV California data based on the measurement of Euclidean distance. We use PCA to reduce the dimensions of the vectorised data. Then, we use distance criteria and K-means approach to evaluate the data. We employ unique data entries in the two datasets, i.e., 40 for DMV California and 24 for DeepScenario.

4.3.1 Evaluation Results

1. *Dimension Reduction.* As described in Section 3.3.3, we perform a Scree test [59] and find 6 the optimal number of target dimensions, which can capture 96.08% of the variance in the original data (i.e., the union of unique data entries in DeepScenario and DMV California). Then, we use PCA from *sklearn* library and transform the vectorised data of DeepScenario and DMV California from 32 into 6 dimensions for subsequent evaluation.
2. *Distance-based analysis.*
   (a) In the first iteration, we discover all DeepScenario data are similar to DMV California data as they are close in distance comparison.
      (i) 17 out of 24 unique DeepScenario scenarios find an identical copy in DMV California data. In addition, the maximum distance for a DeepScenario data to find its nearest DMV California neighbour (1.07) is much smaller than the maximum distance for a DMV California data to find the nearest neighbour of its own kind (1.40). Similarly, the mean distance for a DeepScenario data to find its nearest DMV California neighbour

(0.17) is also much smaller than the mean distance for a DMV California data to find the nearest neighbour of its own kind (0.68), with the same standard deviation (0.3). That indicates, while DMV California data is more scattered in distance in general, DeepScenario data appear to stick around the DMV California data and are easier to find a DMV California neighbour than DMV California data.

(ii) The mean distance for a DeepScenario data to all DMV California data (maximum is 2.23) is smaller than the maximum mean distance (2.54) for a DMV California data to the rest data of its own kind. That implies every DeepScenario data has a shorter average distance to DMV California data than the mean distance of the farthest DMV California data to rest data in its own group. We conclude that every DeepScenario data is still within the boundary of the DMV California data.

(b) In the second iteration, we first use the Elbow method and find 6 the optimal number of clusters for our data. Then, we use K-means from the *sklearn* library to cluster the two datasets. The results indicate all six clusters are a mixture of data from both datasets, not a single cluster that contains DeepScenario data exclusively. The clustering does not separate DeepScenario and DMV California data, given the attributes we extract.

4.3.2 Observations and Recommendations

Based on the distance evaluation in Section 4.3.1, we formulate our observations and recommendations. Except for confirming our findings and giving general feedback, the first author of DeepScenario supplies no specific insights or concerns for this part.

O1 – DeepScenario identifies new collision scenarios that are not recorded by DMV California, but are still considered similar to DMV California data according to the distance analysis.

O2 – The current datasets provide limited attributes that are extractable, especially DeepScenario due to no functioning simulators available. Given more relevant attributes are available or extractable such as vehicle maneuvers, we expect a further evaluation of the realism of DeepScenario and more sophisticated comparison of the two datasets to be feasible.

1. Further elaborating on the distribution discrepancies presented in Sections 4.1 and 4.2, the distance evaluation discloses that DeepScenario contains different scenarios from DMV California, but *they are not overly different* from scenarios reported in DMV California from a distance perspective (O1). As reported in the distance-based analysis in Section 4.3.1, DeepScenario data is close to DMV California data in a vectorized space, and the Euclidean distance from DeepScenario data to DMV California data are generally lower than the Euclidean distances within DMV California data. Therefore, they are considered not significantly different from DMV California data from a distance point of view. Also, we refer to Section 3.1.2 again, discrepancies in attribute distribution does not necessarily mean scenarios contained in the two datasets are different.

We also observe that the *current datasets we analyse contain limited attributes that are available or extractable*, thus not allowing more thorough evaluations (O2). As we mentioned in Section 3, DMV California data contains additional information such as *Accident details – Description*, *Movement Proceeding Collision*, and *Other Associated Factors*, which can be useful to reveal further insights, but are not included in this study due to the limitations of DeepScenario and SVL simulator. Particularly, the *Accident details – Description* describes such as the location, traffic situation, cause of the collision, trajectories, maneuvers, and post-accident actions for each involved entity. *Movement Proceeding Collision* has 18 options (e.g., Stopped, Making right turn, Backing, Changing lanes, and Parking) and clearly indicates the movement and behaviour of each entity. Similarly, *Other Associated Factors* has 12 options (e.g., Vision obscurement, Entering/Leaving ramp) and gives additional information about a collision.

R1 – Develop guidelines of which attributes and in which formats should a scenario contain for future scenario exploration and collection.
R2 – Include more attributes pertaining to the dynamics and kinematics of vehicles, their relative positions and maneuvers in future scenario evaluation and analyses.

2. In response to O2, we recommend *developing guidelines for defining future critical scenarios* regarding which attributes (e.g., weather and lighting) are required, their formats (e.g., categorical or numeric forms), and collection frequency (R1). More specifically, how each attribute and attribute value are collected, and in which format and precision each attribute is defined, are also important to understand. For example, DeepScenario used three floating points' precision for the GPS coordinates of each entity, which does not form an accurate location. Otherwise, the GPS coordinates can be mapped back to a map of San Francisco and obtain such as the vehicle trajectory, location (e.g., an intersection), and some maneuvers (e.g., lane switching) involved in a scenario even without a simulator.

   Furthermore, we recommend *evaluating and analysing the realism of synthetic scenarios using more attributes* from R1, to close the reality gap for synthetically generated scenarios (R2). Given that DeepScenario and DMV California can provide accurate information regarding those attributes, more perspectives of the scenarios can be compared for realism.

5 Discussion

Going beyond the factual results and analysis in Section 4, we discuss our findings, limitations, and future work concerning our research goals and questions as well as their implications. Besides, we present threats to the validity of our study and how we have mitigated them appropriately.

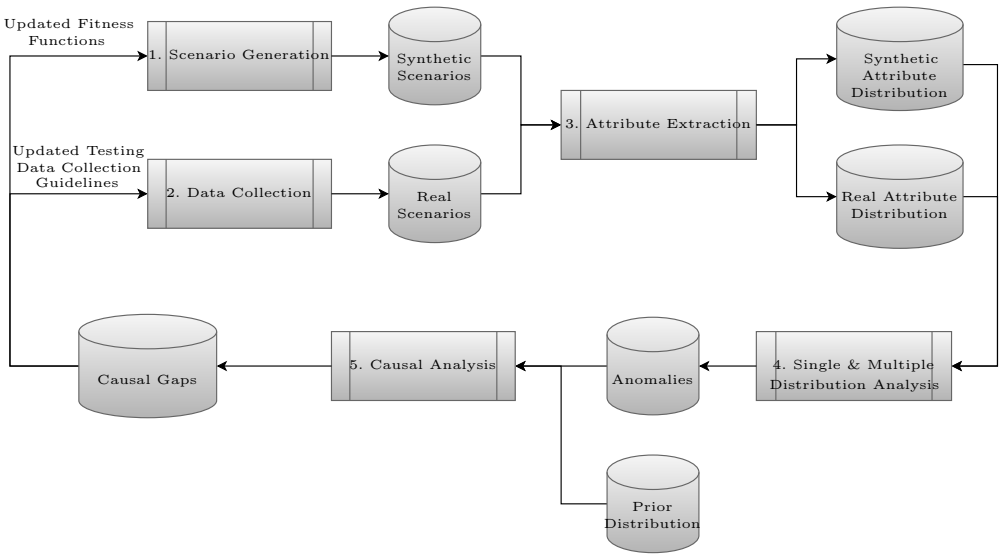

Fig. 6: Process for continuous evaluation and improvement of realism for critical scenarios, based on the findings of our pilot study. The process consists of several steps (in rectangles and annotated with 1–5) and output (in cylindrical) from each step.

5.1 Findings and Implications

Realism is an essential quality to evaluate not only to critical scenarios, but all relevant test scenarios in general [8]. However, it is not sufficiently addressed in the current research. In this study, we propose a methodology using two metrics, i.e., *attribute distribution* and *Euclidean distance*, to evaluate the realism of synthetic critical scenarios. As a proof of concept, we employ two AV collision sets, including a synthetic one from *DeepScenario* and a realistic one from *DMV California* for empirical evaluation of our methodology. We evaluate the similarity of DeepScenario data by analysing the attribute distribution with and distance to DMV California data. After the evaluation, we assess our results by interviewing the DeepScenario author. Based on our findings from the pilot study, we propose a continuous process, depicted in Figure 6. In this process, we propose to use our methodology for a continuous analysis of realism and using the findings, potentially after a causal analysis, to address the gaps by updating the fitness functions for the synthetic scenarios and guidelines for field testing and data collection for real scenarios.

In addition to this general recommendation, we analyse the specific findings to answer our research questions below.

1. Regarding **RQ1.1**, we observe *the attribute distribution between the two datasets differs significantly*. While DMV California had substantial AV collisions in *Clear* weather, *Daylight* lighting, *Dry* roadway surface, and *No unusual* roadway conditions, DeepScenario explored and identified collision scenarios evenly in each selected condition. For example, DMV California has 86.53% of collision scenarios recorded in Clear weather while DeepScenario has 50.81%. However, that does

not necessarily mean the scenarios contained in the two datasets are different. In the distance-based analysis, we take each individual scenario from DeepScenario and find whether a similar scenario appears close by in DMV California. The result reveals that the two datasets, although possessing several different scenarios, are still similar. Overall, *attribute distribution and Euclidean distance are considered effective metrics to quantify similarities between two critical scenario sets, further indicating the realism of the synthetic critical scenarios.*

As we reported in Section 4.1 and 4.2, we find no evident impact of inclement weather or lighting on AV collisions; it requires, however, further investigation for **RQ1.2** to identify which attributes are critical for causing more collisions for AVs in the real world, and subsequently, **RQ1.3** – how the identified casual parameters from real world are reflected in synthetic collision scenarios. Therefore, we are unable to answer these research questions in this study, given the datasets and contextual information we have, e.g., test arrangements for DMV California data.

In summary, for **RQ1**, we conclude, based on our methodology and results, that *DeepScenario is similar and exposes no big differences to realistic collisions, although it does identify new scenarios not recorded in DMV California.* However, that depends on how different AV manufacturers have arranged their tests. Also, as we discussed in Section 3.3, the current study evaluates limited attributes while other attributes of interest are not available or not feasible to extract.

2. For **RQ2**, we observe that more realistic attributes and values should be incorporated in the exploration of synthetic critical scenarios, and the quality of simulation should be properly evaluated with respect to real-world testing before using it for testing AVs/ADS.

   Specifically, DeepScenario used only a small subset of realistic *weather*, *lighting*, *roadway surface*, and *roadway conditions*. To close the reality gaps in **RQ2.1**, *more realistic attributes should be used and their representation, as well as effects in simulation, should be evaluated.* The author of DeepScenario confirms such gaps, and it corroborates the findings by Song et al. [8] that systematic selection of parameters and faithful representation of them in simulation is required for critical scenario exploration. We need to validate and ensure at least attributes that are critical for collisions (as concerned in **RQ1.1**) are precisely simulated.

   As for **RQ2.2**, field testing needs to be compensated by simulation testing as they identify new critical scenarios that might be rare in real-world traffic. Correspondingly, field testing can provide useful guidance for improving the realism of simulation and synthetic scenarios as well. Therefore, we propose *developing a feedback control loop for real-world and simulation testing* and combining them in an effective way for testing AVs/ADS. While such a concept has been studied and reported [14, 73], how the two testing approaches should be divided or combined effectively in AV testing is not entirely evident yet.

Overall, this is the first step towards a data-centric evaluation of synthetic critical scenarios. We observe that a decisive outcome of our analysis is hampered by the fact that 1) *the two datasets used are arguably too small to be statistically significant*, 2) *some contextual information such as test arrangement and simulation quality are unavailable*, and 3) *prior distributions of some attributes such as roadway surface*

and conditions are unavailable. Those limitations have hindered us from performing a comprehensive evaluation of the two datasets and exposed some threats to our study.

Nevertheless, we maximise the findings of this study by 1) *selecting the best datasets available to us*, 2) *using all extractable attributes in our evaluation*, and 3) *deriving general recommendations and guidelines for evaluating and improving the realism of synthetic critical scenarios*. Given realism is an essential quality for test scenarios for AVs/ADS, and limited approaches, empirical evaluation, and insights have been reported so far, our study *sheds some light on this urgent topic* and *serves as a basis for future studies* by making four main contributions, as already described in Section 1:

1. A methodology for evaluating the realism of synthetic scenarios from realistic scenarios, using two metrics – *attribute distribution* and *Euclidean distance*. The metrics provide both macroscopic and microscopic views of the realism of a synthetic critical scenario set.
2. An empirical evaluation of how well a synthetic scenario set *DeepScenario* generates realistic AV collisions as recorded in *DMV California*, revealing findings and insights from empirical perspectives.
3. Observing existing *shortcomings* and possible *future improvements*, serving as *guidelines* for recording realistic scenarios, and generating and evaluating synthetic critical scenarios. The recommendations are general and not specific to the datasets used in this study.
4. We include *human assessment* in the loop to provide further insights and guidelines for evaluating the realism of synthetic critical scenarios on top of the empirical evaluation. The assessment strengthens the need to evaluate the realism of synthetic scenarios for testing AVs/ADS and the findings of our study.

5.2 Limitations

Although our methodology is defined to be generic and is meant to be applicable to various datasets, generalising our results particularly concerning our proof of concept is prone to some limitations (as also mentioned in Section 4 and 5.1).

In our proof of concept, one limitation concerns the different ADS and ODDs involved in the two datasets. As introduced earlier in Section 3, DMV California collects collision reports from several manufacturers, such as Waymo and Cruise, on different roads and weather in California. In contrast, DeepScenario uses Apollo ADS and four roads in San Francisco. Ideally, we want both (synthetic and realistic) datasets generated by the same system, in the same places, under the same weather and road conditions, and so on, to get a fair comparison between them. Otherwise, the discrepancies in collisions between the two datasets may simply be attributed to the different ADS and ODDs involved. That, however, is very challenging for the time being due to the unavailability of two perfectly matching datasets and thus; we acknowledge that as an inherent limitation for the current study. As a proof of concept, we use the two selected datasets (from DeepScenario and DMV California), which are the closest and publicly available datasets that we can find, to analyse how well DeepScenario (with the combination of Apollo ADS and SVL simulator) can produce similar critical scenarios as recorded in DMV California. This has been useful in revealing some insights and deriving general recommendations about realism from them.

Another limitation stems from the small size of the two datasets, further; raising general concerns about the usefulness of comparing their similarities. A small real-world dataset, e.g., DMV California data, owns the risk of not being statistically significant or not reflecting the ground truth distribution of real-world accidents. Further, the risks potentially lead to uncertainties and lack of confidence in findings we obtain from comparing DeepScenario from DMV California. For example, our distribution analysis reveals DeepScenario has a significantly different attribute distribution from DMV California, as described in Section 4, suggesting collisions take place to a different extent in DeepScenario from DMV California, under the same conditions (e.g., weather, road conditions). However, different results may emerge if a larger number of realistic AV collisions were collected in DMV California and expose a different distribution of attributes. Therefore, the evaluation analysis and some observations, as reported in Section 4, are restricted to the selected datasets only. Still, we derive general recommendations for improving the realism of synthetic critical scenarios.

Lastly, the limitation also lies in that only few attributes are extractable from the two selected datasets and their prior distribution in the real world is lacking. As discussed earlier in Section 4.3.2, there are several relevant attributes defined in DMV California data, such as movement proceeding collision and other associated factors, not used, due to inability to extract them from DeepScenario. When more attributes are included in the evaluation, we may obtain different results and analysis on the similarity between the selected datasets. Additionally, more contextual information such as test arrangements and prior distribution of the extractable attributes could enable a causal analysis and identify critical attributes for AV collisions. Those, unfortunately, are not available at the moment and thus; causal relations cannot be established.

5.3 Threats to Validity

As our goal is to devise a methodology and demonstrate a proof of concept for evaluating the realism of synthetic collision scenarios from realistic AV collisions, we strive to maximise the *construct* and *internal* validity throughout the study. Several limitations are discussed in Section 5.2, herein we focus on the validity of this study, especially regarding possible threats and their impacts as well as how we mitigate them.

- *Construct validity* refers to how well the constructs under study are measured [86, 87]. Our primary focus in this study is the realism of synthetic critical scenarios for testing AVs/ADS.

  One threat to construct validity is how *the concept of realism* is defined and measured for synthetic scenarios. As discussed in Section 2, no standard definitions or metrics are reported. To mitigate that, we *define realism as the degree of similarity* of synthetic scenarios to realistic scenarios, and *propose two evaluation metrics* to analyse realism from macroscopic and microscopic perspectives. Also, we *employ various analyses based on the metrics* to capture a better view of realism, including single- and multiple-attribute distribution analysis, and distance-based analysis via several criteria and K-means categorisation.

  Another threat comes from the *datasets and attributes* we select for evaluation, especially concerning how well they can be used to measure realism. To mitigate that, we *use the best datasets that are available* to us and *use all extractable*

attributes in our analysis. DeepScenario is a fairly large scenario set generated using various optimisation strategies, weather, road users, and user behaviours on a map of San Francisco. DeepScenario is open and provides a structured specification for each scenario. DMV California collects and releases all real AV collisions, in a standard template, for 10 years in California. Although they are not big currently, the two datasets represent synthetic and realistic critical scenario sets from the same place (although not entirely exactly the same) to some extent, and are live projects that are continuously maintained and updated.

- *Internal validity* refers to the validity of the results internal to the study [86, 87], e.g., how we have analysed the data and derived the findings and conclusions.

  One threat to internal validity is the *small size of the two datasets* for attribute distribution analysis. We acknowledge that as an inherent limitation as critical scenarios like collisions are rare, and contextual information such as how manufacturers have arranged their tests in different conditions is missing. As a preliminary step attempt to evaluate realism, we *explore different perspectives and maximise insights or current limitations we can learn* from them. Also, we manually *compensate data with missing field(s) and maximise the data* for evaluation. Despite the evaluation results being subject to the datasets used, the recommendations are somewhat general, and provide a basis and considerations for future studies.

  Another threat concerns about the *limited attributes* we use for evaluation. To mitigate that, we *analyse both datasets and use all extractable attributes* in our evaluation. As described in Section 3.2.1 and 4.3.2, some additional attributes are available from DMV California data, but are not used due to the limitations in DeepScenarios and SVL simulator. We expect new results may emerge when more attributes are used, and we do not claim the evaluation results are general. Nevertheless, this study reveals some empirical observations by evaluating the two datasets, and provides general recommendations for future studies.

5.4 Future Work

Given the current limitations and validity concerns for the proof of concept, we identify several tasks, ideas, and potential research directions for future work.

One future work is to update the datasets to incorporate recently included collision scenarios, and explore new datasets and attributes available for evaluation and analysis of realism. We should also synchronize with the DeepScenario team to work out additional attributes, or explore new datasets with more relevant attributes. Also, to develop guidelines on what attributes need to be collected for AVs/ADS test scenarios.

Another idea for future studies is to consider a different strategy to evaluate realism of synthetic scenarios. Instead of comparing everything in a driving scenario as a whole, we could separate the ADS behaviour, surrounding environment, and interactions between ADS and other road users into different parts. For example, we may only focus on evaluating how well a simulation environment represents regular road traffic and surrounding environment, i.e., human traffic without ADS, and evaluating whether simulation reflects ADS behaviour in the real world in another study. The formal one would only require real-world dataset for human traffic, which are already available at a large scale, such as SHRP2 NDS [88] and GIDAS [89] datasets.

The evaluation of realism can significantly impact the selection of test scenarios for ADS based on their realism. It would be important to study how realism evaluation can be used for test scenario selection for ADS, and further; improve the test efficiency and effectiveness. This suggests the testing should only focus on realistic and relevant scenarios for the ADS under test. Another research direction would be studying how realism impacts the scenario coverage. Although critical scenario identification is, in general, expected to identify critical scenarios and improve the overall test coverage effectively, the realism of such scenarios remains a fundamental issue to understand.

6 Conclusion

Critical scenarios are significant and have received considerable attention in research for testing AVs/ADS. Such scenarios are identified and analyzed to determine where, how, and why AVs fail, providing insights into their safety performance and helping to prevent similar incidents in the future. While extensive studies of critical scenario identification for testing AVs/ADS have been reported, the realism of resulting scenarios is rarely explored and their relevance to testing is unclear. In this study, we propose two metrics and evaluate the similarity of a synthetic collision set DeepScenario from realistic AV collisions collected by DMV California. We analyse the distribution of different attributes such as weather and lighting conditions, and observe a significant difference between the two datasets. We also perform a distance-based analysis and find that DeepScenario generates new collision scenarios that have not been recorded by DMV California, but they are still similar to the scenarios from DMV California from a distance perspective. Based on the evaluation results, we derive several recommendations for improving the realism of synthetic critical scenarios, concentrating on including more realistic parameter values (e.g., snowy weather) for critical scenario exploration, evaluating the quality of simulation, performing causal analysis on attributes for AV collisions, and developing a guideline of which attributes to collect for test scenarios. The study is limited by the data and attributes available, and we expect future studies to incorporate more attributes to evaluate the realism of synthetic critical scenarios.

Acknowledgements. This work was supported in part by the Wallenberg AI, Autonomous Systems and Software Program (WASP). Thanks to Chengjie Lu (the first author of DeepScenario) for their support and for sharing knowledge of DeepScenario and the SVL simulator. Mohammad Reza Mousavi has been partially supported by the UKRI Trustworthy Autonomous Systems Node in Verifiability, Grant Award Reference EP/V026801/2 and the EPSRC grant on Verified Simulation for Large Quantum Systems (VSL-Q), Grant Award Reference EP/Y005244/1. The authors were also partially supported by the SeedCorn granted provided by the EPSRC Model-Driven Engineering Network (MDE-Net).

Statements and Declarations

- Conflict of interest/Competing interests: There is no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

- Data availability: The data that support the findings of this study are openly available on Zenodo [51].

References

[1] Neurohr, C., Westhofen, L., Henning, T., Graaff, T., Möhlmann, E., Böde, E.: Fundamental considerations around scenario-based testing for automated driving. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 121–127 (2020). https://doi.org/10.1109/IV47402.2020.9304823

[2] Lou, G., Deng, Y., Zheng, X., Zhang, M., Zhang, T.: Testing of autonomous driving systems: where are we and where should we go? In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ESEC/FSE 2022, pp. 31–43. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3540250.3549111

[3] Riedmaier, S., Ponn, T., Ludwig, D., Schick, B., Diermeyer, F.: Survey on scenario-based safety assessment of automated vehicles. IEEE Access **8**, 87456–87477 (2020) https://doi.org/10.1109/ACCESS.2020.2993730

[4] Tang, S., Zhang, Z., Zhang, Y., Zhou, J., Guo, Y., Liu, S., Guo, S., Li, Y.-F., Ma, L., Xue, Y., *et al.*: A survey on automated driving system testing: Landscapes and trends. ACM Transactions on Software Engineering and Methodology (2023) https://doi.org/10.1145/3579642

[5] Road vehicles — safety of the intended functionality. Standard, International Organization for Standardization (2022). https://www.iso.org/obp/ui/#iso:std:77490:en

[6] Scenario-based verification and validation of self-driving vehicles: relevant safety metrics. White paper, Siemens Digital Industries Software (2022)

[7] Zhang, X., Tao, J., Tan, K., Törngren, M., Sánchez, J.M.G., Ramli, M.R., Tao, X., Gyllenhammar, M., Wotawa, F., Mohan, N., Nica, M., Felbinger, H.: Finding critical scenarios for automated driving systems: A systematic mapping study. IEEE Transactions on Software Engineering **49**(3), 991–1026 (2023) https://doi.org/10.1109/TSE.2022.3170122

[8] Song, Q., Engström, E., Runeson, P.: Industry practices for challenging autonomous driving systems with critical scenarios. ACM Trans. Softw. Eng. Methodol. **33**(4) (2024) https://doi.org/10.1145/3640334

[9] Hallerbach, S., Xia, Y., Eberle, U., Koester, F.: 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 (2018) https://doi.org/10.4271/2018-01-1066

[10] Commission implementing regulation (eu) 2022/1426. Regulation, European Parliament and the Council (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1426

[11] California Department of Motor Vehicles: Autonomous vehicles collision reports. Available online at: https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/ (last accessed: September 1 2023) (2023)

[12] Lu, C., Yue, T., Ali, S.: Deepscenario: An open driving scenario dataset for autonomous driving system testing. In: 2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR), pp. 52–56 (2023). https://doi.org/10.1109/MSR59073.2023.00020

[13] Lu, C., Yue, T., Ali, S.: DeepScenario: An Open Driving Scenario Dataset for Autonomous Driving System Testing. Zenodo (2023). https://doi.org/10.5281/zenodo.7714194

[14] Stocco, A., Pulfer, B., Tonella, P.: Mind the gap! a study on the transferability of virtual versus physical-world testing of autonomous driving systems. IEEE Transactions on Software Engineering **49**(4), 1928–1940 (2023) https://doi.org/10.1109/TSE.2022.3202311

[15] Riedmaier, S., Danquah, B., Schick, B., Diermeyer, F.: Unified framework and survey for model verification, validation and uncertainty quantification. Archives of Computational Methods in Engineering **28**, 2655–2688 (2021)

[16] Riedmaier, S., Schneider, J., Danquah, B., Schick, B., Diermeyer, F.: Non-deterministic model validation methodology for simulation-based safety assessment of automated vehicles. Simulation Modelling Practice and Theory **109**, 102274 (2021)

[17] Reisgys, F., Elgharbawy, M., Schwarzhaupt, A., Sax, E., Kemeny, A., Chardonnet, J., Colombet, F.: Argumentation on adas simulation validity using aleatory and epistemic uncertainty estimation. In: Proceedings of the Driving Simulation Conference 2021 Europe VR, pp. 25–32 (2021). Munich, Germany

[18] Neurohr, B., Graaff, T., Eggers, A., Bienmüller, T., Möhlmann, E.: Providing evidence for the validity of the virtual verification of automated driving systems. In: European Dependable Computing Conference, pp. 5–13 (2024). Springer

[19] Sargent, R.G.: Verification and validation of simulation models. In: Proceedings of the 2010 Winter Simulation Conference, pp. 166–183 (2010). IEEE

[20] Braun, T., Fuchs, J., Reisgys, F., Ries, L., Plaum, J., Schütt, B., Sax, E.: A review of scenario similarity measures for validation of highly automated driving. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems

(ITSC), pp. 689–696 (2023). IEEE

[21] Neelofar, N., Aleti, A.: Towards reliable ai: Adequacy metrics for ensuring the quality of system-level testing of autonomous vehicles. In: Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, pp. 1–12 (2024)

[22] Yan, X., Zou, Z., Feng, S., Zhu, H., Sun, H., Liu, H.X.: Learning naturalistic driving environment with statistical realism. Nature communications **14**(1), 2037 (2023)

[23] Sun, J., Zhang, H., Zhou, H., Yu, R., Tian, Y.: Scenario-based test automation for highly automated vehicles: A review and paving the way for systematic safety assurance. IEEE Transactions on Intelligent Transportation Systems **23**(9), 14088–14103 (2021) https://doi.org/10.1109/TITS.2021.3136353

[24] Abbas, H., O'Kelly, M.E., Rodionova, A., Mangharam, R.: A Driver's License Test for Driverless Vehicles. Mechanical Engineering **139**(12), 13–16 (2017) https://doi.org/10.1115/1.2017-Dec-9

[25] Tenbrock, A., König, A., Keutgens, T., Weber, H.: The conscend dataset: Concrete scenarios from the highd dataset according to alks regulation unece r157 in openx. In: 2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops), pp. 174–181 (2021). https://doi.org/10.1109/IVWorkshops54471.2021.9669219

[26] National Transportation Safety Board: Automated Vehicles - Investigations. Available online: https://www.ntsb.gov/Advocacy/safety-topics/Pages/automated-vehicles-investigations.aspx (last accessed: 1 September 2023) (2018)

[27] National Highway Traffic Safety Administration: Standing General Order on Crash Reporting - Data. Available online: https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting (last accessed: 1 September 2023) (2023)

[28] California Department of Motor Vehicles: Autonomous vehicles disengagement reports. Available online at: https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/disengagement-reports/ (last accessed: September 1 2023) (2022)

[29] Favarò, F.M., Nader, N., Eurich, S.O., Tripp, M., Varadaraju, N.: Examining accident reports involving autonomous vehicles in california. PLoS one **12**(9), 0184952 (2017) https://doi.org/10.1371/journal.pone.0184952

[30] Petrovic, D., Mijailovic, R., Pesic, D.: Traffic accidents with autonomous vehicles: type of collisions, manoeuvres and errors of conventional vehicles' drivers. Transportation research procedia **45**, 161–168 (2020) https://doi.org/10.1016/j.trpro.2020.03.003

[31] Pokorny, P., Høye, A.: Descriptive analysis of reports on autonomous vehicle collisions in california: January 2021–june 2022. Traffic Safety Research **2** (2022) https://doi.org/10.55329/xydm4000

[32] Sinha, A., Chand, S., Vu, V., Chen, H., Dixit, V.: Crash and disengagement data of autonomous vehicles on public roads in california. Scientific Data **8**(1), 298 (2021) https://doi.org/10.1038/s41597-021-01083-7

[33] Klischat, M., Althoff, M.: Generating critical test scenarios for automated vehicles with evolutionary algorithms. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 2352–2358 (2019). https://doi.org/10.1109/IVS.2019.8814230

[34] Song, Q., Engström, E., Runeson, P.: An empirically grounded path forward for scenario-based testing of autonomous driving systems. In: Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering, pp. 232–243 (2024)

[35] Gambi, A., Huynh, T., Fraser, G.: Generating effective test cases for self-driving cars from police reports. ESEC/FSE 2019, pp. 257–267. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3338906.3338942

[36] Gambi, A., Nguyen, V., Ahmed, J., Fraser, G.: Generating critical driving scenarios from accident sketches. In: 2022 IEEE International Conference On Artificial Intelligence Testing (AITest), pp. 95–102 (2022). https://doi.org/10.1109/AITest55621.2022.00022

[37] Gambi, A., Huynh, T., Fraser, G.: Automatically reconstructing car crashes from police reports for testing self-driving cars. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), pp. 290–291 (2019). https://doi.org/10.1109/ICSE-Companion.2019.00119

[38] Xinxin, Z., Fei, L., Xiangbin, W.: Csg: Critical scenario generation from real traffic accidents. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 1330–1336 (2020). https://doi.org/10.1109/IV47402.2020.9304609

[39] Abdessalem, R.B., Panichella, A., Nejati, S., Briand, L.C., Stifter, T.: Testing autonomous cars for feature interaction failures using many-objective search. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. ASE '18, pp. 143–154. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3238147.3238192

[40] Araujo, H., Hoenselaar, T., Mousavi, M.R., Vinel, A.: Connected automated driving: A model-based approach to the analysis of basic awareness services. In: 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1–7 (2020). https://doi.org/10.1109/PIMRC48278.

2020.9217142

[41] Nejati, S., Sorokin, L., Safin, D., Formica, F., Mahboob, M.M., Menghi, C.: Reflections on surrogate-assisted search-based testing: A taxonomy and two replication studies based on industrial ADAS and simulink models. Inf. Softw. Technol. **163**, 107286 (2023) https://doi.org/10.1016/j.infsof.2023.107286

[42] Li, G., Li, Y., Jha, S., Tsai, T., Sullivan, M., Hari, S.K.S., Kalbarczyk, Z., Iyer, R.: Av-fuzzer: Finding safety violations in autonomous driving systems. In: 2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE), pp. 25–36 (2020). https://doi.org/10.1109/ISSRE5003.2020.00012 . IEEE

[43] Luo, Y., Zhang, X.-Y., Arcaini, P., Jin, Z., Zhao, H., Ishikawa, F., Wu, R., Xie, T.: Targeting requirements violations of autonomous driving systems by dynamic evolutionary search. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 279–291 (2021). https://doi.org/10.1109/ASE51524.2021.9678883 . IEEE

[44] Calò, A., Arcaini, P., Ali, S., Hauer, F., Ishikawa, F.: Generating avoidable collision scenarios for testing autonomous driving systems. In: 2020 IEEE 13th International Conference on Software Testing, Validation and Verification (ICST), pp. 375–386 (2020). https://doi.org/10.1109/ICST46399.2020.00045 . IEEE

[45] Abdessalem, R.B., Nejati, S., Briand, L.C., Stifter, T.: Testing vision-based control systems using learnable evolutionary algorithms. In: Proceedings of the 40th International Conference on Software Engineering, pp. 1016–1026 (2018). https://doi.org/10.1145/3180155.3180160

[46] Birchler, C., Khatiri, S., Bosshard, B., Gambi, A., Panichella, S.: Machine learning-based test selection for simulation-based testing of self-driving cars software. Empir. Softw. Eng. **28**(3), 71 (2023) https://doi.org/10.1007/s10664-023-10286-y

[47] Birchler, C., Ganz, N., Khatiri, S., Gambi, A., Panichella, S.: Cost-effective simulation-based test selection in self-driving cars software. Sci. Comput. Program. **226**, 102926 (2023) https://doi.org/10.1016/j.scico.2023.102926

[48] Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK (2009). https://doi.org/10.1017/S0266466603004109

[49] Huang, R., Cui, C., Sun, W., Towey, D.: Poster: Is euclidean distance the best distance measurement for adaptive random testing? In: 2020 IEEE 13th International Conference on Software Testing, Validation and Verification (ICST), pp. 406–409 (2020). https://doi.org/10.1109/ICST46399.2020.00049

[50] PyPI: PyPDF2 3.0.1. Available online at: https://pypi.org/project/PyPDF2/ (last accessed: June 2 2023) (2022)

[51] Song, Q., Avner, B., Mohammad, R.M.: ADT-CSE. Zenodo (2024). https://doi.org/10.5281/zenodo.11179407 . https://doi.org/10.5281/zenodo.11179407

[52] Zhou, Y.: A review of text classification based on deep learning. In: Proceedings of the 2020 3rd International Conference on Geoinformatics and Data Analysis. ICGDA '20, pp. 132–136. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397056.3397082

[53] Lu, C., Shi, Y., Zhang, H., Zhang, M., Wang, T., Yue, T., Ali, S.: Learning configurations of operating environment of autonomous vehicles to maximize their collisions. IEEE Transactions on Software Engineering **49**(1), 384–402 (2023) https://doi.org/10.1109/TSE.2022.3150788

[54] The SVL Simulator team: SVL Simulator: An Autonomous Vehicle Simulator. Available online: https://github.com/lgsvl/simulator (last accessed: 21 August 2023) (2020)

[55] Yuqi Huai: SORA-SVL: Local Cloud built for SVL Simulator. Available online: https://github.com/YuqiHuai/SORA-SVL (last accessed: 21 August 2023) (2022)

[56] California collision manual, chapters 1-13, revised 2003. Regulation, United States Department of Transportation (2003). https://www.nhtsa.gov/document/california-collision-manual-chapters-1-13-revised-2003

[57] United States Department of Transportation: FARS EncyclopediaL Help - Terms. Available online: https://www-fars.nhtsa.dot.gov/help/terms.aspx (last accessed: 21 August 2023) (2023)

[58] Farjo, J., Abou Assi, R., Masri, W., Zaraket, F.: Does principal component analysis improve cluster-based analysis? In: 2013 IEEE Sixth International Conference on Software Testing, Verification and Validation Workshops, pp. 400–403 (2013). https://doi.org/10.1109/ICSTW.2013.52 . IEEE

[59] Brown, J.: Choosing the right number of components or factors in pca and efa. JALT Testing & Evaluation SIG Newsletter **13**(2) (2009)

[60] Bryant, F.B., Yarnold, P.R.: Principal-components Analysis and Exploratory and Confirmatory Factor Analysis., pp. 99–136. American Psychological Association, Washington, DC, US (1995)

[61] Li, S., Wang, W., Mo, Z., Zhao, D.: Cluster naturalistic driving encounters using deep unsupervised learning. In: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1354–1359 (2018). https://doi.org/10.1109/IVS.2018.8500529

[62] Hauer, F., Gerostathopoulos, I., Schmidt, T., Pretschner, A.: Clustering traffic

scenarios using mental models as little as possible. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 1007–1012 (2020). https://doi.org/10.1109/IV47402.2020.9304636

[63] Humaira, H., Rasyidah, R.: Determining the appropiate cluster number using elbow method for k-means algorithm. EAI, Padang, Indonesia (2018). https://doi.org/10.4108/eai.24-1-2018.2292388

[64] Liu, F., Deng, Y.: Determine the number of unknown targets in open world based on elbow method. IEEE Transactions on Fuzzy Systems **29**(5), 986–995 (2021) https://doi.org/10.1109/TFUZZ.2020.2966182

[65] Cui, M.: Introduction to the k-means clustering algorithm based on the elbow method. Accounting, Auditing and Finance **1**(1), 5–8 (2020) https://doi.org/10.23977/accaf.2020.010102

[66] WorldClimate.com: Average Weather Data for San Francisco, California. Available online: http://www.worldclimate.com/climate/us/california/san-francisco (last accessed: 21 August 2023) (2023)

[67] WorldClimate.com: Average Weather Data for San Diego, California. Available online: http://www.worldclimate.com/climate/us/california/san-diego (last accessed: 21 August 2023) (2023)

[68] spicy: scipy.stats.chisquare. Available online at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html (last accessed: June 2 2023) (2023)

[69] timeanddate: San Francisco, California, USA — Sunrise, Sunset, and Daylength, September 2023. Available online at: https:https://www.timeanddate.com/sun/usa/san-francisco (last accessed: June 2 2023) (2023)

[70] Song, Q., Tan, K., Runeson, P., Persson, S.: Critical scenario identification for realistic testing of autonomous driving systems. Software Quality Journal **31**(2), 441–469 (2023) https://doi.org/10.1007/s11219-022-09604-2

[71] Bagschik, G., Menzel, T., Maurer, M.: Ontology based scene creation for the development of automated vehicles. In: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1813–1820 (2018). https://doi.org/10.1109/IVS.2018.8500632

[72] Scholtes, M., Westhofen, L., Turner, L.R., Lotto, K., Schuldes, M., Weber, H., Wagener, N., Neurohr, C., Bollmann, M.H., Körtke, F., *et al.*: 6-layer model for a structured description and categorization of urban traffic and environment. IEEE Access **9**, 59131–59147 (2021) https://doi.org/10.1109/ACCESS.2021.3072739

[73] Stocco, A., Pulfer, B., Tonella, P.: Model vs system level testing of autonomous

driving systems: a replication and extension study. Empirical Software Engineering **28**(3), 73 (2023) https://doi.org/10.1007/s10664-023-10306-x

[74] Beringhoff, F., Greenyer, J., Roesener, C., Tichy, M.: Thirty-one challenges in testing automated vehicles: Interviews with experts from industry and research. In: 2022 IEEE Intelligent Vehicles Symposium (IV), pp. 360–366 (2022). https://doi.org/10.1109/IV51971.2022.9827097 . IEEE

[75] Bärgman, J., Svärd, M., Lundell, S., Hartelius, E.: Methodological challenges of scenario generation validation: a rear-end crash-causation model for virtual safety assessment. Transportation Research Part F: Traffic Psychology and Behaviour **104**, 374–410 (2024)

[76] Wimmer, P., Düring, M., Chajmowicz, H., Granum, F., King, J., Kolk, H., Camp, O., Scognamiglio, P., Wagner, M.: Toward harmonizing prospective effectiveness assessment for road safety: Comparing tools in standard test case simulations. Traffic injury prevention **20**(sup1), 139–145 (2019)

[77] Knauss, A., Schröder, J., Berger, C., Eriksson, H.: Paving the roadway for safety of automated vehicles: An empirical study on testing challenges. In: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 1873–1880 (2017). https://doi.org/10.1109/IVS.2017.7995978 . IEEE

[78] Daza, I.G., Izquierdo, R., Martínez, L.M., Benderius, O., Llorca, D.F.: Sim-to-real transfer and reality gap modeling in model predictive control for autonomous driving. Applied Intelligence, 1–17 (2022) https://doi.org/10.1007/s10489-022-04148-1

[79] Zhong, Z., Tang, Y., Zhou, Y., Neves, V.d.O., Liu, Y., Ray, B.: A survey on scenario-based testing for automated driving systems in high-fidelity simulation. arXiv preprint arXiv:2112.00964 (2021)

[80] (grva) - new assessment/test method for automated driving (natm) guidelines for validating automated driving system (ads). Standard, United Nations Economic Commission for Europe (UNECE) (2023). https://unece.org/transport/documents/2023/04/working-documents/grva-new-assessmenttest-method-automated-driving-natm

[81] Erdogan, A., Kaplan, E., Leitner, A., Nager, M.: Parametrized end-to-end scenario generation architecture for autonomous vehicles. In: 6th International Conference on Control Engineering & Information Technology (CEIT), pp. 1–6 (2018). https://doi.org/10.1109/CEIT.2018.8751872 . IEEE

[82] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on Robot Learning, pp. 1–16 (2017). PMLR

[83] Ji, P., Li, R., Xue, Y., Dong, Q., Xiao, L., Xue, R.: Perspective, survey and trends: Public driving datasets and toolsets for autonomous driving virtual test. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), pp. 264–269 (2021). https://doi.org/10.1109/ITSC48978.2021.9564428 . IEEE

[84] Kang, Y., Yin, H., Berger, C.: Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments. IEEE Transactions on Intelligent Vehicles **4**(2), 171–185 (2019) https://doi.org/10.1109/TIV.2018.2886678

[85] Rosique, F., Navarro, P.J., Fernández, C., Padilla, A.: A systematic review of perception system and simulators for autonomous vehicles research. Sensors **19**(3), 648 (2019) https://doi.org/10.3390/s19030648

[86] Verdecchia, R., Engström, E., Lago, P., Runeson, P., Song, Q.: Threats to validity in software engineering research: A critical reflection. Information and Software Technology **164**, 107329 (2023) https://doi.org/10.1016/j.infsof.2023.107329

[87] Lago, P., Runeson, P., Song, Q., Verdecchia, R.: Threats to validity in software engineering – hypocritical paper section or essential analysis? In: Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. ESEM '24, pp. 314–324. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3674805.3686691

[88] Papazikou, E., Quddus, M., Thomas, P., Kidd, D.: What came before the crash? an investigation through shrp2 nds data. Safety Science **119**, 150–161 (2019)

[89] Otte, D., Jänsch, M., Haasper, C.: Injury protection and accident causation parameters for vulnerable road users based on german in-depth accident study gidas. Accident Analysis & Prevention **44**(1), 149–153 (2012)