Do Not Forget the Obvious - RISC: A Risk-Informed Slice-Coverage Protocol for Safe Autonomous Driving

Fabian Hüger

CARIAD SE, 38440 Wolfsburg, Germany
fabian.hueger@cariad.technology

Abstract. Aggregate metrics may not fully reflect performance in insufficiently examined high-risk driving conditions. We propose *RISC* (Risk-Informed Slice Coverage), a practical protocol for *risk-guided stress testing* and *coverage-qualified evaluation*. Here, risk-guided stress testing means that a finite audit budget is not sampled uniformly, but directed toward risk-relevant sub-datasets, called *risk slices*; coverage-qualified evaluation means that results are reported together with explicit statements about which slices are sufficiently or insufficiently covered. The protocol translates safety concerns into machine-readable risk slices, uses lightweight signals to tag candidate data, selects a compact audit set by risk, and qualifies the reported results by coverage evidence in each critical slice. An LLM can optionally support this process by surfacing relevant but potentially overlooked conditions for consideration during test planning, thereby helping engineers to *not forget the obvious*. The protocol is model-agnostic and can be applied to perception modules, driving models, or other autonomous-driving subsystems whose safety claims depend on adequate exposure to high-risk conditions. We instantiate the protocol for monocular pedestrian perception in urban driving using 1,000 frames from the Zenseact Open Dataset (ZOD), image statistics, and a YOLO-based detector proxy. In this proof-of-concept study, risk-guided selection substantially increases critical failure discovery from 34.0% under random sampling to 98.5%. We position the method as a lightweight, assurance-oriented evaluation layer that complements scenario categorization, scenario coverage assessment, and broader testing-and-verification workflows for safe autonomous driving.

Keywords: autonomous driving · stress testing · safety evaluation · coverage qualification · pedestrian detection · data selection

1 Motivation

Safety-critical autonomous driving systems are evaluated using finite datasets that may not fully represent all relevant operating conditions within their defined operational design domain (ODD), i.e., the operating conditions under which the system is intended to function [1]. Under these constraints, aggregate metrics

may provide an incomplete picture of performance: a subsystem may perform well on average, while conditions associated with potential data-coverage gaps can remain challenging to evaluate. For urban driving, such conditions may include night scenes, partial occlusion by vehicles, small or distant pedestrians, dense crowds, and image degradations such as glare or motion blur. During development and assurance, these conditions should therefore be considered explicitly when evaluating system performance and the available supporting evidence. Their relevance to safety should be assessed in the context of the overall system architecture and its associated safety mechanisms.

Recent progress in large language models creates a practical opportunity for making such evaluation spaces more systematic. LLMs cannot replace safety-critical engineering judgment, but they can help elicit "obvious" risk conditions that are easy to overlook in manual test planning, structure them into reviewable slice candidates, and translate them into machine-readable evaluation specifications [2]. In this paper, we use this capability as an assurance-support mechanism: the LLM acts as a front-end for surfacing candidate stressors, while all slices, weights, proxy signals, and reported claims remain subject to engineering review and coverage qualification.

We therefore argue for a simple principle:

> *Do not forget the obvious, and do not interpret results without stating how well these risks are covered.*

Contributions. This paper contributes *RISC* (Risk-Informed Slice Coverage), a practical protocol for risk-guided stress testing and coverage-qualified evaluation of autonomous-driving subsystems. The protocol comprises: (1) a prompt-to-slice step that derives machine-readable stress-test slices from driving safety concerns and applies them to an existing evaluation pool; (2) a risk-weighted selection rule for constructing compact safety-relevant audit subsets; (3) a coverage-qualified reporting step that keeps reported results but explicitly states where evidence is insufficient; and (4) an empirical pedestrian-perception instantiation showing that the protocol increases critical failure discovery from 34.0% under random sampling to 98.5% under Risk Top-K, thereby exposing safety-relevant hazards for safe autonomous driving.

2 State of the Art

Safety assurance in autonomous driving already recognizes that aggregate metrics alone are insufficient. Scenario-based safety-assurance frameworks and recent reviews of scenario metrics emphasize that safety evidence should be structured around scenario coverage, criticality, exposure, completeness, and related forms of contextual evidence rather than being reduced to a single average number. Complementing this perspective, recent testing-and-verification reviews for connected and autonomous vehicles underline that scalable evaluation requires systematic scenario generation, validation approaches that scale to finite engineering budgets,

and explicit reasoning about which situations are actually covered by the available evidence.[3,4,5,6]

A second relevant line of work studies evaluation beyond raw perception accuracy. Planner-centric or downstream-sensitive evaluation asks whether perception errors actually matter for subsequent driving behavior. This perspective is closely aligned with safe driving: a missed pedestrian in or near the drivable area, especially under occlusion or at long distance, can require different downstream behavior than an error in an otherwise empty, low-risk scene. Complementary work on uncertainty, calibration, and out-of-distribution behavior in autonomous-driving perception further stresses that reliable deployment depends on knowing where confidence is unwarranted and where data coverage is weak.[7,8,9,10]

A third line of work is particularly close to the present paper's proof-of-concept instantiation. Prior work on *safety assurance of machine-learning perception functions* argues that convincing assurance cases require both quantitative performance evidence and structured treatment of ML-specific insufficiencies. Likewise, recent work on *performance limiting factors* for pedestrian detection studies which concrete factors—including distance, occlusion, crowdedness, and visibility effects—systematically contribute to detector misbehavior. These findings motivate our choice to operationalize risk-critical conditions as explicit slices, i.e., named subsets of the data pool, and to bias audit budgets toward the subsets most likely to surface safety-relevant misses.[11,12]

Closely related are automated error-slice discovery approaches such as HiBug2 [13], which use task-specific visual attributes and efficient slice enumeration to identify coherent failure subsets for model debugging and repair. Our focus is complementary: rather than discovering new slices or repairing the model, we study how pre-specified, safety-motivated slice proxies can be used to prioritize samples and report audit evidence under a finite evaluation budget.

The gap addressed in this paper is therefore not a new detector or driving model, but an evaluation protocol that contributes to answering three practical questions: how to translate safety concerns into operational slices, how to allocate a finite audit budget toward these slices, and how to report results without overclaiming when the evidence base is thin. We position the proposed protocol as a model-agnostic layer for safety-oriented evaluation, applicable to perception modules today and, in principle, to larger driving-model pipelines as well.

3 Protocol

Figure 1 gives a schematic overview of the proposed testing workflow. In short, the protocol translates safety concerns into interpretable risk slices, operationalizes them through lightweight tagging rules, ranks candidate samples by risk, and reports the resulting evaluation together with explicit coverage qualifications.

Let $\mathcal{D} = \{x_1, \ldots, x_N\}$ be a pool of frames, clips, or episodes from an operational design domain $\mathcal{O}$. A risk slice $S_i \subseteq \mathcal{D}$ is an interpretable subset of this pool, for example night scenes, small pedestrians, or pedestrians close to vehicles. We define K such slices, $\mathcal{S} = \{S_1, \ldots, S_K\}$, with binary indicators $z_i(x) \in \{0, 1\}$

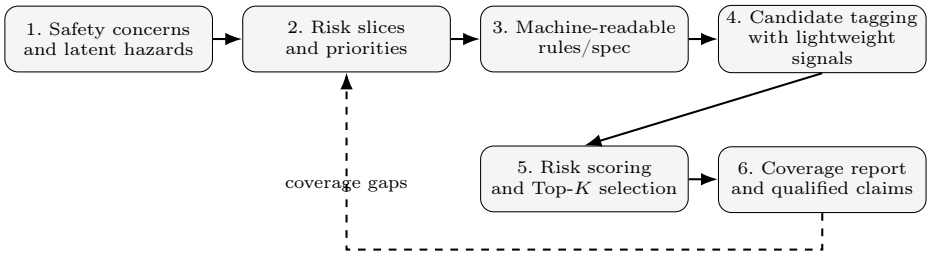

Fig. 1. Schematic overview of the proposed testing workflow. (1) Safety concerns and latent hazards are collected, (2) translated into prioritized risk slices, (3) specified as machine-readable rules, (4) used for candidate tagging with lightweight signals, (5) ranked by risk scoring and Top-K selection, and (6) reported together with coverage qualifications and explicit coverage gaps.

for $i = 1, \ldots, K$. Each slice receives a weight

$$w_i \propto \text{Severity}_i \times \text{Exposure}_i \times \text{Detectability}_i^{-1} \times \text{Overreliance}_i. \tag{1}$$

Here, severity describes the potential consequence of a missed hazard, exposure describes how plausible the condition is in the intended ODD, detectability describes how difficult the condition is expected to be for the evaluated function, and overreliance captures the risk that good average performance leads to excessive trust in the function. Samples are ranked by

$$r(x) = \sum_{i=1}^{K} w_i z_i(x) + \lambda u_{\text{proxy}}(x), \tag{2}$$

where a higher $r(x)$ means that sample x has higher priority for inclusion in the audit set. The optional term $u_{\text{proxy}}(x)$ can encode lightweight evidence of uncertainty, image degradation, detector confidence, or disagreement between multiple candidate taggers or model variants. The top-B samples form a stress-test or audit set.

3.1 Risk-slice elicitation from safety concerns

The protocol begins with a structured safety prompt that asks which hazards should matter for evaluation. The purpose is not to generate a universal ontology, but to create an inspectable initial taxonomy of risk-critical slices that can later be reviewed and operationalized by engineers. A standard large language model is useful in this step because it already encodes many common, obvious failure modes from public autonomous-driving and perception knowledge; these suggestions are treated as candidates, not as authoritative safety requirements. For example, a prompt may ask for pedestrian-perception hazards in urban driving and yield slices such as *night/low light*, *vehicle occlusion*, and *small or distant pedestrians*, together with severity rationale, plausible exposure, expected detectability, overreliance risk, and candidate tagging heuristics.

3.2 Machine-readable specification

The elicited slices are translated into a compact, structured specification, for example in JSON- or YAML-like form. This machine-readable layer is essential because it decouples the safety reasoning from the implementation and enables reusability across curation, sampling, monitoring, and audit pipelines.

3.3 Candidate tagging with lightweight signals

Each slice is approximated with lightweight signals, for example timestamps, image statistics, detector outputs, or simple geometry heuristics. For instance, low-light candidates can be tagged by low mean brightness and a high dark-pixel ratio, while a vehicle-occlusion proxy can be tagged by small distances or overlap between detected person and vehicle boxes. The goal is not perfect semantic labeling, but a low-cost candidate-generation stage that can be audited and improved later.

3.4 Risk-weighted selection

Once candidate tags are available, each sample receives a risk score. The score combines the selected slice weights with optional proxy terms such as image degradation, uncertainty, or disagreement between taggers or model variants. The original evaluation pool is not used only as-is because finite audit budgets can be dominated by easy or frequent cases; Top-K selection instead concentrates manual inspection and detailed analysis on samples expected to be rich in safety-critical failures. Random and heuristic baselines are retained as controls.

3.5 Coverage diagnostics and qualified reporting

The final step is not to suppress results, but to qualify them. If a slice remains undercovered, the evaluation still reports the measured performance, but it additionally states that the evidence base for that slice is weak. This distinction is important: the protocol is intended to prevent overinterpretation, not to block all reporting.

4 Pedestrian-Perception Instantiation for Safe Driving

We instantiate the protocol for monocular pedestrian perception in urban driving. Image statistics are used to tag low light, blur, and glare, while predictions from a frozen detector proxy are used to derive candidate tags for small/distant pedestrians, crowds, and vehicle-occlusion proxies. These detector-derived tags are used only for candidate generation and risk-based sampling; ZOD [14] annotations are used for the subsequent performance evaluation. This separation is important because proxy tags are not ground truth and may share failure modes with the evaluated detector. In this framing, pedestrian-perception failures are interpreted as failures to surface hazards that should trigger more defensive downstream driving behavior.

Table 1. Compact slice specification and implementation signals used in the proof-of-concept study.

ID	Slice	Operational signal used in prototype
S1	Night / low light	mean brightness and dark-pixel ratio
S2	Vehicle occlusion proxy	person–vehicle proximity or box overlap
S4	Small/distant pedestrians	person-box height or image-area ratio
S5	Motion blur	Laplacian-variance sharpness below threshold
S6	Glare/backlight	bright-pixel ratio and high contrast
S7	Dense crowd	number of detected persons

4.1 Why pedestrian perception matters for safe driving

Reliable pedestrian detection is relevant to the development of safe automated driving systems. The perception task can be particularly challenging in situations involving partial occlusion, small or distant pedestrians, or dense interaction zones. Examples include a pedestrian emerging from behind a vehicle, a distant pedestrian at a crossing, or multiple pedestrians in a crowded scene. These conditions should therefore be considered explicitly when evaluating pedestrian-perception functions and the available supporting evidence.

4.2 Prompting procedure and proof-of-concept scope

The slice taxonomy in this paper was produced as a proof-of-concept prompting exercise rather than a universal standard. In a first stage, a structured safety prompt was given to GPT-5.5 Thinking to elicit candidate slices, rationales, and operational heuristics. Details of the prompt and prompting setup can be found in Appendix A. In a second stage, the resulting list was manually reviewed and reduced to a compact subset that could be implemented with the available image statistics, detector outputs, and audit budget. The purpose was to demonstrate a practical workflow from safety concern to operational slice, not to claim that the resulting prompt or slice set is universally optimal.

In Table 1, S3 is intentionally absent because the crosswalk-interaction slice suggested during prompting could not be implemented reliably with the available lightweight signals in this proof-of-concept. The detector-derived slices should therefore be read as proxy candidates for stress testing, not as complete semantic labels.

4.3 Reasoning behind the weights

Table 2 reports the additive slice weights used in the prototype. The weights in this paper are not learned parameters and are not normalized probabilities; they are scoring weights for ranking samples. They are a compact operationalization of a safety argument. We assign higher weight to slices that combine: high potential severity if missed, plausible exposure in urban traffic, low detectability by the perception stack, and elevated risk of human or system overreliance.

Table 2. Risk weights used for additive risk scoring in the detector-assisted proof-of-concept.

Slice	Meaning	Weight
S1	Night / low light	0.19
S2	Vehicle occlusion proxy	0.23
S4	Small/distant pedestrians	0.15
S5	Motion blur	0.05
S6	Glare/backlight	0.11
S7	Dense pedestrian crowd	0.08

Concretely, the highest weight is assigned to **S2 vehicle occlusion proxy** because hidden pedestrians can emerge suddenly from behind parked or moving vehicles and are both frequent and safety-critical in urban scenes. **S1 night / low light** receives a high weight because low visibility degrades detection while also encouraging misplaced confidence if aggregate daytime performance is overgeneralized. **S4 small/distant pedestrians** remains highly relevant because safe action often depends on early recognition of distant, low-pixel-footprint pedestrians. **S6 glare/backlight** and **S5 blur** are weighted lower in this proof-of-concept not because they are unimportant, but because they appeared less frequently in the available 1,000-frame subset. **S7 crowd** captures interaction density and ambiguity, which can increase both miss probability and downstream risk.

5 Experiment: Risk-Guided Selection

5.1 Setup

We process 1,000 monocular front-camera frames and compare three sampling strategies with budget $K = 200$: uniform random sampling, a simple human heuristic baseline, and risk-guided Top-K selection. The human heuristic baseline is a simple non-learned rule-based selection intended to represent manually chosen obvious stressors, rather than a trained risk model. The detector configuration uses YOLOv8n, the nano variant of YOLOv8, with detector input size 640 and confidence threshold 0.25; the original images are resized for detector inference and are not assumed to be square. Blur is measured by the variance of the image Laplacian, a common sharpness proxy: lower values indicate stronger blur. We flag the lowest 10% of frames by this sharpness score, which yields an effective threshold of 14.73 in this dataset. Detection performance is evaluated against pedestrian annotations from the Zenseact Open Dataset (ZOD). Predicted and annotated pedestrian boxes are matched using an IoU threshold of ≥ 0.5; objects marked as unclear are excluded.

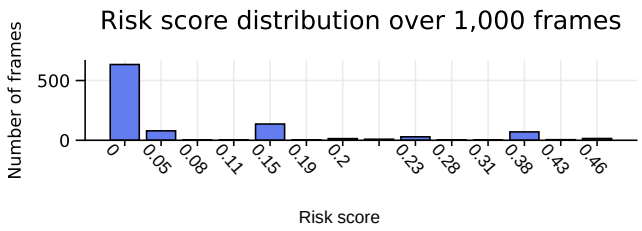

Fig. 2. Risk score distribution over 1,000 processed frames. Most frames are not part of any detected risk slice, while high-risk multi-slice frames form a small tail.

Table 3. Slice coverage comparison. The Manifest column reports prevalence in the full 1,000-frame candidate pool; Random, Human heuristic, and Risk Top-K report prevalence within the selected $K = 200$ audit sets.

Slice	Manifest	Random	Human heuristic	Risk Top-K
S1 night	0.2%	0.0%	1.0%	1.0%
S2 occlusion proxy	12.2%	15.0%	12.0%	**61.0%**
S4 small pedestrian	24.9%	23.5%	23.0%	**83.0%**
S5 blur	10.0%	9.5%	11.5%	10.5%
S6 glare	0.3%	0.5%	1.5%	0.0%
S7 crowd	2.6%	2.0%	3.5%	**12.5%**

5.2 Risk and coverage distribution

As shown in Figure 2, the frame pool is strongly imbalanced: 634 of 1,000 frames receive risk score 0, while only 20 frames reach scores of 0.43 or higher. As can be seen in Table 3 and Figure 3, Risk Top-K enriches detector-derived stress slices: S2 occlusion proxy rises from 15.0% under random sampling to 61.0%, S4 small pedestrians from 23.5% to 83.0%, and S7 crowd from 2.0% to 12.5%. Night and glare do not increase substantially because these conditions are rare in the available 1,000-frame pool and can only be enriched if enough candidate frames are present.

5.3 Performance metrics

Coverage alone does not measure perception quality. We therefore evaluate the same frozen detector on all three selected test sets using pedestrian annotations from ZOD and report the standard set-level counts and rates—precision, recall, and miss rate—together with two audit-oriented metrics that are directly tied to the proposed protocol.

Standard metrics. Precision, recall, and miss rate are computed in the standard way from TP, FP, and FN after IoU-based matching. Miss rate is emphasized because the practical concern is how many pedestrians remain undetected in the selected audit sets.

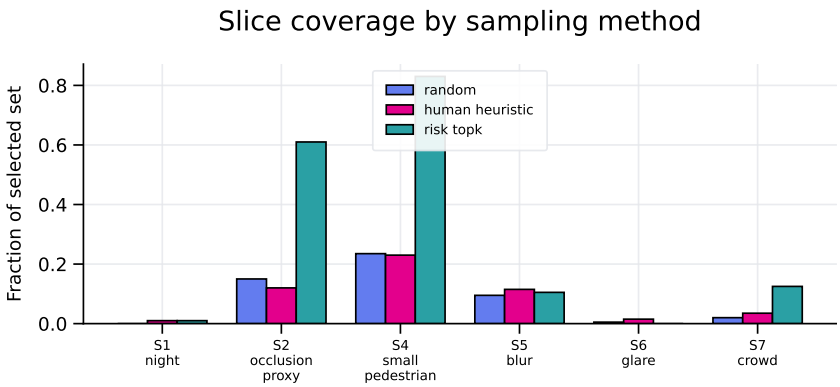

Fig. 3. Slice coverage by sampling method. Risk-guided selection concentrates the audit budget on occlusion, small-pedestrian, and crowd slices.

Critical Failure Discovery Rate (CFDR@K). Let $\mathcal{A}_K$ denote the selected audit set and let $c(x) = 1$ if frame x contains at least one false negative and belongs to at least one non-zero-risk slice, and $c(x) = 0$ otherwise. We compute

$$\text{CFDR@}K = \frac{1}{K} \sum_{x \in \mathcal{A}_K} c(x).$$

CFDR@K therefore measures the fraction of selected audit frames that actually surface missed pedestrians in risk-relevant conditions. This metric is intentionally conditioned on the risk-slice concept and should not be interpreted as an unbiased general-performance metric. Instead, it answers the budget-oriented audit question: *how often does the selected audit set surface failures in risk-relevant conditions?*

Risk-Weighted Failure Discovery (RWFD@K). CFDR@K treats all critical-failure frames equally. RWFD@K further weights each critical-failure frame by its risk score $r(x)$ and normalizes by the audit budget:

$$\text{RWFD@}K = \frac{1}{K} \sum_{x \in \mathcal{A}_K} r(x)c(x).$$

In this experiment, risk scores range from 0 to 0.46; examples include 0.15 for a small-pedestrian-only frame, 0.23 for an occlusion-proxy-only frame, and 0.46 for frames combining small pedestrians, crowds, and occlusion proxies. RWFD@K therefore captures not only whether failures are found, but whether they are found in higher-priority regions of the risk space.

Table 4 reports the resulting set-level performance and audit-yield metrics.

Interpreting performance across sets. The evaluation confirms that the risk-guided set is substantially harder and more failure-dense than the baselines.

Table 4. Detector performance and failure-discovery metrics on the three selected test sets. GT denotes pedestrian instances in ZOD; matching uses IoU ≥ 0.5.

Test set	GT	TP	FP	FN	Prec.	Rec.	Miss	CFDR@K	RWFD@K
Random ($K = 200$)	2029	92	69	1937	57.1%	4.5%	95.5%	34.0%	7.56%
Human heuristic ($K = 200$)	1892	100	58	1792	63.3%	5.3%	94.7%	35.0%	7.21%
Risk Top-K ($K = 200$)	2560	326	207	2234	61.2%	12.7%	87.3%	98.5%	27.81%

Table 5. Slice-level detector performance within the Risk Top-K set. We report slices with sufficient selected frames and available ground-truth pedestrians for meaningful interpretation; S1 night and S6 glare are omitted here because they are too rare or absent in the selected subset. The hardest selected slices contain many pedestrian instances and high miss rates, which is precisely the intended stress-test behavior.

Slice	Frames	GT	TP	FP	FN	Prec.	Rec.	Miss
S2_occlusion_proxy	122	1660	207	146	1453	58.6%	12.5%	87.5%
S4_small_pedestrian	166	2210	298	178	1912	62.6%	13.5%	86.5%
S5_blur	21	185	23	17	162	57.5%	12.4%	87.6%
S7_crowd	25	574	147	68	427	68.4%	25.6%	74.4%

Compared with random sampling, Risk Top-K increases recall from 4.5% to 12.7% because the selected set contains more pedestrian instances and more frames in which the detector produced at least one pedestrian prediction; this should not be interpreted as improved detector robustness. The miss rate nevertheless remains high at 87.3%, indicating that the selected frames expose many missed pedestrians. Most importantly for stress testing, CFDR@K rises from 34.0% for random sampling and 35.0% for the heuristic baseline to 98.5% for Risk Top-K; the risk-weighted failure discovery rate increases from 7.56% to 27.81%. Thus, the main result is not that the detector becomes better on the risk-selected set. Rather, risk-guided selection makes the fixed audit budget more informative by selecting frames in which missed pedestrians occur more often in safety-relevant conditions.

Table 5 further qualifies the Risk Top-K result by breaking down detector performance inside the selected risk slices. The occlusion-proxy and small-pedestrian slices contain many pedestrian instances and retain high miss rates, indicating that the selected audit set indeed concentrates on difficult safety-relevant conditions. The crowd slice shows higher recall but still exposes substantial remaining misses. These slice-level results should not be interpreted as independent robustness estimates, because frames may belong to multiple slices and rare slices such as night or glare remain undercovered. Instead, the table identifies which covered risk slices drive the audit yield and where additional evidence would be needed.

6 Coverage-Qualified Reporting

For each slice S_i, we report its prevalence in the candidate pool and in the selected audit sets. The reporting rule is:

Coverage qualification. Report the measured result, but explicitly state when a risk-critical slice remains undercovered and therefore provides only limited evidence.

The experiment illustrates why this qualification is necessary. Risk-guided Top-K improves S2, S4, and S7 coverage substantially, but it does not automatically cover all relevant slices. In particular, glare (S6) and night (S1) remain rare in the 1,000-frame subset. Therefore, the reported results are informative, but they should not be read as comprehensive evidence for driving safety outside the covered conditions.

7 Discussion and Limitations

RISC is an assurance-support protocol, not a replacement for scenario engineering, verification, or safety-case development. Its main contribution is to separate three concerns that are often mixed in aggregate evaluation: *what to test* through risk slices, *what to prioritize* through risk-weighted selection, and *what can be claimed* through coverage-qualified reporting.

The pedestrian-perception experiment is only an exemplary instantiation of the protocol, not a comprehensive benchmark of detector robustness. It illustrates how safety concerns can be translated into operational slice proxies, how a finite audit budget can be directed toward higher-risk samples, and how the resulting findings can be reported without overclaiming.

The proof-of-concept remains limited by approximate detector-derived slice tags, partial coupling between proxy tagging and detector evaluation, and an intentionally unbalanced Risk Top-K audit set. Consequently, the reported performance numbers should be interpreted as audit-yield diagnostics rather than unbiased robustness estimates. Rare slices such as night and glare also remain undercovered in the available 1,000-frame pool.

These limitations reinforce the purpose of coverage-qualified reporting: RISC does not turn limited evidence into broad safety claims, but makes the boundary of the available evidence explicit. Future work should combine global Risk Top-K selection with slice-balanced quotas, random controls, temporal driving scenes, and downstream planning metrics.

8 Conclusion

We presented *RISC*, a compact risk-informed slice-coverage protocol for risk-guided stress testing and coverage-qualified evaluation in safe autonomous driving. In an empirical pedestrian-perception experiment, risk-guided selection enriched occlusion, small-pedestrian, and crowd-related stress slices compared with random and heuristic baselines, while increasing critical failure discovery from 34.0% to 98.5%. The resulting evaluation combines coverage, perception performance, failure discovery, and explicit coverage qualifications, enabling more transparent reporting of evaluation evidence and its limitations for autonomous driving systems.

Acknowledgement

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project "Safe AI Engineering – Sicherheitsargumentation befähigendes AI Engineering über den gesamten Lebenszyklus einer KI-Funktion". The author would like to thank the consortium for the successful cooperation.

References

1. SAE International: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE J3016 (2018).
2. Hüger, F.: Can Agentic AI Transform Safety Engineering for the Development of Complex Systems? SAFECOMP 2025 Position Paper (2025).
3. de Gelder, E., Singh, T., Hadj-Selem, F., Vidal Bazan, S., Op den Camp, O.: Scenario Metrics for the Safety Assurance Framework of Automated Vehicles: A Review of Its Application. Vehicles 7(3), 100 (2025).
4. Westhofen, L., Neurohr, C., Koopmann, T., Butz, M., Schütt, B., Utesch, F., Kramer, B., Gutenkunst, C., Böde, E.: Criticality Metrics for Automated Driving: A Review and Suitability Analysis of the State of the Art. Archives of Computational Methods in Engineering 30(1), 1–35 (2022).
5. Alemayehu, H., Sargolzaei, A.: Testing and Verification of Connected and Autonomous Vehicles: A Review. Electronics 14(3), 600 (2025).
6. ISO: Road vehicles – Test scenarios for automated driving systems – Scenario categorization. ISO 34504:2024 (2024).
7. Philion, J., Kar, A., Fidler, S.: Learning to Evaluate Perception Models Using Planner-Centric Metrics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020).
8. Araújo, B., Teixeira, J.F., Fonseca, J., Cerqueira, R., Beco, S.C.: The Road to Safety: A Review of Uncertainty and Applications to Autonomous Driving Perception. Entropy 26(8), 634 (2024).
9. Sicking, J., Akila, M., Schneider, J.D., Hüger, F., Schlicht, P., Wirtz, T., Wrobel, S.: Tailored Uncertainty Estimation for Deep Learning Systems. arXiv:2204.13963 (2022).
10. Pintz, M., Sicking, J., Poretschkin, M., Akila, M.: A Survey on Uncertainty Toolkits for Deep Learning. In: ICLR Workshop on Setting up ML Evaluation Standards to Accelerate Progress (2022).
11. Burton, S., Hellert, C., Hüger, F., Mock, M., Rohatschek, A.: Safety Assurance of Machine Learning for Perception Functions. In: Deep Neural Networks and Data for Automated Driving, pp. 335–358. Springer (2022).
12. Bayzidi, Y., Smajic, A., Schneider, J.D., Hüger, F., Moritz, R., Knoll, A.C.: Performance Limiting Factors of Deep Neural Networks for Pedestrian Detection. In: Proceedings of the British Machine Vision Conference (BMVC) (2022).
13. Chen, M., Zhao, C., Xu, Q.: HiBug2: Efficient and Interpretable Error Slice Discovery for Comprehensive Model Debugging. In: International Conference on Learning Representations (ICLR) (2025).
14. Alibeigi, M., et al.: Zenseact Open Dataset: A Large-Scale and Diverse Multimodal Dataset for Autonomous Driving. In: ICCV (2023).

Appendix

A Prompting Setup and Scope

The prompt used in this paper is a *proof-of-concept prompt*, not a generally valid template. Its purpose was to test whether a modern foundation model can help bootstrap an inspectable slice taxonomy for autonomous-driving safety evaluation.

The prompting workflow had two stages. In Stage 1, **GPT-5.5 Thinking** was asked to generate risk-relevant slices, rationales, priorities, and candidate heuristics for pedestrian perception in urban driving. The emphasis was on operationalizability rather than completeness: each slice had to be approximable from timestamps, image statistics, detector outputs, or simple heuristics. In Stage 2, the resulting slice set was manually reviewed, pruned, and aligned with the signals available in the actual prototype pipeline. The final subset retained in the paper should therefore be read as a compact demonstration set for this study, not as a domain-independent ontology.

B Stage-1 Prompt for Risk-Slice Elicitation

The following prompt was used to elicit the initial risk-critical slice taxonomy. It is included to make the slice construction process inspectable and reproducible.

> You are a Safety and ML Assurance Engineer for a camera-based pedestrian detection system deployed in urban driving environments.
> **Goal.** Identify, structure, and prioritize risk-relevant data slices, i.e., operational design domain subspaces, for pedestrian detection, such that they can later be used for: (a) dataset curation and focused sampling, (b) safety monitoring and triggering, (c) audit and assurance documentation, and (d) automated tagging and processing in downstream pipelines.
> **Context and assumptions.**
> – Task: pedestrian detection, either bounding boxes or presence detection, using monocular RGB camera input.
> – Operational domain: urban Germany, EU context, mixed-density city traffic, speeds between 0–60 km/h, frequent pedestrian interactions.
> – Environmental conditions: day/night cycles, weather variations, and typical European urban infrastructure such as crosswalks, sidewalks, and intersections.
> – System context: human-in-the-loop operation possible; risk of operator over-reliance on model outputs should be explicitly considered.
> – Constraints: outputs must be structured to enable simple tagging heuristics without requiring large-scale labeling or complex models.
> – Important: all slices must be defined in a way that allows them to be approximated using simple signals such as metadata, timestamps, image statistics, detector outputs, or lightweight heuristics.
> **Task.**
> 1. Generate a list of 15–25 risk-relevant data slices for pedestrian detection.

2. For each slice, provide a compact operational specification with: slice ID and name; short definition; risk rationale; risk factors including severity, exposure, detectability, and overreliance trigger; tagging and measurement heuristics; typical failure modes; suggested evaluation criterion; priority; and assumptions.
3. Additionally provide: (A) a Top-8 shortlist of critical slices, (B) a compact subset suitable for a proof-of-concept study, (C) ten high-risk combinatorial slices, and (D) a simple weighting scheme for slice importance using

$$w_i \propto \text{Severity}_i \times \text{Exposure}_i \times \frac{1}{\text{Detectability}_i} \times \text{Overreliance}_i.$$

4. After the human-readable section, generate a machine-readable representation of all slices in valid JSON with fields for ID, name, definition, numeric risk factors, priority, heuristics, failure modes, acceptance criterion, tags, and combinable slices. Optionally also provide YAML suitable for pipeline configuration.

Constraints and style. Avoid vague categories; every slice must be operationalizable. Do not assume proprietary signals; use only camera input and common metadata. Focus on real-world urban driving conditions in Germany. Emphasize safety-critical pedestrian scenarios. Use precise engineering language and ensure consistency between human-readable and machine-readable outputs.