Rare Conditions and Coverage in Driving-Video Evaluation

Jack Warren (DK), Jacob King (AL), Evan Holmes (AF)

Abstract

An aggregate driving-video score can look stable while offering little evidence about rare conditions. This article proposes a coverage-first evaluation design that records condition support, generation success, and assessability before interpreting performance. OmniDrive motivates the multi-view generation setting. No rare scenes are generated or scored. The design distinguishes deliberately enriched challenge sets from naturally sampled data and keeps absent evidence visible. It explains how a limited success can remain useful without implying reliability in every uncommon combination of weather, geometry, visibility, and traffic behavior.

Rarity depends on a population and an observation rule

A condition can be rare in one benchmark and common in another environment. It can also appear rare because an annotation process excludes difficult observations. An evaluation should therefore state the population and measurement rule behind the label. Calling a scene long-tail does not identify its frequency, severity, or scientific relevance. The present article develops a design in which observed support and intended use define the rare-condition claim before the generator's performance is examined.

OmniDrive proposes coordinated multi-view driving-video generation (Meng et al., 2026d). Its broader world-model framing motivates questions about coverage beyond common scenes. The current article does not test the architecture on rare cases or infer its failure frequency from another dataset. It proposes an evaluation protocol. The central issue is how to distinguish absence from poor performance: a condition with no assessable outputs has not demonstrated reliability, but neither has every possible error within that unobserved condition been directly established.

Constructing a coverage record before a score

A proposed benchmark would define relevant conditions and their intersections using a documented source. Some attributes may come from known scene specifications, others from annotations or measurements with uncertainty. These origins should remain distinguishable. The record would count requested scenes, generated attempts, completed outputs, and assessable observations for each condition. A favorable score on the assessable subset would then be presented alongside its coverage rather than described as though every requested rare case contributed valid evidence.

Natural sampling and deliberate enrichment answer different questions. A naturally sampled collection can describe observed frequencies under its source process, while an enriched challenge set can provide more information about selected difficult conditions. Its unweighted average does not estimate their frequency in ordinary use. Both designs can be useful when their purpose is explicit. If results are reweighted toward a target population, the weights need an appropriate source and cannot create missing combinations that the study never observed.

Intersectional conditions require particular attention to support. A benchmark can contain rain, night scenes, and pedestrians separately while having almost no observations combining them. Reporting only marginal counts can make coverage appear broader than it is. The proposed protocol would retain relevant joint cells and identify sparse or absent combinations. It would not fill them through an assumption that separate successes automatically combine into reliable performance under the joint condition. That generalization is a further empirical question.

Assessing performance where observations are scarce

Rare-condition outcomes should retain uncertainty appropriate to their independent scene count. Many frames from one unusual scene do not replace several independently acquired or generated scene contexts. Repeated generation draws can describe stochastic variation for that scene, but broader transfer requires other evidence. The report should state which source of variation its interval represents. A numerical score with many decimal places should not conceal that its apparent rare-condition evidence rests on a small number of underlying environments.

The assessor's coverage is another possible bottleneck. Small objects, heavy occlusion, or unusual rendering can defeat a scoring tool even when the video contains useful content. The evaluation should distinguish failed generation, unassessable output, and observed scene inconsistency. Independent review can help when adequate evidence exists, while unresolved cases should remain unresolved. Automatically assigning every unavailable score to the most favorable or unfavorable category would introduce a decision rule whose consequences need to be made explicit.

The final claim should identify the conditions, combinations, sampling process, and effort actually tested. A limited positive finding can be valuable if it establishes performance on a defined challenge. It does not establish operational safety, natural error frequency, or every form of out-of-distribution reliability. A later downstream study can investigate specified practical consequences on independent real data. The proposed design keeps that future evidence separate while making the present coverage sufficiently clear to guide which rare conditions deserve the next collection or experiment.

Counting both rare conditions and assessor coverage

A comparison of naturally sampled common conditions and explicitly enriched joint challenge cases can clarify joint-condition coverage, generation completion and conditional failure. A benchmark includes night, rain and cyclists individually, with very little support at their intersection. This hypothetical example would be recorded at the level of independently sampled scene request. A low failure rate on assessable outputs can exclude precisely the scenes where recognition breaks. Keep attempt and measurement denominators by joint condition and preserve a separate naturally weighted estimate where supported.

Construct the challenge inventory from combinations rather than one marginal label at a time. A scene with a cyclist at night in rain belongs to a different support cell from separate clips showing each condition alone. Record requests, completed outputs and assessable outputs in every cell. Increasing the prevalence of rare combinations can be appropriate for stress testing. A naturally weighted summary would need weights from a defined observation population. Present those quantities separately so that deliberate enrichment does not become an implicit prevalence estimate.

An initial table can record the eligibility and observation status of independently sampled scene request, before any performance statistic is calculated. For the independently sampled scene request, distinguish sensor ineligibility from failures that occur despite sufficient visibility.

Store the camera identifiers, timestamps and spatial relationships associated with independently sampled scene request. To evaluate joint-condition coverage, generation completion and conditional failure, keep paired observations together during resampling, since independent resampling would lose their shared source of variation. For explicitly enriched joint challenge cases, examine whether the finding depends on the allowed timestamp tolerance or geometric matching distance.

For joint-condition coverage, generation completion and conditional failure, define one setting for the primary estimate and show the neighboring settings as a separate comparison. Comparisons involving naturally sampled common conditions can report whether disagreement starts with projection, detection or assignment across views. Document the sampling boundary for joint-condition coverage, generation completion and conditional failure and reserve a new cohort of cases for a distinct replication.

Samples and evidence

Documenting construction and intended use clarifies which exclusions occurred during collection, which arose from a model failure and which were imposed during analysis. Time linkage supports paired cross-view checks once sensor visibility and timestamp differences have been accounted for in the denominator. One study asks dataset creators to record collection, intended use and limitations (Gebru et al., 2021). During validation, another study offers time-linked multimodal driving observations (Caesar et al., 2019).

Spatial organization of urban scenes permits splits by drive or neighborhood so that closely adjacent frames do not dominate nominal transfer results. Calibration and timing metadata make it possible to distinguish disagreement caused by projection assumptions from disagreement in the observed road objects. For a prospective comparison, one published study provides camera-based driving benchmarks and related annotations (Geiger et al., 2012). Earlier research provides spatially organized urban scene observations and annotations (Liao et al., 2021).

A methodological precedent provides data for driving perception and forecasting tasks (Wilson et al., 2023). During validation, multiple annotation types allow an analyst to distinguish performance on visible road objects from performance on lane or scene context. A methodological precedent provides annotated driving video for several perception tasks (Yu et al., 2018). Motion and perception records allow separate assessment of what was observable and what could be forecast over the subsequent trajectory.

One published study provides road-scene data for perception evaluation at scale (Sun et al., 2019). In an evaluation, closed-loop assessment can reveal compounding consequences of small initial errors; open-loop frame evaluation answers a different question about recorded scenes. Recording camera and scene conditions permits a stratified comparison instead of assuming that the overall perception score applies equally to rare encounters. A methodological precedent develops a closed-loop benchmark for vehicle planning (Caesar et al., 2021).

Measurement and error

Confidence calibrated on the original collection may change under a new capture or generation regime, so calibration belongs in transfer testing. A detector baseline makes localization auditable by separating confidence and box agreement from errors in a subsequent scene-level or tracking judgment. During validation, earlier research examines uncertainty estimates under dataset shift (Ovadia et al., 2019). Related work formulates object detection as direct image-to-box prediction (Redmon et al., 2016).

Low prevalence can make a seemingly small false-positive rate operationally costly, so report precision at plausible recall levels and retain class counts. During validation, another study compares choices of forecast error scale across series (Hyndman and Koehler, 2006). One approach contrasts precision-recall and receiver operating characteristic summaries under class imbalance (Saito and Rehmsmeier, 2015). Scaling choices determine how units of different magnitude contribute to an aggregate error; a transparent report retains the original-scale results where interpretable.

Its separate video dimensions provide a reason to publish a failure profile instead of combining every observation into a single quality index. Report coverage beside error rates. Rejecting difficult cases can improve conditional accuracy while leaving fewer cases with an answer. A methodological precedent organizes video-generation evaluation into distinct measurement dimensions (Huang et al., 2023). During validation, one study reports prediction performance jointly with abstention coverage (Geifman and El-Yaniv, 2017).

A relevant contribution assesses probability calibration and studies temperature scaling (Guo et al., 2017). One published study associates objects using both high- and lower-confidence detections (Zhang et al., 2021). After fitting a decision threshold, reliability curves on an independent split can show whether reported confidence still matches observed success frequency. Retaining weaker detections offers a concrete test of whether missed associations are caused by the confidence filter or by the subsequent motion assignment.

Systems and comparators

An existing method links three-dimensional object detections across time using a tracking baseline (Weng et al., 2020). Simultaneous spatial and temporal consistency motivates keeping a separate record of cross-camera disagreements and of discontinuities within each camera sequence. Its matching step is a sensible control for whether an apparent identity switch arises before or after three-dimensional detections enter the tracker. Related work targets panoramic and temporally coherent driving scene generation (Wen et al., 2023).

A downstream planning evaluation can test whether perception differences affect a decision, with the object-level failures retained for diagnosis. A methodological precedent uses structured conditions from driving scenes to guide video generation (Wang et al., 2023a). During validation, structured inputs permit a comparison of controlled motion while keeping the initial scene and available layout observations explicit in the experimental record. A methodological precedent integrates perception and motion tasks for driving planning (Hu et al., 2022a).

Use camera calibration to assess cross-view consistency. Two plausible images can disagree about the same object's location. For a prospective comparison, the distribution that generated an archive constrains follow-up evaluation, particularly where difficult states or rare recovery actions are sparsely represented. Related work addresses consistency of multi-view driving video synthesis (Li et al., 2023b). Another study reviews the problem of learning policies from a fixed set of experience (Levine et al., 2020).

A simple motion-and-detection tracker offers a transparent control for assessing which failures appear before more complex visual features are introduced. Center-based detections provide a geometric reference for interpreting whether a matching error arises from the object center or its estimated extent. One study represents detected objects as centers for detection and tracking (Yin et al., 2020). In an evaluation, a relevant contribution combines detection association and motion estimation in an online tracker (Bewley et al., 2016).

Analysis and interpretation

Earlier research provides guidance on inference under dependence within specified groups (Cameron and Miller, 2015). In a controlled test, an existing method shows bias from reusing validation for selection and error estimation (Varma and Simon, 2006). Repetitions from a common generating condition can have correlated errors, making the number of independent groups more informative than the nominal row count. Reusing a validation split for repeated configuration selection can make the final error estimate optimistic; retain a separate assessment for the chosen rule.

Linking evidence to the evaluation

A methodological precedent represents static scene elements and moving objects for street-scene reconstruction (Yan et al., 2024). One approach provides a configurable driving simulation environment (Dosovitskiy et al., 2017). In an evaluation, related work formulates classification decisions with explicit costs for different errors (Elkan, 2001). A simulator can vary visibility, traffic and camera placement independently while leaving the underlying events observable for verification against the rendered frames. An application with asymmetric errors calls for costs to be stated separately from the metric so that operating thresholds can be recalculated. Separating static and dynamic content makes it possible to attribute an alignment error to road geometry, object tracking or a combination of both.

Conclusion

Rare-condition evaluation should begin with support and assessability, then interpret scores within that boundary. The proposed protocol reports no new generator performance. It distinguishes natural frequencies, enriched challenges, joint-condition gaps, and independent scene replication. A future result can then substantiate a specific uncommon-condition capability without letting an attractive aggregate or a large correlated frame count imply evidence for conditions that were scarcely observed or never assessed.

References

Yu, Fisher, et al.. "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning" *arXiv preprint arXiv:1805.04687*, 2018. https://arxiv.org/abs/1805.04687.

Sun, Pei, et al.. "Scalability in Perception for Autonomous Driving: Waymo Open Dataset" *arXiv preprint arXiv:1912.04838*, 2019. https://arxiv.org/abs/1912.04838.

Wilson, Benjamin, et al. "Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting." *arXiv preprint arXiv:2301.00493*, 2023. https://arxiv.org/abs/2301.00493.

Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

Wen, Yuqing, et al. "Panacea: Panoramic and Controllable Video Generation for Autonomous Driving." *arXiv preprint arXiv:2311.16813*, 2023. https://arxiv.org/abs/2311.16813.

Meng, Zijie, et al. "OmniDrive: An LLM-Choreographed Multi-Agent World Model with Unified Latent Co-Compression for Multi-View Driving Video Generation." *arXiv preprint arXiv:2606.17536* (2026).

Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

Caesar, Holger, et al. "NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles." *arXiv preprint arXiv:2106.11810*, 2021. https://arxiv.org/abs/2106.11810.

Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

Caesar, Holger, et al.. "nuScenes: A multimodal dataset for autonomous driving" *arXiv preprint arXiv:1903.11027*, 2019. https://arxiv.org/abs/1903.11027.

Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

Huang, Ziqi, et al.. "VBench: Comprehensive Benchmark Suite for Video Generative Models" *arXiv preprint arXiv:2311.17982*, 2023. https://arxiv.org/abs/2311.17982.

Zhang, Yifu, et al.. "ByteTrack: Multi-Object Tracking by Associating Every Detection Box" *arXiv preprint arXiv:2110.06864*, 2021. https://arxiv.org/abs/2110.06864.

Weng, Xinshuo, et al.. "AB3DMOT: A Baseline for 3D Multi-Object Tracking and New Evaluation Metrics" *arXiv preprint arXiv:2008.08063*, 2020. https://arxiv.org/abs/2008.08063.

Wang, Xiaofeng, et al. "DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving." *arXiv preprint arXiv:2309.09777*, 2023. https://arxiv.org/abs/2309.09777.

Hu, Yihan, et al. "Planning-oriented Autonomous Driving." *arXiv preprint arXiv:2212.10156*, 2022. https://arxiv.org/abs/2212.10156.

Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

Levine, Sergey, Aviral Kumar, George Tucker, and Justin Fu. “Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.” *arXiv*, 2020, arXiv:2005.01643. https://arxiv.org/abs/2005.01643.

Bewley, Alex, et al. "Simple Online and Realtime Tracking." *arXiv preprint arXiv:1602.00763*, 2016. https://arxiv.org/abs/1602.00763.

Gebru, Timnit, et al.. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

Liao, Yiyi, et al. "KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D." *arXiv preprint arXiv:2109.13410*, 2021. https://arxiv.org/abs/2109.13410.

Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Yan, Yunzhi, et al. "Street Gaussians: Modeling Dynamic Urban Scenes with Gaussian Splatting." *arXiv preprint arXiv:2401.01339*, 2024. https://arxiv.org/abs/2401.01339.

Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite." *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354-3361. https://doi.org/10.1109/CVPR.2012.6248074.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

Li, Xiaofan, et al. "DrivingDiffusion: Layout-Guided multi-view driving scene video generation with latent diffusion model." *arXiv preprint arXiv:2310.07771*, 2023. https://arxiv.org/abs/2310.07771.

Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

Yin, Tianwei, et al. "Center-based 3D Object Detection and Tracking." *arXiv preprint arXiv:2006.11275*, 2020. https://arxiv.org/abs/2006.11275.

Dosovitskiy, Alexey, et al. "CARLA: An Open Urban Driving Simulator." *arXiv preprint arXiv:1711.03938*, 2017. https://arxiv.org/abs/1711.03938.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You Only Look Once: Unified, Real-Time Object Detection.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788. https://doi.org/10.1109/cvpr.2016.91.