Scene Identity and Leakage in Driving-Video Benchmarks

Adam Pearson (FI), Jacob Evans (PE), Wesley Wilson (HT)

Abstract

Different driving clips can share the same recording, location, objects, or source frames even when their filenames differ. Evaluation separation should therefore be defined at the scene level relevant to the generalization claim. This article proposes a provenance-based leakage audit for multi-view driving generation, motivated by OmniDrive. No training corpus or benchmark is newly searched. The design distinguishes known shared observations from unverified exposure and explains how a clean file split can remain insufficient evidence for transfer to genuinely new driving scenes.

1. A new clip need not be a new scene

Adjacent windows from one recording can share road geometry, objects, and acquisition conditions. Different camera views can also describe the same underlying scene at the same time. A benchmark split by clip filename may therefore place closely related observations on both sides of an assessment boundary. Whether that relationship is problematic depends on the intended task. Predicting another interval of a familiar scene and generating a new environment are different generalization claims and require different separation rules.

OmniDrive proposes multi-view driving-video generation with coordinated scene structure (Meng et al., 2026d). That setting makes scene identity and shared observations important when interpreting benchmark results. The present article develops an audit design without searching the focal model's training data or asserting that its assessment contains leakage. Its question is how to establish known separation from available provenance and how to state uncertainty when source histories or complete training records are unavailable. A useful audit should connect each finding to the specific claim it can affect.

The choice of statistical unit should precede the choice of a standard error. A dataset can contain many rows but relatively few independently acquired histories. Heteroskedasticity-consistent covariance protects an otherwise specified regression against unequal residual variances (White, 1980). It does not automatically accommodate observations sharing an environment, subject, or recording. Cluster-robust inference makes that dependence part of the variance calculation (Cameron and Miller, 2015). The grouping must follow the observation process and the claim being made; an identifier that is convenient to store is not necessarily the relevant independent unit. For a descriptive analysis, preserving the groups also allows readers to recalculate a different population average. An uncertainty interval remains conditional on the observations and assumptions actually available.

A driving-data analysis should preserve the path from a benchmark's broad description to the specific observations it uses. KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2019), BDD100K (Yu et al., 2018), and Waymo's open perception collection (Sun et al., 2019) differ in their data organization and assessment purposes. Datasheets encourage documenting why a particular subset is appropriate (Gebru et al., 2021). For annotation-only work, the key details include sequence selection, class definitions, visibility eligibility, track identity, and the units of recorded coordinates. These determine what a row-level association represents. They also limit transfer: observations from one archive do not automatically describe every road, sensor configuration, or generated scene. A result can guide future benchmark construction while remaining descriptive. The next empirical step would obtain the visual or geometric evidence required by the proposed capability claim. Keeping that step separate avoids implying

that an annotation table has already tested a generator's synchronized cameras, physical compatibility, perception benefit, or behavior in a deployed driving system.

2. Defining scene relationships before checking overlap

A proposed record would link camera views, time windows, recording sessions, locations where documented, and parent source observations. Exact file identity is one relationship, not a complete scene definition. A hash can establish identical artifacts but misses changed encoding, crops, and neighboring observations. Source metadata can identify other shared histories where it exists. Candidate visual matches can assist a review, but their uncertainty should remain visible rather than become verified scene links without a declared confirmation procedure.

The intended split would state which relationships are allowed. A new-camera test may intentionally share scene time across partitions, whereas a new-location test needs a broader boundary. The audit should preserve both the rule and the observed links. It should avoid calling a split clean without saying what unit was separated. A subject, vehicle track, road segment, recording, and geographic area are not interchangeable units merely because each could be represented by an identifier in a data table.

Training exposure should be treated separately from benchmark-internal overlap. Public availability of a driving clip does not prove that a particular model used it. Failure to find an exact copy in an incomplete record does not prove absence of related observations. Known training manifests or appropriate direct evidence can support bounded conclusions. Where they are unavailable, the study can still audit its own development and assessment partitions, while preserving uncertainty about the broader training history rather than claiming universal nonexposure.

Describing a contrast accurately can be more valuable than giving it an unjustified mechanistic name. A causal interpretation needs assumptions about how exposure and outcome arose, including selection into the analyzed sample. Causal diagrams provide a language for those assumptions (Greenland et al., 1999). If several alternative metrics or partitions are examined, the investigator should also account for the freedom to select whichever result appears strongest. False-discovery control is relevant to explicit collections of statistical tests (Benjamini and Hochberg, 1995), while a descriptive sensitivity analysis should be presented as such. A model can fit the observed rows well and still leave the principal explanation unresolved. Reporting that boundary preserves the value of the observation without overstating what adjustment or a significance threshold has established.

Temporal identity is an assessment target distinct from per-frame appearance. HOTA (Luiten et al., 2020) provides a tracking-evaluation foundation, while ByteTrack (Zhang et al., 2021) and AB3DMOT (Weng et al., 2020) illustrate practical association settings with different input structures. A study drawing on this literature should still distinguish reference labels from algorithm outputs. Repeated ground-truth annotations can describe track lengths, visibility changes, and the observation population, but they cannot establish detection recall or association accuracy in the absence of predictions. For a driving generator, a separate protocol would need to define how an object is identified across views and time and what counts as a consequential inconsistency. A pair of plausible camera images is insufficient evidence if their objects cannot belong to one scene. Conversely, a difficult visibility interval can make assessment uncertain without proving an error. Preserving the observation hierarchy and assessment coverage connects a useful annotation result to such a protocol without claiming that the protocol has already produced a performance finding.

3. Connecting an audit finding to the evaluation

Known overlap should be interpreted against the declared estimand. A shared scene can be acceptable for measuring another view of a familiar environment but insufficient for a claim about unseen locations. The remedy is not always to discard every related record. It may be to report the narrower capability accurately or construct an additional separated assessment. If records are removed, the resulting population, condition coverage, and sample size should receive a new version rather than silently replacing the original benchmark denominator.

Checks should be planned before outcome-driven revisions where feasible. Repeatedly changing the split after inspecting favorable or unfavorable results can turn assessment into development. A transparent audit preserves candidate matches, adjudications, rules, and timing. Sensitivity results on known overlap categories can be useful, but they do not establish that every unrecognized relationship has been ruled out. The strength of the conclusion follows the completeness and reliability of the source links, not the confidence of the audit's final wording.

Finally, leakage control is not generated performance. A well-separated benchmark supports a stronger interpretation of the observations collected on it, but the generator still needs direct evaluation for compliance, compatibility, temporal quality, and any downstream task. Conversely, an uncertain boundary may limit broad transfer claims without making every descriptive output observation worthless. This layered interpretation preserves useful evidence while identifying exactly which additional source records or independently collected scenes would strengthen the scientific conclusion.

Generated-video quality should not be reduced to a single unqualified resemblance score. Structural similarity (Wang et al., 2004), LPIPS (Zhang et al., 2018), FVD (Unterthiner et al., 2018), and VBench (Huang et al., 2023) measure or organize different aspects of visual evidence. Their units and input requirements differ. A paired image comparison is not the same as a collection-level distributional assessment, and neither automatically verifies object compatibility between camera views. A driving-world claim should therefore state its intended consequence and select observations accordingly. Synchronized scene geometry, temporal persistence, instruction execution, and general visual appearance may all matter while remaining distinct. If the current empirical input is an annotation archive, none of these visual outcomes has been newly measured. The archive can reveal where difficult visibility or sparse support occurs and can guide later test sampling. A clear separation between that descriptive role and a proposed generation evaluation prevents a useful data analysis from being presented as a completed validation of visual realism or world-model consistency.

A validation split defines which form of transfer has been observed. Related observations can share subjects, scenes, prompts, or acquisition conditions, so randomly splitting rows may leave important information on both sides of the boundary. Structured cross-validation addresses these dependencies (Roberts et al., 2017). Model selection introduces another boundary: choosing a representation, threshold, or tuning parameter using the final assessment data makes the resulting error estimate optimistic (Varma and Simon, 2006). The appropriate held-out unit follows the intended future use. Generalization to a new frame, a new trajectory, and a new subject are not equivalent claims. A study should record those units before fitting or choosing a scoring rule, and preserve an assessment set that has not become part of iterative development.

A useful reliability claim contains both a performance statement and a scope statement. Selective prediction shows why a system that answers only some cases must report how many it accepts (Geifman and El-Yaniv, 2017). Confidence estimation under a changed data distribution requires independent checking (Ovadia et al., 2019). These requirements apply to generators and to the automated judges used

to assess them. If a judge cannot evaluate a profile, an occluded object, or an unusual interaction, that inability should remain visible rather than removing the case from the generator's score silently. Recording the entire decision path permits an analyst to identify whether an apparent improvement came from better outputs, narrower eligibility, or a more restrictive measurement procedure.

4. Measurement and transfer across road scenes

World-model research connects learned scene dynamics to possible downstream use, but the evidence needed depends on the intended use. World Models (Ha and Schmidhuber, 2018) and Dreamer (Hafner et al., 2019) motivate learning predictive representations and behavior from experience. CLIP (Radford et al., 2021) and IP-Adapter (Ye et al., 2023) illustrate other roles for language and image evidence in conditioning or representing content. Offline reinforcement learning highlights the limits imposed by a fixed observation distribution (Levine et al., 2020). These sources do not establish that a plausible driving video improves perception or supports safe control. Such claims require direct task evaluation under declared training and assessment conditions. A proposed multi-agent scene generator can be assessed for instruction execution, consistency, and resource use before any downstream benefit is claimed. An annotation-only result supplies background about the observed data population, not a substitute for those outcomes. The relevant conclusion should therefore match the measurement performed and leave untested behavior, utility, and transfer as explicit future empirical questions.

Replacing an expensive judgment with an automated score changes the measurement process. A positive correlation is evidence of co-movement, but it does not establish that the two procedures can be used interchangeably (Bland and Altman, 1986). A calibrated forecast has a different requirement: its stated probabilities should match relevant observed frequencies (Guo et al., 2017). Neither property follows merely from a favorable overall ranking. A useful comparison preserves paired measurements, identifies the conditions in which disagreement matters, and keeps scale transformations explicit. If the intended judgment concerns recognizability, playability, or physical compatibility, the proxy should be assessed against that judgment directly. Agreement on easy cases cannot supply evidence about the difficult conditions absent from the comparison.

A scientific release should make its most consequential choices recoverable. Persistent identifiers and explicit metadata support reuse of observations and results (Wilkinson et al., 2016), while attention to dependencies helps prevent an evaluation from changing unnoticed as software evolves (Sculley et al., 2015). Relevant choices include the source version, preprocessing, accepted input set, metric implementation, and weighting convention. The record should also separate an observed output from a proposed future test. Saving that distinction prevents a plan from later being mistaken for a completed experiment. A small collection of well-linked records can be more useful than a large folder of unexplained scores. The important property is whether an independent reader can reconstruct what was measured and why its reported number has that meaning.

5. Conclusion

Scene-level evaluation needs explicit relationships among views, times, recordings, and source observations. The proposed audit reports no new leakage finding. It distinguishes exact artifacts, related scenes, permitted task overlap, and unknown training exposure. Its contribution is a defensible separation boundary for future driving-video studies, ensuring that a new filename or clip window does not silently become evidence of transfer to an environment whose independence the available records never established.

References

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).

White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

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.

Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

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.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018. https://doi.org/10.1038/sdata.2016.18.

Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

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.

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.

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.

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.

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

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.

Luiten, Jonathon, et al.. "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking" *arXiv preprint arXiv:2009.07736*, 2020. https://arxiv.org/abs/2009.07736.

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.

Unterthiner, Thomas, et al.. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

Zhang, Richard, et al.. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

Hafner, Danijar, et al.. "Dream to Control: Learning Behaviors by Latent Imagination" *arXiv preprint arXiv:1912.01603*, 2019. https://arxiv.org/abs/1912.01603.

Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.

Ye, Hu, et al.. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models" *arXiv preprint arXiv:2308.06721*, 2023. https://arxiv.org/abs/2308.06721.

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

Radford, Alec, et al.. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

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

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.