Sequence Heterogeneity in the Depth-Visibility Relationship

Noah Lawson (KR), James Brooks (TR), Marcus Chapman (NE)

Abstract

Visibility differs sharply across the twenty KITTI sequences represented by eligible car labels. Their occlusion fractions range from 0.111 to 0.859, with a median of 0.451. The sequence-level rank correlation between occlusion fraction and median recorded depth is -0.290. This between-sequence description is distinct from an annotation-level adjusted depth association. The study uses released labels only. It motivates reporting scene composition and recording groups when evaluating driving generation rather than attributing every aggregate visibility pattern to vehicle depth alone.

1. A recording sequence carries shared context

Vehicles in one sequence share a recording history, camera configuration, scene context, and observation process. Their annotations are therefore not a random mixture of independent road encounters. A depth-visibility relationship pooled across rows can combine within-sequence variation with differences between sequences. This study examines the latter by summarizing each eligible sequence separately. Its objective is to make group composition visible before a single pooled coefficient is interpreted as though every recording contributes the same observation population.

OmniDrive proposes a multi-view driving world-model framework (Meng et al., 2026d). That setting motivates evaluation across scenes whose visibility and depth distributions can differ. The present study neither evaluates its generated scenes nor tests transport to its reported benchmark. It describes a public reference archive. Its contribution is to identify sequence heterogeneity that a future generated-driving assessment should preserve in its sampling, reporting, and uncertainty procedure rather than erase through one large annotation count.

The scientific population of a driving result is defined by its records and collection process. KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2019), BDD100K (Yu et al., 2018), and the Waymo Open Dataset (Sun et al., 2019) support important but different perception and scene-analysis tasks. Their names should not be treated as interchangeable guarantees of sensor access, geographic coverage, or annotation semantics. Datasheets motivate explicit documentation of those boundaries (Gebru et al., 2021). A label-only secondary study should therefore state which archive, classes, visibility codes, and eligibility filters define its rows. It should also preserve sequence and track identifiers because repeated annotations share an acquisition history. Such a study can establish useful distributional facts without supplying unobserved model outputs. A proposed generated-driving evaluation can draw on those facts when selecting tests, but it must separately obtain synchronized views, valid assessment evidence, and generated samples. The distinction keeps a modest annotation result interpretable and prevents the breadth of a benchmark's reputation from substituting for the narrower evidence actually analyzed.

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.

2. Sequence summaries and their support

Eligible annotations are Car rows with nonnegative track identifiers, positive recorded camera depth, and occlusion codes zero, one, or two. Any positive eligible occlusion code defines the binary outcome. The subset contains 26,805 rows from twenty sequences and 574 sequence-specific tracks. Each sequence receives an annotation count, mean binary occlusion fraction, and median depth coordinate. The depth variable is camera-coordinate z, not full radial distance. The analysis operates on labels and does not inspect the corresponding images or estimate scene geometry.

Occlusion fractions range from 0.111 in sequence 0012 to 0.859 in sequence 0016. Their median across sequences is 0.451. Sequence 0012 has 144 eligible rows, while sequence 0016 has 836. Another sequence, 0013, has only forty-seven eligible rows and an occlusion fraction of 0.809. These counts matter because an extreme group mean based on a small observed record does not have the same descriptive support as one based on a larger and more varied recording history.

The largest eligible sequence has 5,324 rows, while the smallest has forty-seven. The rank correlation between sequence mean occlusion and sequence median depth is -0.290 across the twenty summaries. This is an equal-sequence descriptive statistic: each sequence contributes one point regardless of its row count. It is not the same estimand as an annotation-level regression, and it does not establish a statistically confirmed negative depth effect. No new causal or generated-model inference is attached to this between-sequence correlation.

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.

A proposed explanation should be distinguishable from alternative explanations using observations the study can actually collect. Diagramming the relationships among selection, exposure, measurement, and outcome makes that requirement explicit (Greenland et al., 1999). The diagram does not establish its own correctness, but it prevents adjustment choices from remaining invisible. The same transparency is needed when investigators inspect many endpoints before choosing a narrative. Controlling false discovery concerns the behavior of a specified testing family (Benjamini and Hochberg, 1995); it is not a general certificate of scientific validity. Primary endpoints, secondary descriptions, and future mechanism tests should be named separately. Readers can then assess whether an attractive pattern was predicted, explored, or merely noticed after the available comparisons were examined.

The protected assessment population should be defined before investigators inspect detailed failures. Once a collection of cases guides repeated revisions, it has become development information even if its filenames still carry an evaluation label. Cross-validation for dependent data treats the grouping of observations as a substantive design decision (Roberts et al., 2017). Research on model selection likewise shows why tuning and final error estimation should not reuse the same information (Varma and Simon,

2006). A practical record therefore distinguishes fitting data, diagnostic examples, validation data, and the final external assessment. Where only a fixed public table is available, a descriptive analysis remains legitimate, but it should not be renamed a held-out prediction experiment without generating and assessing predictions under a protected split.

3. Why the level of comparison matters

Within a sequence, more distant observed vehicles can have one visibility relationship while sequences with larger median depth have another overall composition. There is no mathematical requirement that the two relationships share a sign or magnitude. A sequence-level median also discards the full depth distribution. The present analysis therefore avoids using its group correlation to overturn or confirm every row-level finding. The quantities answer different questions and should be reported with their levels clearly identified.

Sequence differences can reflect road layout, traffic composition, recording opportunity, and other unmeasured factors. The label summaries do not identify which factor explains the spread. They also cannot establish that the twenty eligible sequences represent a broader driving population. Reweighting them would change the emphasis among observed recordings but would not create absent conditions. A future transport claim needs observations from the target environment and a design that states how the benchmark's scene mixture relates to that intended use.

For generated-video evaluation, the practical implication is to preserve per-scene results and observation counts alongside an aggregate. A generator could perform well on common easy scenes while failing in a smaller subset with different visibility. Detecting that pattern would require actual generated outputs and valid scene-specific assessment. The current archive study supplies only the reference-data motivation. It helps define a future evaluation hierarchy while leaving cross-view compatibility, temporal quality, and downstream perception benefit to separate measurements rather than inferring them from the observed sequence spread.

A complete road-scene assessment needs a profile of outcomes. Structural similarity (Wang et al., 2004), learned perceptual correspondence (Zhang et al., 2018), generative video-distribution comparison (Unterthiner et al., 2018), and VBench's multiple dimensions (Huang et al., 2023) offer different measurement resources. Their relevance depends on the claim, reference availability, and observation unit. A strong score in one representation does not establish that a vehicle occupies a compatible location in every camera or persists coherently through an occlusion. The evaluation should therefore retain targeted scene checks alongside general visual measurements. It should also identify clips or conditions for which a score cannot be obtained validly. Label summaries can help define such difficult conditions, but they are not substitutes for generated visual observations. A secondary analysis of annotations remains a description of recorded objects and visibility. Any later claim about appearance, synchronization, or temporal quality requires the corresponding output data and a separately implemented assessment rather than an inference from the archive's label distribution.

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.

4. Assessing a proposed driving-world claim

A learned world representation can serve several tasks, and success in one does not settle the others. World Models (Ha and Schmidhuber, 2018) and Dreamer (Hafner et al., 2019) provide influential settings linking prediction and behavior. CLIP's visual-language representation (Radford et al., 2021) and IP-Adapter's image conditioning (Ye et al., 2023) concern other information paths. Offline reinforcement learning (Levine et al., 2020) emphasizes that learning from fixed data introduces support and generalization constraints. A driving-generation study should use these distinctions to define its own claim. Producing a semantically appropriate clip, maintaining a compatible multi-view scene, improving a detector, and supporting a control decision require different observations. The present label-based evidence can motivate where to test but cannot establish those downstream outcomes. A future experiment should declare the complete pipeline, independent assessment distribution, and relevant cost. This keeps the interpretation proportional to the available evidence while allowing useful intermediate scene-generation findings to be reported without an unsupported leap to broad driving capability.

Coverage should be reported whenever the system or its evaluator can avoid difficult cases. An abstention policy changes the population on which an accuracy or fidelity score is calculated. Selective classification formalizes the joint reporting of coverage and conditional performance (Geifman and El-Yaniv, 2017). Distribution-shift research adds that estimated uncertainty can behave differently outside the development population (Ovadia et al., 2019). A benchmark should therefore retain which observations were eligible, accepted, rejected, or unavailable. A result on a carefully limited subset can be useful, provided the subset is visible. The challenge is to prevent a narrow success claim from becoming a broad capability claim through omission of the denominator or the circumstances in which the system declines to operate.

The unit of provenance should be fine enough to reconstruct an important disagreement. Reusable identifiers can connect a reported score to its input evidence and the relevant model and evaluator versions (Wilkinson et al., 2016). Pipeline dependencies should be treated as part of the experiment because apparently minor changes can alter output eligibility or metric behavior (Sculley et al., 2015). A report that preserves only aggregate numbers prevents this reconstruction. Keeping per-case results and documented transformations enables alternative weighting and error analysis without repeating expensive generation. Such a record also helps identify whether two papers compare the same scientific objects or merely use similar names for different releases. Versioned evidence makes progress easier to assess across successive evaluations.

5. Conclusion

The eligible KITTI sequences have substantial variation in occlusion fractions, unequal observation counts, and a modest negative between-sequence depth correlation. These are descriptive group-level findings, not a causal depth effect or a test of OmniDrive. Their contribution is to make recording composition visible. A future driving-generation comparison should retain that hierarchy so that an overall score can be interpreted in relation to the scenes and observation opportunities that actually support it.

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.