Curvature and Sparse Tails in Driving-Scene Depth Coverage

Owen Ward (GN), Dylan Bailey (TO), Evan Graham (DJ)

Abstract

A quadratic specification suggests curvature in the association between recorded depth and car occlusion in KITTI tracking labels. The coefficient on squared base-two log depth is -0.0246, with a 95 percent interval from -0.0409 to -0.0083. Its algebraic turning point lies near 96.6 depth units, but only fifteen eligible annotations have depth at least eighty. This secondary analysis emphasizes that fitted curvature and supported tail inference are different matters. It does not establish a physical visibility law or a generated-driving performance result.

1. A straight-line association can miss distributional shape

A single depth coefficient summarizes a relationship under a selected functional form. It can be useful without describing every part of the observed range. A curved association may instead arise from scene composition, visibility selection, or other unmeasured properties. This study adds a quadratic term to a conventional label-based regression and examines the resulting support boundary. Its question is whether the fitted shape can be interpreted responsibly when the extreme depth region contains very few observations.

OmniDrive proposes a multi-view driving world model with coordinated scene generation (Meng et al., 2026d). That setting motivates evaluation over a declared range of object depths. The present analysis does not generate distant vehicles or evaluate the focal method's geometry. It studies reference labels to identify why a seemingly informative fitted curve can become unreliable evidence for rare conditions. The empirical contribution is a supported annotation relationship and an explicit warning about its sparse tail, rather than a new claim of world-model performance.

Driving datasets describe particular observation systems rather than a universal road environment. KITTI provides influential autonomous-driving benchmarks (Geiger et al., 2012); nuScenes supplies a multimodal driving collection (Caesar et al., 2019); BDD100K emphasizes diverse driving-video tasks (Yu et al., 2018); and the Waymo Open Dataset offers another large-scale perception setting (Sun et al., 2019). Their sensors, annotations, sampling, and scene populations differ. Dataset documentation should connect the chosen files to their intended use and known limits (Gebru et al., 2021). An analysis of a released annotation archive should identify the exact subset and label definitions it processes. It cannot inherit every modality or capability associated with the dataset's broader name. In particular, a study using only labels has not measured images, calibrated camera geometry, run a detector, or generated a new road scene. Its empirical contribution can still be useful: it describes the observed annotation population and motivates an assessment condition. Transfer to another collection or a generated-video system requires the corresponding observations rather than a citation alone.

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.

2. The quadratic specification and its observed range

Eligible Car annotations have nonnegative track identifiers, positive recorded camera-depth coordinate, and occlusion codes zero, one, or two. The subset contains 26,805 rows from 574 sequence-specific tracks in twenty sequences. The outcome indicates any positive eligible occlusion code. The model includes base-two log depth, its square, and sequence indicators. Ordinary least squares is used as a linear probability specification, with sequence-clustered standard errors and the saved t-based confidence intervals.

The coefficient on log depth is 0.3245, with a 95 percent interval from 0.2079 to 0.4411. The squared-term coefficient is -0.0246, with an interval from -0.0409 to -0.0083. The reported coefficient of determination is 0.175. The positive first-order and negative quadratic terms describe a concave fitted shape on the chosen log scale. Neither coefficient alone is a constant probability change over the whole depth range, because the fitted slope depends on the value of log depth when the quadratic term is present.

The algebraic turning point is 6.594 on the base-two log scale, corresponding to approximately 96.6 in recorded depth units. Only fifteen eligible annotations have depth at least eighty. That support count is crucial to interpretation. A turning point can be calculated from the coefficients even when the surrounding data are sparse. The study does not attach a turning-point confidence interval or present the calculated location as a precisely observed peak in a physical visibility process. Recorded z is also a camera-depth coordinate, not full radial distance.

Precision and identification are separate properties of a result. A narrow interval can surround an association whose scientific target is unclear. Robust covariance calculations allow particular departures from simple error assumptions, including unequal variance (White, 1980) and dependence within declared clusters (Cameron and Miller, 2015). Their role is to improve uncertainty calculations under those assumptions. They do not establish a causal assignment, guarantee a correct measurement, or expand coverage to conditions missing from the data. The analyst should therefore explain both the averaging population and the source of replication. An evaluation repeated many times on one history may provide information about stochastic variability while supplying little evidence about transfer to another history. Those two sources of uncertainty should remain distinguishable in the final interpretation.

The scientific interpretation should specify what would change under the proposed intervention. Holding an observed covariate fixed statistically is not equivalent to manipulating it in an environment. A causal graph helps separate confounders, mediators, and selection mechanisms (Greenland et al., 1999). It can also reveal why controlling for every recorded variable is not automatically defensible. Meanwhile, an evaluation with multiple conditions and endpoints creates a family of opportunities for apparently striking findings. Multiplicity methods require a declared set of comparisons and an appropriate inferential purpose (Benjamini and Hochberg, 1995). Descriptive tables may be useful without testing every cell. When the data do not identify a mechanism, the appropriate contribution can be a sharper question and a transparent comparison rather than an unsupported causal answer.

Visual plausibility is only one dimension of a generated driving scene. Structural similarity (Wang et al., 2004) and learned perceptual similarity (Zhang et al., 2018) compare particular image properties, while video-distribution evaluation (Unterthiner et al., 2018) and VBench's multidimensional framework (Huang et al., 2023) address other aspects of generated output. None alone establishes that synchronized

camera views depict one compatible scene. A method can produce sharp, individually credible frames with inconsistent object placement across views. It can also preserve geometry while showing visible temporal defects. The evaluation should retain these outcomes separately and connect each instrument to the reference information it requires. A label-only analysis cannot calculate these visual measures because it has not processed the necessary images or clips. Its role is to describe annotation conditions that may deserve targeted generation tests. A subsequent experiment would need actual outputs, declared sampling, synchronized assessment, and an uncertainty policy. Those measurements should be reported as new evidence rather than inferred from favorable distributions in the reference label archive.

3. Curvature is useful only within an evidence boundary

The polynomial is one descriptive specification, not proof that the true relationship has a universal quadratic form. Sequence indicators do not control every scene variable, and object eligibility can vary across depth. A different flexible model could produce another shape, particularly where observations are rare. The current result therefore supports examining nonlinearity, while the tail count limits how strongly the extreme fitted behavior should influence scientific or practical conclusions about distant objects.

Linear probability models also require care when used outside their observed support. The analysis has not established that every possible extrapolated prediction is a valid probability. It uses the model to describe the eligible table under a declared specification. A future study could predefine depth bands, obtain additional distant observations, or compare functional forms under an independent assessment plan. Those steps would add evidence rather than merely make the same sparse tail appear smoother through a more elaborate mathematical representation.

For generated-video evaluation, the practical implication is to report depth coverage alongside performance. A model could receive a favorable overall score with almost no assessable distant objects. A fitted trend in reference labels cannot fill that gap. The generation study would need actual distant-scene outputs, appropriate visibility and correspondence measurements, and an uncertainty policy for small projected objects. The current archive analysis supplies a reason to collect those observations, while leaving the generator's rare-condition behavior untested rather than borrowing confidence from an algebraic turning point.

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.

A score is useful only relative to the question it is supposed to answer. Correlation measures a relationship between varying quantities, whereas agreement concerns the discrepancies produced when one procedure substitutes for another (Bland and Altman, 1986). Confidence estimates add a further question about calibration (Guo et al., 2017). These distinctions matter when a learned evaluator becomes the principal source of evidence about another learned system. The evaluation should preserve examples with conflicting scores and obtain independent target judgments where feasible. Numerical ranges and normalization choices should also be documented. An apparently small difference may have a different

practical meaning near a decision threshold than in an unimportant part of the scale. Reporting the intended consequence gives those comparisons an interpretable purpose.

4. Assessing a proposed driving-world claim

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.

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.

Reproducibility begins with preserving which scientific objects are being linked. Stable identifiers and reusable metadata help connect inputs, reference observations, model versions, and reported outcomes (Wilkinson et al., 2016). The scoring pipeline also has dependencies that can change the interpretation of a result: preprocessing, representation checkpoints, thresholds, and aggregation code can all matter (Sculley et al., 2015). A compact release should retain those decisions rather than only the final mean. Where source data cannot be redistributed, an access description and hashes can still identify the analyzed inputs. This record allows later users to distinguish a change in the underlying system from a change in how the system was measured. Reproducibility is therefore part of interpretation, not simply a final packaging step.

5. Conclusion

The quadratic label regression shows concavity under the chosen specification, but its calculated extreme-depth turning point lies in a sparsely observed region. This is a descriptive finding with a clear support limitation. It does not identify a physical law or validate a driving generator. Its contribution is to connect functional-form interpretation to observed coverage, ensuring that a smooth fitted curve does not make rare depth conditions appear more empirically established than the archive permits.

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.