Distance and Visibility in Driving Video: A Track-Aware Analysis of KITTI Annotations

Benjamin Allen (FI), Jeremy Harper (DK), Mark Mason (BY)

Abstract

Nearly half of the eligible car annotations in KITTI's public tracking labels are marked as occluded. The share rises from 14.5 percent below ten meters to 59.2 percent between twenty and forty meters, then declines in the next distance band. Across 26,805 annotations, a sequence-adjusted linear probability model associates a doubling of depth with 12.8 percentage points more occlusion. Equalizing the influence of tracks gives a similar, slightly larger association. These secondary results show why driving-video evaluation should preserve distance, visibility, and track structure. They do not measure the fidelity or safety of a generated world.

1. A plausible driving scene can hide a weak evaluation

A generated driving video may contain recognizable cars, lanes, and buildings while representing the difficult objects poorly. Distance changes apparent size, and occlusion changes how much evidence is available for maintaining an object's identity and trajectory. If a benchmark mostly presents nearby, unobstructed vehicles, an average quality score can say little about the conditions under which perception becomes difficult. A useful preliminary analysis therefore asks which distance and visibility combinations are actually represented in a public driving annotation set.

OmniDrive describes DRIVE-CHOREO, a multi-agent world-model framework that coordinates language, geometry, and multi-view video through a structured latent representation (Meng et al., 2026d). Its reported Director, Cartographer, and Auditor have distinct roles in interpreting intent, grounding layout, and feeding back cross-view critiques. Co-compression is intended to connect those controls with the generated views. These claims motivate a concrete evaluation requirement: preserve the underlying objects and their visibility conditions when judging the videos. The analysis here does not reproduce the model, inspect its generated clips, or verify its reported downstream improvements.

The empirical contribution is a secondary analysis of KITTI tracking annotations. KITTI was established to support real-world vision evaluation with calibrated sensing and challenging road scenes (Geiger et al., 2012). This article uses the public tracking label archive, rather than the separate object-detection split or a collection of newly generated frames. It asks how occlusion labels co-vary with recorded depth and whether that association changes when long tracks receive less influence.

2. From label rows to eligible vehicle observations

The downloaded archive contains 65,301 label rows across 21 training sequences. Its format records frame number, track identifier, object category, truncation, occlusion, bounding-box information, three-dimensional location, and orientation. The developer documentation is retained with the raw archive. An annotation is a statement about an object in a particular frame; it is not an independent vehicle journey, a fresh recording, or an independent draw from a deployment environment.

Eligibility is defined before modeling: the category must be Car, the track identifier nonnegative, the recorded depth positive, and the occlusion code one of zero, one, or two. The outcome is one when the occlusion code is one or two and zero when it is zero. Other categories, ignored regions, invalid track identifiers, nonpositive depth, and unknown occlusion codes do not enter the estimation sample. These rules leave 26,805 car annotations from 574 sequence-specific tracks and twenty sequences. The sample includes 6,457 distinct sequence-frame combinations containing eligible cars.

The recorded z coordinate is used as depth, in meters. It is not reconstructed from pixel height and is not treated as full radial distance from the sensor. A track key combines sequence and track identifier, since the same numeric identifier can recur in another sequence. No image, video, LiDAR scan, detector output, or vehicle control log is processed. The analysis cannot therefore estimate perception errors, frame quality, collision risk, or behavioral realism.

Documentation should make those absences as visible as the variables that are present. Datasheets provide a framework for recording a dataset's intended scope and limitations (Gebru et al., 2021), while reusable identifiers and provenance records support later integration of related observations (Wilkinson et al., 2016). The processed table preserves the original fields needed for the stated filters and the derived variables used in estimation. The count of eligible annotations can therefore be checked without reconstructing the source archive from a verbal description.

3. A descriptive model with two weighting schemes

The main specification is a linear probability model of any recorded occlusion on log-base-two depth and sequence fixed effects. Its slope describes the average fitted difference associated with doubling recorded depth within the modeled sequence context. It is not a causal effect of physically moving a vehicle farther away. Lane position, traffic density, object selection, camera orientation, and scene composition can influence both depth and visibility. A causal graph would require those processes before supporting an intervention interpretation (Greenland et al., 1999).

Each annotation receives equal weight in the first model. Long visible trajectories can consequently contribute many more rows than short trajectories. The second model gives each annotation the reciprocal of its track's eligible row count. Each of the 574 tracks then contributes equal total weight. This does not make tracks representative of an external vehicle population. It asks whether the relationship is similar when the scientific unit shifts from the average recorded car-frame to the average observed track.

Both covariance calculations cluster by sequence, with a finite-cluster adjustment and a t reference distribution. Vehicles and frames from the same recording share conditions, so treating annotation rows as independent would exaggerate the effective amount of information (Cameron and Miller, 2015). Heteroskedasticity-robust covariance addresses variation in residual variance but not the full recording-level dependence (White, 1980). There are only twenty contributing sequence clusters; the resulting intervals should be read with that limitation in mind.

Additional summaries include five depth bands and leave-one-sequence-out point estimates. The bands show the distribution without requiring a linear relationship. The omission exercise checks whether a single recording drives the fitted slope. It does not create twenty new datasets, and its range is not a confidence interval. The primary model and filters were not selected by searching across those omission results.

4. Visibility changes with depth, but not monotonically across all bands

Among eligible annotations, 49.17 percent have a nonzero occlusion code and median depth is 26.95 meters. The overall fraction combines near and distant objects, long and short tracks, and different sequences. It is a property of this filtered archive rather than an estimate of how frequently an arbitrary car in road traffic is occluded.

Recorded depth in meters	Annotations	Occluded share	Distinct contributing tracks
Below 10	3,141	14.5 percent	319
10 to below 20	5,208	44.7 percent	432
20 to below 40	11,740	59.2 percent	508
40 to below 80	6,701	51.3 percent	296
80 and above	15	40.0 percent	9

Tracks can contribute to more than one band as their recorded depth changes, so the last column must not be summed to recover the number of unique tracks. The tiny final band also illustrates why a percentage alone is insufficient: forty percent there represents only fifteen annotations from nine tracks. There is no basis for a precise claim about very distant vehicles from that cell.

The band pattern is not monotonically increasing. Occlusion is lower between forty and eighty meters than between twenty and forty meters. This can reflect the composition and selection of observed objects as well as scene geometry. The positive regression coefficient summarizes a fitted relationship; it does not replace that observed pattern with a universal law. A flexible model could describe curvature, but the current analysis does not fit additional terms merely to improve the appearance of the result.

The annotation-weighted slope is 0.1280, with a sequence-clustered standard error of 0.0170 and a 95 percent interval from 0.0925 to 0.1635. In probability units, this corresponds to 12.8 percentage points per doubling of depth. The equal-track slope is 0.1427, with an interval from 0.1187 to 0.1667. Agreement in sign and broad magnitude shows that long tracks alone do not account for the positive association. It does not establish that either weighting scheme represents a deployment distribution.

Leaving out each sequence in turn produces slopes from 0.1145 to 0.1338 in the annotation-weighted specification. The main positive result is therefore not dependent on one particular retained recording. The model's in-sample R-squared is 0.169, leaving substantial variation unexplained. No held-out occlusion forecast or detector improvement is reported.

5. What these labels can contribute to world-model evaluation

The results identify a practical stratification variable. If a generated dataset is used to train a perception system, the evaluation should preserve depth and occlusion so that performance can be reported within supported strata. Matching the marginal number of cars does not ensure comparable exposure to partially hidden vehicles. Nor does matching an overall occlusion fraction guarantee a realistic relationship between visibility and location. Joint distributions matter.

Large driving datasets illustrate complementary coverage. nuScenes provides a multimodal autonomous-driving collection (Caesar et al., 2019), BDD100K emphasizes diverse driving conditions and heterogeneous tasks (Yu et al., 2018), and the Waymo Open Dataset expands the scale of driving perception data (Sun et al., 2019). Their existence does not establish that a model trained or tested on one transfers to the others. The KITTI archive analyzed here supplies a historical, geographically and procedurally limited observation set. Its annotations cannot verify multi-camera consistency in OmniDrive's generated scenes.

Tracking metrics also separate endpoints that an average image score can miss. HOTA balances detection and association considerations (Luiten et al., 2020). ByteTrack connects low-confidence detections to the association problem (Zhang et al., 2021), while AB3DMOT demonstrates a comparatively simple

three-dimensional tracking approach and evaluation perspective (Weng et al., 2020). None of these trackers or metrics is executed here. They make clear why retaining an object's identity across temporary invisibility is a different accomplishment from generating an appealing individual frame.

An appropriate follow-up would pair generated views with object-level scene records, including intended positions, visibility, and persistent identities. It would then test whether the generated evidence supports consistent observations across cameras and time. Such a design could reveal whether a structured latent control is obeyed or merely encoded in an input. The present labels help specify the evaluation strata; they are not outputs from that proposed study.

6. Geometry, visual fidelity, and behavior need separate endpoints

Video-distribution metrics can summarize aspects of temporal visual realism (Unterthiner et al., 2018), but they do not automatically determine whether the same vehicle has compatible positions in two cameras. Structural similarity (Wang et al., 2004) and learned perceptual similarity (Zhang et al., 2018) compare visual observations under their own reference assumptions. Their scores should be accompanied by geometry and association endpoints when those are the intended claims. A favorable number on one dimension cannot supply a missing observation on another.

World Models investigates compact representations of environment experience (Ha and Schmidhuber, 2018), and Dreamer connects latent imagination to behavior learning (Hafner et al., 2019). Those studies illustrate why a world model may be judged through a downstream decision as well as visual outputs. In driving, however, behavioral claims require suitable actions, trajectories, and outcomes. This annotation archive contains none of the intervention information needed to show that a generated environment is safe for planning.

Conditioning also needs a clear information boundary. IP-Adapter provides image-based conditioning compatible with text (Ye et al., 2023), and CLIP learns representations connecting language and imagery (Radford et al., 2021). An evaluation should record which geometric, textual, and visual inputs each compared method receives. VBench's multiple video-quality dimensions provide a useful precedent for decomposing the final assessment (Huang et al., 2023). OmniDrive's agent roles create an additional reason to preserve intermediate intended layouts and critiques for inspection, without assuming that an internal agreement establishes external correctness.

Offline reinforcement learning highlights the limitation of a fixed behavioral support (Levine et al., 2020). The parallel here is a limitation in observational support: a label set with very few distant annotations cannot substantiate broad claims about distant-object behavior. Reweighting can alter emphasis among observed records, but it cannot create missing visibility configurations, adverse weather, or intervention histories.

7. Designing a test beyond this archive

A next study should specify a target distribution over scenes, tracks, and frames. If the target is track-level robustness, equal track exposure is defensible; if it is a frame encountered during deployment, exposure should follow a justified traffic distribution. Both should be reported when their difference is scientifically important. Dependence should guide train-test separation at the scene or sequence level (Roberts et al., 2017), and architecture selection should not reuse the final evaluation scenes (Varma and Simon, 2006).

Confidence and rejection policies would need direct observation. Calibration asks whether predicted probabilities correspond to frequencies (Guo et al., 2017), while distribution shift can change uncertainty behavior (Ovadia et al., 2019). Selective prediction distinguishes performance on accepted examples from

the fraction of examples a system accepts (Geifman and El-Yaniv, 2017). The current linear probability fit is a descriptive association and is not offered as a calibrated deployable predictor.

If a proposed geometry measure replaces a human or task endpoint, their agreement should be studied rather than inferred from correlation (Bland and Altman, 1986). Multiple depth bands, object classes, camera pairs, and metrics create opportunities for selective reporting; primary comparisons and multiplicity handling should be declared (Benjamini and Hochberg, 1995). Finally, stable metadata and explicit scoring dependencies help keep later evaluations comparable (Sculley et al., 2015). These records are as important to interpreting progress as the headline score.

8. Conclusion

The public KITTI tracking labels show a positive sequence-adjusted association between recorded depth and car occlusion, with a similar result when tracks receive equal weight. The depth bands also reveal nonlinearity and a very sparse distant tail. Driving-video evaluation should preserve those distinctions instead of treating annotation rows as independent or an overall visibility fraction as sufficient. The analysis provides a reproducible description of observed road-scene evidence. Multi-view fidelity, downstream perception utility, and behavioral safety require their own tests on generated outputs and appropriate external data.

References

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.

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).

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.