Image-Edge Position and Recorded Vehicle Occlusion

Isaac Harper (SS), Daniel Walters (NA), Cameron Barker (AM)

Abstract

Eligible KITTI car boxes with a left coordinate at most five pixels have a raw occlusion fraction of 0.178, compared with 0.516 for other boxes. After depth and sequence adjustment, the estimated difference is -0.079, with a 95 percent interval from -0.184 to 0.027. This analysis concerns a declared left-coordinate boundary, not every image edge. It uses labels only and assumes no right-edge image width. The result illustrates how a strong raw visibility contrast can change under an explicit observational comparison.

1. An edge condition needs a known coordinate rule

An image-edge indicator can be defined in several ways, but each requires appropriate information. A left-coordinate threshold can be constructed directly from released boxes. A right-edge threshold also needs the relevant image width, which this label-only analysis does not assume. The present study therefore uses one deliberately narrow condition: the annotated left coordinate is at most five pixels. It asks how recorded occlusion differs under that rule and how the contrast changes after available adjustment.

OmniDrive proposes generated driving scenes with multiple synchronized views (Meng et al., 2026d). Boundary objects are relevant to that evaluation agenda because their observation coverage can differ across cameras. The present study does not inspect generated views, reconstruct image geometry, or assess the focal method. It describes a public annotation relationship. The empirical contribution is a clearly defined coordinate condition and its observed visibility contrast, without presenting one left-boundary proxy as a complete measurement of all edge positions or truncation events.

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.

Road-scene benchmarks vary in what they observe and what their annotations mean. KITTI (Geiger et al., 2012) and nuScenes (Caesar et al., 2019) provide different autonomous-driving data settings, while BDD100K (Yu et al., 2018) and the Waymo Open Dataset (Sun et al., 2019) broaden the available tasks and collections in other ways. A study should select a source because its records fit the scientific question, not because several benchmark names appear together in the literature. Documentation of collection, processing, and intended use makes that choice assessable (Gebru et al., 2021). The same discipline applies within one source: tracking annotations and detection annotations need not share every label convention. An analyst should consult the relevant archive definition rather than import a familiar interpretation from another task. When only labels are processed, conclusions concern those labels and their observed relationships. Visual realism, cross-view compatibility, and downstream detector utility remain further outcomes that require appropriate inputs and measurements. A clear evidence boundary makes the annotation analysis useful without presenting it as a completed evaluation of a driving generator.

2. Data, threshold, and adjusted specification

The input is the KITTI tracking training-label archive. Eligible rows are Car annotations with nonnegative track identifiers, positive camera-depth coordinate, and occlusion code zero, one, or two. Any positive eligible occlusion code defines the binary outcome. There are 26,805 rows in twenty sequences, linked to 574 sequence-specific tracks. The boundary indicator equals one when the released left box coordinate is at most five pixels. It does not use an inferred camera width, a normalized image fraction, or a symmetric right-edge condition.

The boundary group contains 1,900 rows. Its mean binary occlusion is 0.178, compared with 0.516 for the other eligible rows. These are annotation-weighted summaries. They do not imply that every boundary vehicle is fully visible or that the left-coordinate threshold identifies all forms of truncation. Box location and the archive's truncation and occlusion labels remain distinct recorded properties. The analysis preserves that distinction rather than treating a convenient coordinate rule as a substitute for every visibility annotation.

A linear probability model relates any occlusion to the boundary indicator, base-two log depth, and sequence indicators. It uses sequence-clustered standard errors with the saved t-based interval. The boundary coefficient is -0.0790, with a 95 percent interval from -0.1845 to 0.0265. Thus the adjusted point estimate remains negative but is much smaller than the raw contrast and uncertain under the chosen specification. The interval includes zero, so the result does not establish a precisely negative conditional difference in a broader population.

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.

The tracking literature makes the unit of a temporal claim explicit. HOTA addresses higher-order tracking assessment (Luiten et al., 2020), ByteTrack develops an association approach (Zhang et al., 2021), and AB3DMOT concerns three-dimensional multi-object tracking (Weng et al., 2020). A road-scene study can use those distinctions to separate objects, annotations, trajectories, and predictions. They should not be pooled simply because each occupies a row in a table. A label-only analysis describes the reference trajectories that were recorded; a tracker benchmark requires matching predictions to those references. A generated-video consistency study needs another explicit definition of cross-view and temporal object identity. The relevant records should retain sequence, time, class, and correspondence information wherever available. Missing or ambiguous matches should remain visible rather than become assumed successes. This design connects descriptive visibility evidence to future evaluation while respecting the absence of newly run tracking models or generated scenes in the present annotation archive.

3. What the changing contrast does and does not show

The difference between raw and adjusted summaries indicates that the comparison depends on observed depth and sequence composition. It does not establish that those controls capture every relevant scene property. Boundary position can be associated with vehicle paths, camera motion, object selection, and other unmeasured conditions. The model is observational. Moving an otherwise identical vehicle to the boundary is not an intervention carried out by this study, and the coefficient should not be interpreted as its causal effect on visibility.

The threshold is also a chosen descriptive rule. Five pixels is not claimed to be a universal physical boundary or a validated failure threshold for a perception model. A planned sensitivity study could examine several declared thresholds, but the present result should remain tied to the one actually computed. The analysis does not expand the condition to the right edge or all camera views without additional width and calibration evidence. That restraint is necessary for the coordinate-based finding to remain reproducible.

For generated-road evaluation, a future protocol could collect explicit boundary and truncation conditions across synchronized cameras. It would need actual clips, known image geometry, object correspondences, and valid visibility assessment. A reference-data contrast cannot establish whether a generator creates or preserves compatible boundary objects. The current result instead helps define a narrower sampling condition and shows why raw averages deserve composition checks. It also keeps unavailable visual and geometric evidence separate from the label calculations already completed in the public archive.

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.

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.

Difficult cases can disappear at several stages: data eligibility, generation, measurement, and final scoring. Each exclusion changes the population represented by the reported result. Coverage-aware evaluation makes those decisions visible (Geifman and El-Yaniv, 2017). A confidence score should also be checked when the input distribution changes, because uncertainty under shift is an empirical property

rather than a guarantee of the model interface (Ovadia et al., 2019). An informative record preserves failures to produce an output and failures to obtain a valid measurement. Treating either as absent data without explanation can inflate apparent reliability. A proposed method should state its unsupported conditions and test the boundary between accepted and rejected cases as carefully as the quality of its successful outputs.

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.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

An evaluation remains comparable only if its inputs and transformations can be identified. The FAIR principles emphasize persistent identification and reusable scientific metadata (Wilkinson et al., 2016). Work on machine-learning technical debt highlights how dependencies and hidden couplings accumulate in a pipeline (Sculley et al., 2015). In this setting, a score should be linked to the exact references, generation settings, evaluator version, and aggregation rule that produced it. A replacement checkpoint or a revised eligibility filter may change the scientific question even when the output column keeps the same name. Preserving such changes in a release record permits later comparisons to be interpreted honestly. The objective is not maximum documentation volume, but a recoverable account of the decisions that affect the result.

5. Conclusion

The declared left-boundary group has a much lower raw occlusion fraction, while the adjusted contrast is smaller and uncertain. This is an annotation association under a specific coordinate rule. It does not cover every image edge or establish a causal position effect. Its contribution is a reproducible example of how boundary definitions, composition adjustment, and data availability shape the interpretation of visibility evidence used to design future driving-generation tests.

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.