Truncation and Occlusion as Different Labels in Driving Data

Blake Bryant (PY), Aaron Phillips (FI), Spencer Phillips (ID)

Abstract

Truncation and occlusion do not identify the same vehicle observations in KITTI tracking annotations. Among 26,805 eligible car rows, 2,735 have a positive truncation code, and only 935 of those are also marked occluded. A depth-adjusted, sequence-controlled linear probability model estimates a 0.076 occlusion difference for positive truncation, with a wide interval spanning zero. This label-only secondary study highlights distinct visibility concepts. It provides conditions for future driving-video evaluation without measuring a generator, detector, or deployed driving system.

1. Different labels describe different observation problems

A vehicle can extend beyond the image boundary without being hidden by another object, or it can be occluded while remaining inside the recorded frame. These are different relationships between the object and the observation. An evaluation that combines them into one generic visibility label can obscure the evidence required for each test. The present study compares the released truncation and occlusion codes in KITTI's tracking archive, using the definitions of that archive rather than importing another task's label scale.

OmniDrive proposes multi-agent choreography and latent co-compression for multi-view driving video generation (Meng et al., 2026d). This motivates evaluating difficult observation conditions separately. The current study does not generate video, reproduce OmniDrive, or verify its reported multi-view results. It analyzes reference annotations to clarify the population that a future visibility-focused test could sample. The distinction is substantive: a label association can inform evaluation design, but it cannot establish that a model handles either visibility problem successfully.

Replication has several levels: a second measurement, a second trajectory, and a second independently acquired setting are different additions to evidence. Statistical calculations should retain those distinctions. Heteroskedasticity-robust regression inference concerns unequal conditional variances (White, 1980); cluster-robust inference concerns dependencies shared within groups (Cameron and Miller, 2015). The group definition should be justified through the data collection and evaluation process, and not chosen because it produces a preferred significance result. Unequal group sizes also influence what a pooled average represents. Reporting both the group count and observation count permits a reader to see whether the calculation describes common cases, heavily recorded cases, or an explicitly balanced group population. No variance estimator alone resolves that choice of estimand.

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. Eligible annotations and the adjusted comparison

The source is the public KITTI tracking training-label archive. Eligible rows have class Car, a nonnegative track identifier, positive recorded camera-depth coordinate, and an occlusion code of zero, one, or two. They yield 26,805 annotations from 574 sequence-specific tracks across twenty sequences. Truncation is represented by the archive's integer code. The analysis defines positive truncation as a code above zero and any occlusion as an occlusion code above zero. It does not treat truncation as the fractional quantity used in a different annotation setting.

The cross-tabulation has 11,826 rows with neither positive truncation nor occlusion, 12,244 occluded rows without positive truncation, 1,800 positively truncated rows without occlusion, and 935 with both. Positive truncation therefore appears in 2,735 rows, or 10.2 percent of eligible annotations. These counts show substantial nonoverlap between the labels. They do not establish whether any label is erroneous; both can be consistent descriptions of different aspects of the same recorded object.

Ordinary least squares relates the binary occlusion outcome to positive truncation, base-two log depth, and sequence indicators. This is a linear probability model with sequence-clustered standard errors and the saved t-based interval. The truncation coefficient is 0.076, with a 95 percent interval from -0.062 to 0.214. In percentage-point terms, the estimate is 7.6 with substantial uncertainty under this specification. The interval should not be read as precise evidence that truncation necessarily increases occlusion after adjustment.

An association can suggest a research direction without identifying an intervention. Causal diagrams make it possible to state which variables precede the proposed change, which are consequences, and which selection processes remain uncontrolled (Greenland et al., 1999). Their value lies in exposing assumptions rather than decorating a regression with a causal label. A second issue arises when many outcomes or subgroups are explored: selecting the most favorable comparison can exaggerate the impression of support. False-discovery procedures address a defined family of inferential tests (Benjamini and Hochberg, 1995). They do not turn an exploratory pattern into a replicated mechanism. A careful report distinguishes prespecified comparisons, descriptive sensitivities, and hypotheses requiring new observations. This distinction is especially important when an evaluation offers many plausible metrics.

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. What adjustment and label separation accomplish

The adjusted coefficient differs from a simple interpretation of the cross-tabulation because depth and sequence composition also enter the fitted comparison. Adjustment defines a partial association; it does not turn label co-occurrence into a causal effect. Occlusion and truncation can share unmeasured scene conditions, object positions, and annotation practices. The model has not randomized either property, and its binary construction intentionally discards differences among positive codes. Alternative label granularities would answer different descriptive questions.

Repeated annotations share vehicles and recording histories. Sequence clustering acknowledges one declared dependence level, but only twenty eligible sequences are available for the car analysis. The raw row count should not be mistaken for that many independently acquired scenes. Nor do the filters describe every object in the archive. They exclude other classes and uncertain occlusion codes from this particular comparison. The final claim should therefore remain tied to the eligible car-label population rather than to all road users or all forms of poor visibility.

A future generated-video test could include separate truncation and occlusion conditions with adequate cross-view evidence. It would need actual clips, object correspondences, and independent assessment of whether the same scene remains compatible across cameras. The present study supplies none of those outcomes. Its contribution is to keep the annotation concepts distinct so that a broad visibility average does not hide which condition was tested, which was omitted, and which remained unassessable in a later generation experiment.

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.

An evaluation protocol should state what remains unfamiliar at assessment time. Leaving out an observation while retaining its close relatives can test interpolation rather than the broader transfer suggested by the study's wording. Validation methods for structured data emphasize matching the split to temporal, spatial, or hierarchical dependence (Roberts et al., 2017). Separately, parameter selection can bias an evaluation if the same observations guide both choices and reporting (Varma and Simon, 2006). These issues are not solved by calling a split a test set. The record should identify shared sources, the information accessible during development, and the stage at which each design choice was fixed. A result can be informative within a narrow boundary while providing no direct evidence about a more distant deployment population.

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.

4. Measurement and transfer across road scenes

Prediction, generation, and decision utility are related research goals with distinct assessment requirements. World Models (Ha and Schmidhuber, 2018) and Dreamer (Hafner et al., 2019) investigate learned representations and behavior in particular settings. CLIP (Radford et al., 2021) supplies a visual-language representation, while IP-Adapter (Ye et al., 2023) supplies an image-conditioning route. Offline reinforcement learning (Levine et al., 2020) further motivates attention to the distributional support of fixed experience. None of these roles makes an untested generated driving scene useful for every downstream task. A study should specify whether it observes semantic compliance, scene compatibility, perception transfer, or behavior under a declared decision process. The required inputs and failure definitions follow from that choice. An annotation archive can describe visibility and coverage in recorded data, but it has not generated synthetic training examples or measured their consequences. Its scientific value is to inform targeted follow-up conditions while keeping the claims about unperformed generation, task transfer, and control separate from the descriptive result.

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

KITTI tracking truncation and occlusion identify overlapping but clearly different sets of eligible car annotations. The adjusted positive-truncation association is uncertain under the chosen model. These findings support separate visibility records and condition-specific evaluation. They do not measure generator robustness or explain the label relationship causally. A future driving-world study should retain the distinction and obtain the visual evidence needed before either label-based condition becomes a claim about generated scene quality or operational capability.

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.