Frame Crowding and Vehicle Visibility in Road-Scene Annotations

Steven Hoffman (KE), Spencer Marshall (SV), Connor Allen (KP)

Abstract

Frames containing more eligible cars also contain a higher adjusted frequency of occluded car annotations in KITTI tracking data. A linear probability model estimates a 0.0277 increase per additional eligible car, controlling for recorded log depth and sequence. Its 95 percent interval is 0.0161 to 0.0394. The analysis covers 26,805 car rows in 6,457 frames. The count describes the filtered annotation population rather than total traffic density. It offers a crowding-related test condition without measuring detector performance or generated-scene consistency.

1. Crowding needs an observable definition

Traffic density can refer to vehicles per road length, objects in an image, or several other quantities. A label-only archive does not automatically provide every denominator needed for those definitions. This study uses a narrower measure: the number of eligible car annotations in the same recorded frame. It asks whether an individual eligible car is more often labeled occluded when that frame contains more such cars, after adjustment for depth and sequence in a conventional descriptive model.

OmniDrive motivates evaluating generated road scenes that contain interacting objects across multiple views (Meng et al., 2026d). Crowding is relevant to that agenda, but the present study does not assess generated interactions or reproduce its model. It analyzes one observed relationship in public tracking labels. Its contribution is to define an annotation-based condition that a future test can sample, without assuming that a higher count is a causal intervention or that a model already handles the corresponding visual complexity.

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.

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.

2. Eligible-car counts and the fitted association

The source archive is filtered to Car annotations with nonnegative track identifiers, positive camera-depth coordinate, and occlusion codes zero, one, or two. This yields 26,805 rows in twenty sequences, linked to 574 sequence-specific tracks. Counting eligible rows within each sequence-frame pair produces 6,457 observed frames. Their median eligible-car count is three, and the maximum is sixteen. Frames without any eligible car are outside this analysis, so these counts do not describe the full frame population of the original recordings.

The outcome indicates any recorded occlusion, combining codes one and two. Ordinary least squares relates that binary variable to eligible cars in the frame, base-two log depth, and sequence indicators. The model uses sequence-clustered standard errors and the saved t-based confidence interval. The count coefficient is 0.0277, with a 95 percent interval from 0.0161 to 0.0394. Under this linear specification, one additional eligible car is associated with about 2.77 percentage points more occlusion probability at the stated adjustment boundary.

The regressor includes the focal car itself. Its interpretation is therefore a count of all eligible car observations in that frame, not a separately measured set of physical occluders. Other object classes, excluded annotations, road geometry, and unrecorded visibility conditions are not counted. Nor is the measure divided by road area, image area, or travel distance. Retaining this exact definition prevents an accessible metadata count from being presented as a more comprehensive measurement of traffic density than the archive processing provides.

Tracking evaluation distinguishes object observations from the continuity of their associations. HOTA develops a higher-order tracking assessment (Luiten et al., 2020), ByteTrack studies association using detection information including lower-confidence boxes (Zhang et al., 2021), and AB3DMOT supplies a baseline and evaluation setting for three-dimensional multi-object tracking (Weng et al., 2020). These sources motivate preserving track and frame structure, but they do not convert ground-truth label summaries into detector or tracker performance. A label archive has no newly measured false detections or identity switches unless predictions are actually compared with it. For generated driving video, a separate study could assess whether objects remain compatible across views and time. That would require defined object correspondences and an appropriate reference or consistency criterion. A descriptive visibility analysis can identify conditions worth testing, yet its transition counts remain transitions in annotations rather than observed failures of a tracking algorithm. The report should state the level at which its evidence exists and retain unavailable observations instead of assigning them an unmeasured prediction error.

Describing a contrast accurately can be more valuable than giving it an unjustified mechanistic name. A causal interpretation needs assumptions about how exposure and outcome arose, including selection into the analyzed sample. Causal diagrams provide a language for those assumptions (Greenland et al., 1999). If several alternative metrics or partitions are examined, the investigator should also account for the freedom to select whichever result appears strongest. False-discovery control is relevant to explicit collections of statistical tests (Benjamini and Hochberg, 1995), while a descriptive sensitivity analysis should be presented as such. A model can fit the observed rows well and still leave the principal explanation unresolved. Reporting that boundary preserves the value of the observation without overstating what adjustment or a significance threshold has established.

The phrase generalization needs an object: another observation of the same source, another source within the same collection, or a different acquisition process. Each requires a different assessment boundary. Structured validation makes explicit how correlated observations are assigned across training and

evaluation (Roberts et al., 2017). Selection bias can still arise if the final boundary is repeatedly consulted while choosing a method (Varma and Simon, 2006). Publishing the split identifiers and the information permitted during each stage allows the claimed transfer to be checked. When a study proposes a new architecture but performs only a data audit, it should state that difference directly. Useful evidence about data coverage does not require pretending that a prediction system has already been trained and externally validated.

3. Interpreting crowding without a causal shortcut

The positive association can reflect several correlated scene properties. More eligible cars may occur in settings with different lane layouts, object positions, recording conditions, or annotation visibility. Depth and sequence adjustment do not guarantee that those differences are removed. The model therefore does not estimate what would happen if one car were physically added to an otherwise unchanged scene. Establishing such an intervention effect would require a suitable design and additional evidence beyond the observed label table.

The frame count is repeated for every eligible car in the frame, while the outcome belongs to an object annotation. The regression consequently weights frames with more eligible cars more heavily through their larger number of rows. That is appropriate to an annotation-level estimand when stated explicitly. An equal-frame analysis would answer another question. Sequence clustering acknowledges shared recording histories, but twenty eligible sequences still provide a limited number of independent groups for broader inference about road populations.

A future generated-driving study could stratify tasks by a declared object-count requirement and assess object persistence, visibility, and cross-view compatibility. It would need actual generated clips and an independent correspondence procedure. A reference-label association cannot reveal whether a generator omits vehicles, creates incompatible overlaps, or produces useful perception training examples. The present result instead helps make the future test's condition and denominator explicit, while leaving those unmeasured capabilities to their own direct assessment rather than inferring them from the crowding coefficient.

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. From annotation evidence to generated scenes

The phrase world model should be connected to an observable capability. World Models (Ha and Schmidhuber, 2018) and Dreamer (Hafner et al., 2019) motivate predictive representations in settings that connect experience and behavior. CLIP (Radford et al., 2021) and IP-Adapter (Ye et al., 2023) show additional ways to represent or supply external content. Offline reinforcement learning (Levine et al., 2020) reminds the evaluator that fixed-data support constrains what can be inferred about unseen decisions. For driving generation, these perspectives motivate explicit stages: conditioning, scene production, consistency assessment, and any downstream use. A favorable observation at one stage is not automatic evidence for every later stage. A multi-view clip can be useful for visualization without having demonstrated a perception benefit or a safe control policy. The current annotation analysis should therefore retain its descriptive scope. A later task experiment can test the broader claim using independent real observations and declared training conditions, rather than assigning utility to synthetic scenes solely because their reference data contain realistic label patterns.

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.

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

Eligible-car count is positively associated with recorded occlusion after the available depth and sequence adjustment. The finding describes object annotations in observed frames and is narrower than a traffic-density effect. It does not establish causation or generated-scene performance. Its contribution is a reproducible crowding-related condition for later evaluation, with the object-level weighting, filtered count definition, and missing visual evidence kept explicit in the interpretation.

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.