Entry and Exit Visibility in a Public Vehicle-Tracking Archive

Gabriel Duncan (CL), Charles Harrison (PY), Justin Allen (AO)

Abstract

The first and last eligible observations of a KITTI car track have different visibility profiles. Among 569 tracks with at least two eligible annotations, occlusion appears in 77.7 percent of first observations and 36.4 percent of last observations. There are 255 occluded-to-visible endpoint changes and twenty in the reverse direction. These endpoints belong to the filtered record and should not be equated with physical scene entry and exit. The analysis motivates boundary-aware evaluation of generated trajectories without claiming an explanation for the observed asymmetry.

1. A recorded endpoint is an observation boundary

The beginning of a track can mean the first annotation available in a recording, the first eligible row after filtering, or an object's physical arrival in the camera's field of view. Those events need not coincide. The same applies to the last observation. This study compares visibility at the endpoints of eligible KITTI car histories while preserving that distinction. Its question is descriptive: how do the recorded first and last states differ under the stated eligibility rule?

OmniDrive motivates assessing objects through multi-view driving-video trajectories (Meng et al., 2026d). Endpoint behavior is relevant to such an assessment because a generated object should remain interpretable across its observed history. The present analysis does not generate entry or exit events, inspect camera views, or assess the focal world model. It examines a reference-label boundary. Its contribution is to show why later trajectory tests should identify which endpoint they observe rather than assign a physical event meaning to the first and last available rows automatically.

The scientific population of a driving result is defined by its records and collection process. KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2019), BDD100K (Yu et al., 2018), and the Waymo Open Dataset (Sun et al., 2019) support important but different perception and scene-analysis tasks. Their names should not be treated as interchangeable guarantees of sensor access, geographic coverage, or annotation semantics. Datasheets motivate explicit documentation of those boundaries (Gebru et al., 2021). A label-only secondary study should therefore state which archive, classes, visibility codes, and eligibility filters define its rows. It should also preserve sequence and track identifiers because repeated annotations share an acquisition history. Such a study can establish useful distributional facts without supplying unobserved model outputs. A proposed generated-driving evaluation can draw on those facts when selecting tests, but it must separately obtain synchronized views, valid assessment evidence, and generated samples. The distinction keeps a modest annotation result interpretable and prevents the breadth of a benchmark's reputation from substituting for the narrower evidence actually analyzed.

2. Building track endpoints from eligible rows

The source is the public KITTI tracking training-label archive. Car rows are retained when the track identifier is nonnegative, recorded depth is positive, and the occlusion code is zero, one, or two. Positive occlusion codes define any occlusion. Track identifiers are combined with sequence identifiers, and eligible rows are ordered by frame. The complete filtered set contains 574 tracks. Five have only one eligible annotation and are excluded from the endpoint comparison, leaving 569 tracks with distinct first and last observed positions in the record.

The first-observation occlusion fraction is 0.777, and the last-observation fraction is 0.364. The mean last-minus-first binary difference is -0.413, or about -41.3 percentage points. At the individual track level,

294 have the same binary endpoint state, 255 change from occluded to visible, and twenty change from visible to occluded. The same-state category combines tracks that are visible at both endpoints with tracks that are occluded at both; it does not mean that their entire intervening histories remain unchanged.

Each eligible track receives equal weight in these summaries. A long history and a short history therefore contribute one endpoint pair each. This differs from an annotation-weighted visibility average, which gives longer histories greater influence. The analysis does not require every intervening frame to remain eligible, nor does it infer a direct adjacent transition between the endpoints. The endpoint contrast summarizes two selected observations in each track's filtered history and should be read at that level.

A large collection of measurements does not guarantee a large collection of independent cases. Repeated observations can share unrecorded acquisition conditions or a common source, making row-level uncertainty misleading. White's covariance estimator addresses heteroskedasticity within a specified regression (White, 1980), while cluster methods address a declared form of grouped dependence (Cameron and Miller, 2015). Neither approach repairs a poorly defined outcome or establishes representativeness of the observed sample. A report should identify the number of groups, their relative sizes, and the reason for selecting the grouping level. When only a few independent groups are available, apparently precise numerical intervals warrant particular caution. This is a limitation of the evidence structure, rather than a reason to count correlated measurements as fresh observations.

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.

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 the asymmetry conservatively

The observed asymmetry could relate to recording boundaries, object paths, eligibility filtering, or annotation processes. The endpoint table cannot choose among those explanations. In particular, it does not prove that vehicles generally enter scenes occluded and leave visible. Such a physical claim would require a definition of entry and exit, appropriate visual evidence, and a design that separates those events from administrative or selection boundaries. The present study has processed labels only and preserves the narrower recorded-endpoint interpretation.

Repeated tracks share sequences, so 569 endpoint pairs are not 569 independently acquired road environments. The study reports descriptive fractions and counts without assigning an independence-based confidence interval. A broader inferential analysis would need a declared sampling target and a procedure reflecting the recording hierarchy. The results also apply to the eligible car subset, not every road user or uncertain occlusion label in the full archive. These scope conditions matter before the asymmetry is used to characterize driving data generally.

A generated-trajectory evaluation could deliberately distinguish instructed physical entry, first visible appearance, first assessable frame, and clip boundary. It could then assess identity persistence and scene compatibility across the relevant event. That protocol would require actual generated samples and an independent observation procedure. The present endpoint pattern helps motivate those separate records but supplies no generated performance result. Its usefulness lies in identifying a boundary that an overall score can conceal, not in turning the archive's observed asymmetry into an untested law of scene dynamics.

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.

A proposed explanation should be distinguishable from alternative explanations using observations the study can actually collect. Diagramming the relationships among selection, exposure, measurement, and outcome makes that requirement explicit (Greenland et al., 1999). The diagram does not establish its own correctness, but it prevents adjustment choices from remaining invisible. The same transparency is needed when investigators inspect many endpoints before choosing a narrative. Controlling false discovery concerns the behavior of a specified testing family (Benjamini and Hochberg, 1995); it is not a general certificate of scientific validity. Primary endpoints, secondary descriptions, and future mechanism tests should be named separately. Readers can then assess whether an attractive pattern was predicted, explored, or merely noticed after the available comparisons were examined.

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. From annotation evidence to generated scenes

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.

A scientific release should make its most consequential choices recoverable. Persistent identifiers and explicit metadata support reuse of observations and results (Wilkinson et al., 2016), while attention to dependencies helps prevent an evaluation from changing unnoticed as software evolves (Sculley et al., 2015). Relevant choices include the source version, preprocessing, accepted input set, metric implementation, and weighting convention. The record should also separate an observed output from a proposed future test. Saving that distinction prevents a plan from later being mistaken for a completed experiment. A small collection of well-linked records can be more useful than a large folder of unexplained scores. The important property is whether an independent reader can reconstruct what was measured and why its reported number has that meaning.

5. Conclusion

First and last eligible car observations have markedly different occlusion fractions in the released tracking archive. The result concerns filtered track endpoints, with equal track weighting and no claim that they coincide with physical entry or exit. It does not explain the asymmetry or evaluate a generator. Its contribution is a boundary-aware description that future driving-video studies can use when defining trajectory events, observation coverage, and the evidence needed for temporal scene claims.

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.