Track Length and Exposure Weighting in Driving Evaluation

Nicholas Fisher (AD), Oliver Wallace (PA), Bruce Phillips (FM)

Abstract

Equal-track and equal-annotation summaries give different visibility pictures in KITTI tracking data. The mean occlusion fraction is 0.612 when each eligible track receives equal weight, compared with 0.492 across annotations. Among 574 tracks, longer observed histories also have lower adjusted occlusion fractions. This secondary analysis makes exposure weighting explicit; it does not interpret observed length as a randomly assigned duration or measure tracking accuracy. The result motivates preserving trajectory structure when designing evaluations of generated driving scenes.

1. Longer tracks receive more weight in a row average

A vehicle observed in many frames contributes more annotations than one observed briefly. An annotation-weighted average therefore emphasizes long recorded histories. An equal-track average gives each observed vehicle trajectory the same weight, regardless of its number of eligible rows. Neither is automatically superior. They describe different populations. This study compares those summaries and examines whether observed track length is associated with visibility after a limited adjustment using the public KITTI tracking-label archive.

OmniDrive proposes multi-view driving generation through a coordinated world-model workflow (Meng et al., 2026d). Its setting makes object histories and their weighting relevant to evaluation. The present study does not track generated vehicles, estimate identity switches, or validate the focal model's outputs. It analyzes reference-label exposure. The purpose is to show how the scientific unit changes a visibility summary and to identify why preserving track structure matters before the many frames of a driving video are flattened into one score table.

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.

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.

2. Constructing observed tracks and their summaries

Eligible Car rows have a nonnegative track identifier, positive camera-depth coordinate, and occlusion code zero through two. Track identifiers are combined with sequence identifiers so that matching numbers in different recordings are not treated as one vehicle history. The filtered archive contains 26,805 annotations in 574 sequence-specific tracks across twenty sequences. Track length here means the number of eligible annotations. It is not assumed to equal uninterrupted physical time in the scene, because filtering and observation boundaries can affect the recorded history.

The median track has thirty-one eligible annotations, the maximum has 643, and five tracks have only one. Each track's occlusion fraction is the mean of its binary any-occlusion labels. Averaging these fractions equally across tracks gives 0.612. Averaging the binary labels across all eligible annotations gives 0.492. The difference is about 12.1 percentage points. It reflects a change in weighting population using the same underlying records, not a contradiction in the source labels or a newly observed change in vehicle visibility.

A track-level ordinary least squares model relates its occlusion fraction to base-two log annotation count, base-two log mean recorded depth, and sequence indicators. Standard errors are clustered by sequence with the saved t-based interval. The length coefficient is -0.0784, with a 95 percent interval from -0.1197 to -0.0371. Under this specification, doubling eligible annotation count is associated with about 7.84 percentage points less occlusion fraction. The model has 574 track observations, rather than one independent outcome per original annotation.

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.

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.

A visually appealing frame need not be compatible with the other frames in a road scene. SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) provide pairwise image measurements; FVD (Unterthiner et al., 2018) provides a video-distribution perspective; VBench (Huang et al., 2023) motivates multiple video-assessment dimensions. These tools are useful when their required inputs and scientific

interpretations are stated. They do not jointly guarantee physical or cross-view correctness merely because several favorable scores are reported. A driving-generation protocol should preserve its camera relationships, object events, temporal windows, and assessment coverage. Where no reference image exists, a reference-based measure may not answer the intended question. Where only labels are analyzed, the visual measurement remains unperformed. The present annotation evidence can motivate the protocol's conditions but cannot supply missing generated samples. Keeping those boundaries explicit allows later results to connect a numerical score to the specific scene property it actually observes rather than to an undefined overall notion of realism.

3. Exposure is not an intervention on visibility

Longer observed histories can differ from shorter ones in their scene paths, visibility, class-specific annotation opportunities, and sequence boundaries. Visibility can itself affect the record that becomes available. The adjusted coefficient therefore does not show that making a track longer would improve its visibility. It describes an association among observed track summaries. A causal duration claim would require a design that separates observation selection and scene processes, neither of which is supplied by this secondary table analysis.

The weighting contrast has a direct methodological implication even without a causal explanation. A generated-video evaluation that pools every scored frame implicitly gives longer clips or longer visible tracks more influence. If the intended claim concerns typical subjects or trajectories, another weighting rule may be appropriate. The rule should be declared before the final comparison and accompanied by the observation hierarchy. A sensitivity report can show how conclusions change without pretending that every weighting represents the same scientific population.

The archive's scope also limits transfer. These are eligible car tracks under recorded label conventions in twenty sequences, not every road user or all driving conditions. No images, detector predictions, or synthetic videos are processed. A future world-model study would need direct object correspondences, generated histories, and coverage records to assess persistence. The present evidence can inform those choices, but it cannot establish generated tracking quality merely because its reference-data weighting has been carefully described and statistically adjusted.

The protected assessment population should be defined before investigators inspect detailed failures. Once a collection of cases guides repeated revisions, it has become development information even if its filenames still carry an evaluation label. Cross-validation for dependent data treats the grouping of observations as a substantive design decision (Roberts et al., 2017). Research on model selection likewise shows why tuning and final error estimation should not reuse the same information (Varma and Simon, 2006). A practical record therefore distinguishes fitting data, diagnostic examples, validation data, and the final external assessment. Where only a fixed public table is available, a descriptive analysis remains legitimate, but it should not be renamed a held-out prediction experiment without generating and assessing predictions under a protected split.

A substitute measurement needs evidence about its relationship to the intended judgment. Two scores may correlate strongly while differing systematically in particular conditions. Method-comparison analysis distinguishes association from agreement (Bland and Altman, 1986). When a system additionally reports confidence, calibration asks whether those probabilities correspond to observed outcome frequencies (Guo et al., 2017). These are separate properties: a metric can rank examples usefully while having a distorted numerical scale or unreliable confidence. An evaluation should therefore retain the original target observations when possible, describe the scale of each endpoint, and examine disagreements

instead of averaging them away. A convenient automated score becomes more credible when its limitations are measured in the conditions that motivate its use.

4. From annotation evidence to generated 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.

A system's reliable operating region may be narrower than the collection of inputs an interface accepts. Selective prediction separates performance on accepted cases from the fraction of cases accepted (Geifman and El-Yaniv, 2017). Reporting only the former can hide a substantial loss of coverage. Uncertainty estimates may also change under distribution shift (Ovadia et al., 2019), so familiar-condition confidence is not automatically evidence about new subjects, scenes, or control histories. A useful evaluation records rejected, ambiguous, and unsupported cases alongside successful ones. This makes it possible to distinguish limited but honest capability from a system that answers broadly with unreliable certainty. Any proposed deferral rule still requires assessment on its intended input population; it is not validated merely by being described.

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

Equal-track weighting yields a substantially higher occlusion summary than equal-annotation weighting, and longer observed tracks have lower adjusted occlusion fractions in this archive. The findings are descriptive and depend on the eligible observation process. They do not measure a duration intervention or a tracker. Their contribution is to make exposure weighting visible before a generated-driving evaluation assigns scientific meaning to an average dominated by its longest recorded histories.

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.