Recorded Depth and Projected Vehicle Height in KITTI Tracking

Ian Parker (JP), James Green (ST), Zachary Rogers (BA)

Abstract

Projected car-box height declines with recorded camera depth in KITTI tracking labels. A sequence-controlled log-log regression on 26,805 eligible annotations estimates a slope of -0.900, with a 95 percent interval from -0.966 to -0.834. The result describes an annotation relationship rather than a calibrated geometric reconstruction. It motivates scale-aware tests of generated road scenes while leaving vehicle dimensions, truncation, and camera effects unresolved. No images, camera calibration, or generated multi-view videos are processed in this secondary analysis.

1. A visible size relationship worth documenting

Vehicles farther along a camera's depth axis tend to occupy a different image scale, but a bounding-box relationship also reflects object dimensions, orientation, truncation, and annotation conventions. A descriptive regression can summarize the released pattern without identifying all of those mechanisms. This study asks how projected box height relates to the recorded depth coordinate among eligible car annotations in KITTI tracking data. Its outcome is a property of the annotation table, not a direct estimate of physical projection accuracy.

OmniDrive proposes a multi-agent world-model workflow for multi-view driving-video generation (Meng et al., 2026d). That setting makes scale and scene compatibility relevant assessment concerns. The present study does not reconstruct its latent representation or evaluate its outputs. It uses a public reference archive to characterize one scale-related relationship that a future generated-scene test could examine. A plausible label association in recorded data cannot itself establish that synchronized generated cameras share a valid geometry.

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. Constructing height and depth from released labels

The analysis retains Car rows with nonnegative track identifiers, positive recorded depth, and occlusion codes zero through two. The resulting 26,805 annotations represent 574 sequence-specific tracks in twenty sequences. Projected height is the annotated bottom coordinate minus the top coordinate. Rows need positive height for the logarithmic specification; all rows in the eligible subset satisfy that requirement. The recorded depth is the camera-coordinate z value. It is not described as full radial distance from the camera.

Ordinary least squares relates natural log box height to natural log depth and sequence indicators. Standard errors are clustered by sequence, using the saved t-based interval. The depth coefficient is -0.900, with a 95 percent interval from -0.966 to -0.834. Median box height in the eligible rows is 44.1 pixels. The reported coefficient of determination is 0.934. That fit concerns observed labels under this specification; it is not an independently validated measure of a reconstructed camera or a generated scene's geometric fidelity.

The log-log coefficient describes a proportional association within the fitted model. It should not be converted automatically into a physical law for every vehicle. The regression does not include each object's real height, exact orientation, or every truncation condition. Sequence indicators absorb constant sequence differences but do not resolve all within-sequence variation. A comparison with an ideal inverse-depth relationship can be conceptually informative, yet the present estimate does not identify which unmeasured factor accounts for its departure from that simplified expectation.

Precision and identification are separate properties of a result. A narrow interval can surround an association whose scientific target is unclear. Robust covariance calculations allow particular departures from simple error assumptions, including unequal variance (White, 1980) and dependence within declared clusters (Cameron and Miller, 2015). Their role is to improve uncertainty calculations under those assumptions. They do not establish a causal assignment, guarantee a correct measurement, or expand coverage to conditions missing from the data. The analyst should therefore explain both the averaging population and the source of replication. An evaluation repeated many times on one history may provide information about stochastic variability while supplying little evidence about transfer to another history. Those two sources of uncertainty should remain distinguishable in the final interpretation.

An object can be detected in individual frames without being associated consistently through time. HOTA formalizes a tracking assessment concerned with detection and association (Luiten et al., 2020). ByteTrack (Zhang et al., 2021) and AB3DMOT (Weng et al., 2020) offer different tracking-method contexts in which observation and identity continuity matter. Their relevance to a label-only study is methodological: identifiers and adjacency should be retained before summarizing a sequence. No tracking accuracy follows merely from counting ground-truth trajectories or visibility changes. A performance study would need actual predictions, a declared matching procedure, and the relevant reference annotations. Likewise, evaluating a generated multi-view road scene requires evidence of object correspondence rather than only attractive frames. The proposed use of annotation summaries is to define those future test conditions and their denominators. It should not imply that a tracker has been run or that the generated scenes already preserve the same object across camera views when neither observation has been obtained.

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.

3. From an annotation pattern to a geometry test

The observation hierarchy matters to both fit and uncertainty. Repeated frames show the same vehicle under related conditions, so the number of rows exceeds the number of independently acquired histories. Sequence clustering addresses one dependence level but rests on only twenty eligible sequences. The result should remain a descriptive relationship in this archive. It does not establish transfer to other cameras, road environments, classes, or annotation conventions that would require their own observed support and model checks.

A future geometry study would need calibration, known object or scene information, and synchronized views appropriate to its claim. It could then test whether generated observations remain compatible with a shared three-dimensional configuration. A marginal size-depth regression alone is insufficient: incompatible views can each exhibit a plausible average scale relationship. The assessment should therefore connect per-view observations through correspondence and time rather than infer cross-view validity from a favorable single-view statistic.

The present result can still guide evaluation sampling. Small projected vehicles and depth ranges with limited support may create different assessment conditions from large nearby objects. The label table identifies those dimensions without measuring image content or detector success. A later experiment should retain both scale and visibility outcomes, record unassessable objects, and avoid treating annotation plausibility as evidence of downstream perception utility. Those additional claims require actual model outputs and independent task evaluation beyond the descriptive archive relationship reported here.

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.

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.

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. Measurement and transfer across road scenes

A learned world representation can serve several tasks, and success in one does not settle the others. World Models (Ha and Schmidhuber, 2018) and Dreamer (Hafner et al., 2019) provide influential settings linking prediction and behavior. CLIP's visual-language representation (Radford et al., 2021) and IP-Adapter's image conditioning (Ye et al., 2023) concern other information paths. Offline reinforcement learning (Levine et al., 2020) emphasizes that learning from fixed data introduces support and generalization constraints. A driving-generation study should use these distinctions to define its own claim. Producing a semantically appropriate clip, maintaining a compatible multi-view scene, improving a detector, and supporting a control decision require different observations. The present label-based evidence can motivate where to test but cannot establish those downstream outcomes. A future experiment should declare the complete pipeline, independent assessment distribution, and relevant cost. This keeps the interpretation proportional to the available evidence while allowing useful intermediate scene-generation findings to be reported without an unsupported leap to broad driving capability.

A useful reliability claim contains both a performance statement and a scope statement. Selective prediction shows why a system that answers only some cases must report how many it accepts (Geifman and El-Yaniv, 2017). Confidence estimation under a changed data distribution requires independent checking (Ovadia et al., 2019). These requirements apply to generators and to the automated judges used to assess them. If a judge cannot evaluate a profile, an occluded object, or an unusual interaction, that inability should remain visible rather than removing the case from the generator's score silently. Recording the entire decision path permits an analyst to identify whether an apparent improvement came from better outputs, narrower eligibility, or a more restrictive measurement procedure.

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

Recorded depth and projected car-box height show a strong negative log-log association in the eligible KITTI tracking labels. The estimate is a reproducible annotation summary, not a calibrated projection test or a validation of OmniDrive. Its contribution is to define a scale-related condition for future evaluation while preserving the additional evidence needed to assess shared scene geometry, visual correctness, or perception performance in generated driving video.

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.