Cross-View Calibration as an External Test of a Driving World Model

Kyle Thompson (GE), Mason Baker (ME), Mason Hayes (SB)

Abstract

Individually plausible camera views can still depict an incompatible road scene. An external calibration-based assessment would test whether generated observations support a common geometric account rather than rely solely on internal consistency scores. This article specifies the evidence, correspondences, and uncertainty such a test would require. OmniDrive provides the motivating multi-view generation setting. The proposed study has not been implemented. Its contribution is a separation between visual plausibility, compatibility under declared camera relationships, and broader claims about the physical correctness of a generated world.

1. A shared scene is an external claim

A collection of camera images can look convincing one image at a time while placing the same vehicle differently across views. A multi-view world-model claim therefore needs a relationship among observations, not only favorable per-view quality. Calibration offers one possible external structure for that relationship. Its use must be defined carefully: a known camera arrangement supplies constraints, but applying those constraints still requires valid correspondences and a clear account of what uncertainty is introduced by measurement.

OmniDrive proposes coordinated generation through a language-model director, layout-related scene organization, auditing, and latent co-compression (Meng et al., 2026d). These components motivate a test of whether the resulting views can be interpreted as one compatible scene. The present article does not run that architecture or reproduce its reported results. It proposes an independent assessment whose criteria are not supplied solely by the generator's own latent representation or agent critiques. Such independence can make a consistency claim more interpretable without assuming that internal checks have no useful role.

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.

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. What a calibration-based protocol would need

A proposed study would declare the camera relationships available to the generator and the evaluator. Known calibration, estimated calibration, and calibration inferred from generated content provide different evidence. If parameters are estimated, their method and uncertainty should be retained. A favorable fit after freely adjusting every camera can answer a weaker question than compatibility under a fixed intended arrangement. The protocol should therefore state which quantities are given, which are estimated, and which are allowed to vary across clips.

Object or feature correspondences would need an independently specified procedure. Incorrect matches can create apparent geometric failure, while selective matching can hide incompatible objects. The study should retain unmatched and uncertain cases as coverage outcomes. It should also define whether the target concerns static scene structure, moving objects, or both. Synchronization becomes essential for the latter: observations at different times can legitimately place a vehicle differently even when the scene is physically compatible.

The assessment would use declared residuals or compatibility criteria with a development stage for operational thresholds. Those thresholds should reflect an intended consequence and measurement quality rather than be selected to maximize a preferred method difference. A reserve assessment set could then support the principal comparison. Independent visual or geometric review of disputed cases would help distinguish generated inconsistency from failures in calibration estimation, correspondence, or time alignment. No such measurements or reviews are conducted in the present conceptual study.

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.

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 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.

3. Interpreting a successful external check

Passing a selected geometric test would establish a bounded compatibility result under its assumptions. It would not prove every physical property of the generated environment. A scene can satisfy a sparse set of correspondences while containing incompatible unmeasured objects. The report should therefore identify coverage, tested object classes, visibility states, and temporal intervals. A favorable result on easily matched regions should not be described as whole-scene validation when the difficult portions remain absent from the assessment.

Conversely, a failed check needs a diagnostic path. The source can be the generator, the measurement pipeline, or a mismatch between the declared task and supplied evidence. Repeating a test under improved correspondence or calibration can be informative, but the revised procedure should receive a version record. Otherwise a change in assessment may look like a change in generator quality. The original observations and reasons for revision should remain available so that the meaning of the result can be reconstructed.

Comparisons among generators should preserve camera conditions, instructions, reference access, output duration, and generation effort. A method allowed additional scene information faces another task. Resource and selection differences can be part of an end-to-end comparison, provided they are explicit. The useful final claim is compatibility under declared constraints and observed coverage. Downstream perception benefit or safe control would require separate task evidence rather than being inferred automatically from passing an external geometric check.

Visual plausibility is only one dimension of a generated driving scene. Structural similarity (Wang et al., 2004) and learned perceptual similarity (Zhang et al., 2018) compare particular image properties, while video-distribution evaluation (Unterthiner et al., 2018) and VBench's multidimensional framework (Huang et al., 2023) address other aspects of generated output. None alone establishes that synchronized camera views depict one compatible scene. A method can produce sharp, individually credible frames with inconsistent object placement across views. It can also preserve geometry while showing visible temporal defects. The evaluation should retain these outcomes separately and connect each instrument to the reference information it requires. A label-only analysis cannot calculate these visual measures because it has not processed the necessary images or clips. Its role is to describe annotation conditions that may deserve targeted generation tests. A subsequent experiment would need actual outputs, declared sampling, synchronized assessment, and an uncertainty policy. Those measurements should be reported as new evidence rather than inferred from favorable distributions in the reference label archive.

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. Evaluation design beyond the observed archive

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 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.

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

Cross-view calibration can provide an external evaluation structure when camera relationships, correspondences, synchronization, and uncertainty are explicit. The proposed protocol reports no new geometry result. It connects a shared-scene claim to measurable constraints while retaining unassessed regions and alternative sources of failure. A future successful test would support bounded multi-view compatibility, leaving broader physical validity and downstream utility to the additional observations those claims require.

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.