Semantic Compliance and Physical Compatibility in Generated Road Scenes

Stephen Young (SM), Nathaniel Anderson (SC), Justin Graham (AM)

Abstract

A driving video can contain every requested object while depicting a scene that is incompatible across cameras or time. It can also be physically plausible while failing the instruction. This article develops a joint assessment design that preserves both outcomes, motivated by OmniDrive. No generated scenes are evaluated. The design links semantic requirements to observable content and physical compatibility to declared geometric and temporal constraints. It explains why neither a fluent scene description nor an attractive video establishes the other dimension of a driving-world claim.

1. Two kinds of success can disagree

An instruction might require a car to turn behind a bus while a pedestrian waits near a crossing. A video can include those semantic elements yet place the objects incompatibly across views. Another video can depict a coherent road scene while omitting the turn or changing the pedestrian's role. The two failures concern different requirements. An evaluation that collapses them into one unexplained realism score risks obscuring which capability the generated output actually demonstrates.

OmniDrive combines structured scene choreography, multi-agent roles, and latent co-compression for driving-video generation (Meng et al., 2026d). That architecture motivates checking both instruction realization and scene compatibility. The present article proposes the assessment without running the model or verifying its claimed results. It does not assume that an internal auditor supplies an independent physical test. The question is how external observations could establish the two dimensions separately and then describe their coexistence under a declared task and information boundary.

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.

2. Defining independent evidence for each dimension

A proposed semantic assessment would preserve the original request, structured scene specification, and any revisions. Requirements would identify objects, actions, relationships, and temporal events that can be assessed under the available camera coverage. The protocol would distinguish well-formed instructions from realized content. An independent assessor would check the video rather than infer execution from the presence of a script field. Ambiguous or unobservable requirements would receive explicit status instead of being counted automatically as successful.

The compatibility assessment would declare the constraints it can test. These might concern cross-view correspondence, temporal object persistence, or consistency under known camera relationships. Each requires appropriate evidence and measurement uncertainty. A sparse geometric check does not prove all physical laws, and a plausible action description does not establish valid dynamics. The study should therefore name the tested property rather than use physical correctness as an unrestricted label for whatever a convenient instrument happens to measure.

Both paths would preserve their input access and validation. If an assessor uses the generator's own representation or critique, shared sensitivities should be examined through an independent endpoint. Human review can contribute where the task is observable, but reviewers need adequate views, synchronization, instructions, and uncertainty responses. A judgment about visual appeal should remain distinct from a judgment about geometric or semantic compatibility. The final report should retain those distinctions before considering a combined application-specific success criterion.

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.

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.

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. Reporting coexistence rather than substituting one score

A joint outcome table could show clips that satisfy both dimensions, only one, neither, or remain partly unassessable. This preserves information that separate marginal averages can hide. Two methods can have similar semantic and compatibility means while differing in whether both succeed on the same clips. The joint criterion should be declared before final assessment, with explicit treatment of uncertain cases. Otherwise the definition of overall success can drift toward whichever combination favors a preferred generator.

Sampling and effort also matter. A selected clip found after many attempts represents a different workflow from an unselected first generation. The study should preserve candidates, revisions, and stopping rules. Repeated views and frames of the same scene remain correlated observations. Uncertainty should reflect the relevant scene and task hierarchy rather than treating every requirement check as an independent new environment. This makes a joint success rate interpretable in relation to the actual breadth of tested conditions.

The resulting claim should remain bounded to the instructions, constraints, and coverage observed. Passing semantic and selected compatibility checks would be useful evidence of a defined generation capability. It would not establish downstream perception benefit, real-world frequency of errors, or safe control. Those are further tasks with their own measurements. The proposed design allows a meaningful intermediate result while resisting an unsupported leap from a plausible generated road scene to every practical consequence associated with the phrase world model.

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.

An association can suggest a research direction without identifying an intervention. Causal diagrams make it possible to state which variables precede the proposed change, which are consequences, and which selection processes remain uncontrolled (Greenland et al., 1999). Their value lies in exposing assumptions rather than decorating a regression with a causal label. A second issue arises when many outcomes or subgroups are explored: selecting the most favorable comparison can exaggerate the impression of support. False-discovery procedures address a defined family of inferential tests (Benjamini and Hochberg, 1995). They do not turn an exploratory pattern into a replicated mechanism. A careful report distinguishes prespecified comparisons, descriptive sensitivities, and hypotheses requiring new observations. This distinction is especially important when an evaluation offers many plausible metrics.

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. Measurement and transfer across road 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.

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.

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

Semantic compliance and physical compatibility need separate evidence and a declared joint interpretation. The proposed design reports no new generation result. It preserves requirements, geometric and temporal constraints, uncertainty, coverage, and effort. A future assessment can then show where both capabilities coexist in the same outputs, while keeping attractive appearance, internal descriptions, downstream utility, and broad safety claims distinct from the specific scene properties that were actually observed.

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.