Checking Whether a WorldScript Is Obeyed in Generated Driving Video

Adam Lewis (PE), Gregory Lawson (BJ), Joseph Franklin (CI)

Abstract

A structured scene instruction is useful only if its intended objects, events, and relationships appear in the generated output. This article proposes an independent compliance audit for WorldScript-style driving generation. It distinguishes a well-formed script from a realized scene and keeps semantic execution separate from visual and geometric quality. OmniDrive supplies the motivating architecture. No scripts or videos are newly evaluated. The design specifies how requirements, assessment evidence, ambiguous cases, and generation attempts could be linked in a reproducible compliance study.

1. A valid instruction is not an observed outcome

A structured script can express objects, locations, actions, and temporal relationships more explicitly than an unconstrained prompt. That structure makes requirements easier to inspect, but it does not establish that a generator fulfills them. The output needs its own evidence. A script parser can confirm that a field exists while the corresponding object is missing, misplaced, or temporally inconsistent in the video. Evaluation should therefore distinguish instruction validity from execution of the instruction in visible content.

OmniDrive describes a director-generated WorldScript within a multi-agent driving-generation workflow (Meng et al., 2026d). Its structured scene organization motivates an audit that connects script requirements to outputs. The present article proposes that audit without generating scripts or assessing the focal system. It does not assume that every proposed requirement maps to a simple detector label. Its question is how a future study could verify compliance independently, especially when an instruction involves relationships across cameras or events unfolding through time.

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.

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.

2. Turning requirements into assessable records

A proposed protocol would preserve the original request, structured script, any revisions, model version, and generated attempts. Each requirement would receive a declared interpretation and assessment rule before final comparison. Object presence, relative position, action timing, and cross-view identity may require different instruments. The record should retain those distinct endpoints rather than reduce every requirement to one broad semantic score whose sensitivity to the intended condition has not been checked.

Ambiguous or unsupported requirements should be identified explicitly. Some may lack an adequate visual test, while others may be impossible under the requested camera coverage. A script can still be useful, but its untestable portion should not become an assumed success. Independent reviewers or instruments would assess realized content without relying solely on the generator's own declaration of compliance. Where an automated tool is used, its validity and coverage in the generated population need separate evidence.

The study would also distinguish script generation from video generation. If a director changes the user's request to make it easier, a compliant output under the revised script may still fail the original intent. Preserving both versions makes that divergence visible. A controlled experiment could hold scripts fixed to examine the video stage, while an end-to-end study could assess the complete request-to-output workflow. These are different contrasts and should retain their respective information and revision opportunities.

The choice of statistical unit should precede the choice of a standard error. A dataset can contain many rows but relatively few independently acquired histories. Heteroskedasticity-consistent covariance protects an otherwise specified regression against unequal residual variances (White, 1980). It does not automatically accommodate observations sharing an environment, subject, or recording. Cluster-robust inference makes that dependence part of the variance calculation (Cameron and Miller, 2015). The grouping must follow the observation process and the claim being made; an identifier that is convenient to store is not necessarily the relevant independent unit. For a descriptive analysis, preserving the groups also allows readers to recalculate a different population average. An uncertainty interval remains conditional on the observations and assumptions actually available.

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. Compliance across views, time, and attempts

A requirement can be satisfied in one camera and violated in another. Likewise, an object can appear at the correct time once but fail to persist through the declared event. A proposed compliance table would therefore retain view and temporal localization where the requirement demands them. A clip-level average can be reported after those records exist, but its aggregation rule should be explicit. Otherwise a short or spatially concentrated violation may disappear in a favorable overall score.

Repeated attempts introduce another distinction. A workflow that searches until one compliant clip appears has a different reliability and effort profile from one that succeeds on its first request. Both can be evaluated, but the attempt budget and selection rule must be visible. The test should preserve failed generations and unassessable requirements. Reporting only requirements from accepted outputs would condition the result on prior success and obscure the practical cost of obtaining the displayed compliant scene.

Finally, semantic compliance is not complete physical or downstream validity. A video can contain the requested vehicles and actions while depicting incompatible geometry. It can also be geometrically plausible while failing an instruction. A useful evaluation would report these dimensions separately and connect any combined utility to an application preference. The proposed audit supplies a way to verify what was asked and what appeared, leaving perception improvement, control utility, and broad physical correctness to additional direct experiments rather than assigning them from a successful script check.

An evaluation protocol should state what remains unfamiliar at assessment time. Leaving out an observation while retaining its close relatives can test interpolation rather than the broader transfer suggested by the study's wording. Validation methods for structured data emphasize matching the split to temporal, spatial, or hierarchical dependence (Roberts et al., 2017). Separately, parameter selection can bias an evaluation if the same observations guide both choices and reporting (Varma and Simon, 2006). These issues are not solved by calling a split a test set. The record should identify shared sources, the information accessible during development, and the stage at which each design choice was fixed. A result can be informative within a narrow boundary while providing no direct evidence about a more distant deployment population.

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. 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.

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.

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

A WorldScript compliance audit should link original requests, script revisions, requirements, generated attempts, and independent output evidence. The proposed design reports no new compliance rate. It distinguishes instruction structure from execution and preserves view, time, ambiguity, and effort. A future result can then establish which declared requirements were fulfilled under the observed workflow without allowing a well-formed script or an internal assertion to substitute for verification of the generated road scene.

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.