Counterfactual Traffic Interventions for Generative Driving Models

Lucas Green (MY), James Harrison (FM), Adrian Hall (TD)

Abstract

Changing one traffic element in a generated scene can alter many others. A useful intervention test must therefore verify both the intended change and the properties meant to remain stable. This article develops a counterfactual evaluation design for driving-world generation, motivated by OmniDrive. It distinguishes a controlled request to a model from a causal intervention in the real world. No traffic scenes are generated here. The proposed protocol records scene identity, intervention compliance, spillovers, and assessment uncertainty before interpreting a paired output difference.

1. A changed instruction is only the beginning of an intervention

A request to add a vehicle, move a pedestrian, or change a signal can produce broad scene changes. The generator may alter camera motion, road layout, or unrelated objects along with the intended element. Comparing two resulting clips therefore does not automatically isolate the requested factor. A counterfactual-style evaluation should specify the intervention and the preservation boundary, then verify both in the outputs. The term counterfactual should not make an uncontrolled pair appear causally clean merely because its prompts differ by one phrase.

OmniDrive proposes structured scene choreography through multiple agents and a WorldScript-oriented workflow (Meng et al., 2026d). That organization motivates testing whether a declared scene change can be made while other requirements remain compatible. The present article proposes such a test without running the architecture or measuring intervention fidelity. It treats the generated comparison as a model behavior study. Any claim about real traffic consequences would require additional evidence linking the model's changes to an appropriate real-world causal setting.

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. Defining the intended change and protected properties

A proposed experiment would preserve a base scene specification, an intervention specification, model versions, sampling settings, and all generated attempts. The intervention would identify the object, attribute, relationship, or event to change. Protected properties would identify what should remain stable, allowing for consequences explicitly required by the task. A binary image mask alone may be insufficient because an added object can legitimately affect visibility or shadows. The protocol should describe intended semantic consequences rather than demand identical pixels everywhere else.

Paired generation could use shared settings or random-state information where available, but a shared seed does not prove that every unintended source of variation is held fixed. The study should assess actual output changes. It could distinguish compliance with the intervention from preservation of road layout, other object identities, camera relationships, and temporal structure. Uncertain or unobservable properties should remain in the coverage record rather than becoming assumed invariants because their specification was copied between requests.

The intervention set would be selected before final outcome inspection and should include conditions relevant to the intended use. A deliberately challenging test can be valuable, but it does not estimate the natural frequency of those conditions in driving. A practical workflow may also include iterative correction, whose effort and stopping rule should be reported. A controlled single-attempt comparison and a user-assisted correction process provide different evidence and should not share an unexplained reliability denominator.

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.

Temporal identity is an assessment target distinct from per-frame appearance. HOTA (Luiten et al., 2020) provides a tracking-evaluation foundation, while ByteTrack (Zhang et al., 2021) and AB3DMOT (Weng et al., 2020) illustrate practical association settings with different input structures. A study drawing on this literature should still distinguish reference labels from algorithm outputs. Repeated ground-truth annotations can describe track lengths, visibility changes, and the observation population, but they cannot establish detection recall or association accuracy in the absence of predictions. For a driving generator, a separate protocol would need to define how an object is identified across views and time and what counts as a consequential inconsistency. A pair of plausible camera images is insufficient evidence if their objects cannot belong to one scene. Conversely, a difficult visibility interval can make assessment uncertain without proving an error. Preserving the observation hierarchy and assessment coverage connects a useful annotation result to such a protocol without claiming that the protocol has already produced a performance finding.

The phrase generalization needs an object: another observation of the same source, another source within the same collection, or a different acquisition process. Each requires a different assessment boundary. Structured validation makes explicit how correlated observations are assigned across training and evaluation (Roberts et al., 2017). Selection bias can still arise if the final boundary is repeatedly consulted

while choosing a method (Varma and Simon, 2006). Publishing the split identifiers and the information permitted during each stage allows the claimed transfer to be checked. When a study proposes a new architecture but performs only a data audit, it should state that difference directly. Useful evidence about data coverage does not require pretending that a prediction system has already been trained and externally validated.

3. Interpreting paired changes in generated scenes

The analysis would first verify that the requested intervention occurred. A difference in downstream score without intervention compliance does not answer the intended question. It would then examine protected properties and permitted consequences. These dimensions should remain separate so that a successful object addition does not conceal a changed road or incompatible camera view. A combined utility can be defined for an application, but its weighting should not erase the component observations needed to understand the tradeoff.

Repeated scenes and samples should retain their hierarchy. Several interventions on one base scene are not equivalent to several independently sampled environments. Uncertainty should reflect the intended generalization unit and the dependence introduced by pairing. The study should also retain all failed or unassessable attempts. Selecting only visually plausible pairs after generation can make the intervention appear more controlled than the complete workflow actually is, particularly if difficult spillovers are disproportionately excluded from the displayed examples.

Finally, synthetic interventions should be interpreted within the model's tested support. A plausible change in a clip does not establish the physical response of a real road system or a valid safety estimate. A later task study could assess whether such outputs improve a declared downstream model on independent real data. That would be new evidence with its own design. The present proposal instead defines a bounded capability: controlled, assessable scene modification under a documented preservation boundary and generation budget.

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.

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.

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

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.

Difficult cases can disappear at several stages: data eligibility, generation, measurement, and final scoring. Each exclusion changes the population represented by the reported result. Coverage-aware evaluation makes those decisions visible (Geifman and El-Yaniv, 2017). A confidence score should also be checked when the input distribution changes, because uncertainty under shift is an empirical property rather than a guarantee of the model interface (Ovadia et al., 2019). An informative record preserves failures to produce an output and failures to obtain a valid measurement. Treating either as absent data without explanation can inflate apparent reliability. A proposed method should state its unsupported conditions and test the boundary between accepted and rejected cases as carefully as the quality of its successful outputs.

Reproducibility begins with preserving which scientific objects are being linked. Stable identifiers and reusable metadata help connect inputs, reference observations, model versions, and reported outcomes (Wilkinson et al., 2016). The scoring pipeline also has dependencies that can change the interpretation of a result: preprocessing, representation checkpoints, thresholds, and aggregation code can all matter (Sculley et al., 2015). A compact release should retain those decisions rather than only the final mean. Where source data cannot be redistributed, an access description and hashes can still identify the analyzed inputs. This record allows later users to distinguish a change in the underlying system from a change in how the system was measured. Reproducibility is therefore part of interpretation, not simply a final packaging step.

5. Conclusion

Counterfactual-style driving tests need verified intervention compliance and explicit protected properties. The proposed design records scene versions, spillovers, attempts, and uncertainty, while distinguishing model behavior from real-world causation. It reports no new generation results. Its contribution is an interpretable paired comparison in which a requested traffic change becomes an observed, bounded capability rather than an assumed isolated intervention inferred solely from a small difference in the input script.

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.