Uncertainty Budgets for Comparing Multi-View Driving Generators

Adrian Roberts (MM), Cameron Hall (TZ), Jack Porter (VN)

Abstract

Uncertainty in a driving-generator comparison can arise from scene sampling, stochastic generation, annotation, automated measurement, and evaluator choice. Repeating one stage cannot resolve all the others. This article proposes an uncertainty budget that links each source to a corresponding observation or sensitivity analysis. OmniDrive provides the motivating multi-view setting. No new uncertainty estimates are calculated. The design helps allocate evaluation effort and prevents a narrow interval over many correlated frames from being interpreted as broad confidence about new scenes, measurement validity, or downstream driving utility.

1. One interval cannot summarize every unresolved question

A reported confidence interval usually reflects a particular sampling model. It does not automatically incorporate uncertain labels, unobserved scene conditions, evaluator bias, or changes in the task definition. A driving-video comparison can therefore be numerically precise while remaining scientifically limited. The present article proposes an uncertainty budget: a record of the main sources of variation and ambiguity, the evidence addressing each, and the portions that remain outside the reported statistical calculation.

OmniDrive proposes multi-agent coordination and a shared latent approach for multi-view driving-video generation (Meng et al., 2026d). Its setting creates several linked observation levels, including requests, scenes, camera views, clips, and frames. The present article does not estimate uncertainty for the focal model or rerun its benchmark. It develops a comparison design. The goal is to connect a future interval to the evidence it represents while keeping other unresolved questions visible rather than allowing one favorable numerical summary to imply certainty about the complete world-model claim.

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.

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.

2. Mapping uncertainty to observable replication

A proposed study would identify its primary estimand before allocating evaluation effort. Scene sampling supports one type of generalization; repeated generation draws for a fixed scene support another. Multiple camera views improve scene evidence but do not create independent environments. Repeated measurements of one clip can reveal evaluator instability while saying little about new subjects or roads. The record should preserve these levels so that the statistical procedure does not count repeated frames as independent scene replication.

Annotation uncertainty would receive a separate record. Independent judgments, disagreement, and unassessable cases can describe the stability of an endpoint under the chosen evidence presentation. Automated measurements would need their own validation population and coverage checks. If a score fails under difficult visibility, removing those cases changes the population represented by the final result. A comparison should report coverage alongside conditional performance rather than treating missing measurements as an issue already resolved by a standard error on the retained rows.

Generation and workflow variation would include sampling settings, model versions, reference access, revisions, and selection effort. A best-of-many result and a first-attempt result have different uncertainty and cost interpretations. The study should declare which workflow it measures and retain all attempts. Where practical conditions vary, a planned sensitivity analysis can show their consequences. It should not be described as probabilistic coverage over every unknown factor unless the design actually supplies a justified sampling model for that broader statement.

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.

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.

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.

3. Allocating effort to the unresolved source

An uncertainty budget can guide collection. If the main unresolved question is scene transfer, more independent scenes may be more useful than many additional draws of one scene. If the assessor is unstable, repeated measurement or a better reference protocol may be the appropriate next step. If the benchmark omits a condition entirely, narrower intervals on common cases will not fill the gap. These decisions should follow the intended claim and observed evidence structure rather than the convenience of producing more rows.

Comparisons between methods can use paired tasks where appropriate, preserving common scene and reference conditions. Pairing can improve precision for the defined contrast, but it does not repair invalid outcomes or unequal information access. The report should state which uncertainty components are represented in its primary interval and which are explored through separate sensitivity summaries. A table of unresolved boundaries can be useful when it names concrete missing observations rather than providing a generic list of hypothetical risks disconnected from the study.

The final interpretation should distinguish precision, validity, coverage, and transport. A narrow interval can support a stable difference on the observed benchmark while leaving a new road population untested. Agreement among evaluators can coexist with a shared blind spot. A visually plausible output can remain unassessed for downstream utility. The proposed design keeps these claims separate while allowing each positive result to contribute. Its purpose is not to demand impossible certainty, but to make the next informative measurement and the present conclusion equally clear.

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.

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. From annotation evidence to generated scenes

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.

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.

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

An uncertainty budget connects each unresolved question to the observation level capable of addressing it. The proposed framework reports no new generator comparison. It preserves scene replication, generation variation, annotation, measurement coverage, and workflow effort. A future result can then state what its interval actually supports and where additional evidence is needed, avoiding the assumption that repeated scoring of many related frames establishes broad confidence about every aspect of a generated driving world.

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.