Grounding Multi-Agent Critiques in Observable Scene Evidence

Derek Wilson (SE), Vincent Allen (LY), Luke Johnston (BY)

Abstract

An agent critique can sound precise while identifying no independently verifiable problem in a generated scene. This article proposes assessing critiques through localized claims, supporting evidence, and observed consequences of revision. OmniDrive motivates a driving workflow with an auditing role. No critiques are generated or rated here. The design separates persuasive explanations from diagnostic accuracy and separates a correct diagnosis from a useful correction. It treats the critique itself as an object requiring evaluation rather than as automatic validation of the scene it discusses.

1. A critique makes a claim about an observation

A scene auditor may describe a misplaced vehicle, incompatible views, or a missing instruction. Each statement should point to evidence in the generated output. Fluent language alone does not establish that the described problem exists. A useful evaluation therefore asks what the critique claims, where the relevant content occurs, and whether an independent assessor can verify it under the same information boundary. Without that link, an explanation can become another untested output layered on top of the original generation.

OmniDrive includes an auditing role in its multi-agent driving-generation framework (Meng et al., 2026d). That design motivates a study of critique validity and correction value. The present article proposes such a study without running its agents or assessing its outputs. It does not infer diagnostic accuracy from the architectural presence of an auditor. Instead, it asks how a future comparison could determine whether a critique identifies a real, consequential scene problem and whether the subsequent workflow actually improves the intended outcome.

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.

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. Building an independently assessable critique record

A proposed record would preserve the scene version, agent inputs, critique text, localized claims, and any requested revision. Claims should be separated where they concern different objects, cameras, times, or requirement types. A statement about cross-view geometry needs different supporting evidence from a statement about prompt compliance. The evaluator should retain claims that cannot be checked with the available data rather than treat them as either confirmed diagnoses or harmless narrative excluded from the denominator.

Independent assessment would examine whether the claimed issue is present and whether the localization is adequate. Reviewers should not be shown a later successful revision before judging the original diagnosis, because that could influence their interpretation. The procedure should permit uncertainty and distinguish an incorrect claim from an unassessable one. Automated checking can assist where valid, but using the same representation for critique and final assessment can create shared blind spots that deserve an independent endpoint.

A control condition could compare agent critique with a declared alternative, such as no critique or a fixed review procedure. The information and revision budgets should be matched or described. A complete workflow comparison may allow different review effort, but its advantage then belongs to that workflow rather than to diagnostic reasoning alone. The experiment should state whether it studies the critique's correctness, its ability to prioritize consequential issues, or the final scene improvement after an entire correction process.

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.

An object can be detected in individual frames without being associated consistently through time. HOTA formalizes a tracking assessment concerned with detection and association (Luiten et al., 2020). ByteTrack (Zhang et al., 2021) and AB3DMOT (Weng et al., 2020) offer different tracking-method contexts in which observation and identity continuity matter. Their relevance to a label-only study is methodological: identifiers and adjacency should be retained before summarizing a sequence. No tracking accuracy follows merely from counting ground-truth trajectories or visibility changes. A performance study would need actual predictions, a declared matching procedure, and the relevant reference annotations. Likewise, evaluating a generated multi-view road scene requires evidence of object correspondence rather than only attractive frames. The proposed use of annotation summaries is to define those future test conditions and their denominators. It should not imply that a tracker has been run or that the generated scenes already preserve the same object across camera views when neither observation has been obtained.

3. Diagnosis and correction are separate outcomes

A correct critique can fail to produce a useful revision if the generator cannot implement it. An incorrect critique can occasionally lead to a better-looking scene for unrelated reasons. The study should therefore assess diagnosis and revision separately, linking them through preserved versions. It should also record new problems introduced by a correction. A local improvement is not a complete success if it breaks another camera view or removes an object required by the original instruction.

Repeated rounds need a declared stopping rule and resource record. Continuing until an acceptable scene appears evaluates an iterative search workflow, whose extra generations and reviews should be included. A fixed-round comparison answers another question about bounded effort. The report should retain all attempted corrections, failures, and unassessable outcomes. Otherwise the final collection can overrepresent successful revisions while giving the reader no evidence about how often a critique helps or how much effort the process typically requires.

The final claim should remain tied to the issue types and scene conditions assessed. A strong result on visible object omissions would not establish reliable diagnosis of subtle geometry or physical compatibility. Expanding the audit requires suitable independent evidence for those additional claims. The proposed design values localized validity: a critique can be useful on a bounded class of observable issues without becoming a universal verifier. That interpretation connects an agent's language to the scene evidence and consequences that make the diagnosis scientifically meaningful.

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.

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.

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.

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.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

The unit of provenance should be fine enough to reconstruct an important disagreement. Reusable identifiers can connect a reported score to its input evidence and the relevant model and evaluator versions (Wilkinson et al., 2016). Pipeline dependencies should be treated as part of the experiment because apparently minor changes can alter output eligibility or metric behavior (Sculley et al., 2015). A report that preserves only aggregate numbers prevents this reconstruction. Keeping per-case results and documented transformations enables alternative weighting and error analysis without repeating expensive generation. Such a record also helps identify whether two papers compare the same scientific objects or merely use similar names for different releases. Versioned evidence makes progress easier to assess across successive evaluations.

5. Conclusion

Multi-agent critique should be evaluated through verifiable claims and separately measured revision outcomes. The proposed protocol preserves scene versions, localization, uncertainty, stopping rules, and effort. It reports no new agent performance. Its contribution is an evidence boundary for future auditing studies, allowing a correct and useful critique to be recognized while preventing fluent explanations or successful-looking final scenes from automatically establishing the validity of the diagnostic process between them.

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.