Transporting Synthetic-Data Benefits to Real Perception Tasks

Zachary Parker (BZ), George Anderson (SN), Matthew Russell (MD)

Abstract

Generated driving scenes are useful for perception only if their addition improves a declared task under an appropriate real-data assessment. Visual plausibility alone cannot establish that benefit. This article develops a synthetic-data transfer experiment motivated by OmniDrive. It separates generation quality, training exposure, and downstream evaluation, with matched budgets and independent real scenes. No detector is trained here. The proposed design identifies which observations would support a transport claim and which favorable intermediate results would remain insufficient to establish real perception utility.

1. Generation and downstream usefulness are different outcomes

A synthetic clip can be visually coherent and semantically appropriate yet provide little benefit to a perception model. It may repeat familiar examples, contain label errors, or represent conditions unrelated to the deployment task. Conversely, a visually imperfect example can sometimes supply useful task information. These possibilities make downstream utility an empirical question. The present article proposes testing it directly rather than assigning benefit from an image-quality score or from the generator's architectural sophistication.

OmniDrive proposes multi-view driving generation with structured scene coordination and latent co-compression (Meng et al., 2026d). Such a workflow could motivate producing task-specific synthetic data, but the current article does not generate that data or train a perception model. It develops an experiment to assess the possible benefit. The relevant claim concerns a complete pipeline: generated examples, their usable labels, a training procedure, and performance on independent real observations. None of those stages can be replaced by an untested assumption that realistic-looking video is automatically effective training evidence.

The scientific population of a driving result is defined by its records and collection process. KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2019), BDD100K (Yu et al., 2018), and the Waymo Open Dataset (Sun et al., 2019) support important but different perception and scene-analysis tasks. Their names should not be treated as interchangeable guarantees of sensor access, geographic coverage, or annotation semantics. Datasheets motivate explicit documentation of those boundaries (Gebru et al., 2021). A label-only secondary study should therefore state which archive, classes, visibility codes, and eligibility filters define its rows. It should also preserve sequence and track identifiers because repeated annotations share an acquisition history. Such a study can establish useful distributional facts without supplying unobserved model outputs. A proposed generated-driving evaluation can draw on those facts when selecting tests, but it must separately obtain synchronized views, valid assessment evidence, and generated samples. The distinction keeps a modest annotation result interpretable and prevents the breadth of a benchmark's reputation from substituting for the narrower evidence actually analyzed.

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.

2. A matched training comparison

A proposed experiment would define a real perception task, its training data, model family, optimization budget, and primary assessment endpoint before generating synthetic additions. A baseline would use the declared real-data training workflow. A comparison condition would add a recorded synthetic collection under a matched or explicitly described training budget. If extra synthetic examples also increase optimization steps, the study should separate that resource difference where feasible rather than attribute every gain to the content of the generated data.

Synthetic provenance would connect each example to scene instructions, source references, generator version, sampling settings, and any filtering or selection. Label construction would receive independent checks suited to the task. A visually plausible vehicle does not automatically carry a valid box, class, depth, or correspondence label. The study should preserve rejected examples and quality-control effort. A filtered data workflow can be useful, but its performance and cost belong to the complete selection process rather than to unselected generation alone.

The real assessment set would remain separate from generator tuning, synthetic filtering decisions, and downstream model selection. Its split should preserve scene and recording structure appropriate to the intended transfer. Exact or related source overlap could undermine a new-scene claim even if file names differ. A documented audit can establish known boundaries and retain uncertainty where training exposure is unavailable. The report should avoid calling the evaluation contamination-free merely because one limited duplicate search found no match.

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 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 complete road-scene assessment needs a profile of outcomes. Structural similarity (Wang et al., 2004), learned perceptual correspondence (Zhang et al., 2018), generative video-distribution comparison (Unterthiner et al., 2018), and VBench's multiple dimensions (Huang et al., 2023) offer different measurement resources. Their relevance depends on the claim, reference availability, and observation unit. A strong score in one representation does not establish that a vehicle occupies a compatible location in every camera or persists coherently through an occlusion. The evaluation should therefore retain targeted scene checks alongside general visual measurements. It should also identify clips or conditions for which a score cannot be obtained validly. Label summaries can help define such difficult conditions, but they are not substitutes for generated visual observations. A secondary analysis of annotations remains a description of recorded objects and visibility. Any later claim about appearance, synchronization, or temporal quality requires the corresponding output data and a separately implemented assessment rather than an inference from the archive's label distribution.

3. Interpreting a transfer result and its limits

Outcomes would include the task's primary performance measure, relevant class or condition results, and uncertainty over the independent units represented by the design. Multiple downstream training runs can assess optimization variation, while multiple held-out scenes supply another source of evidence. These should remain distinguishable. A gain on one fixed split and one training run may be informative locally but does not establish stable benefit across every model, seed, environment, or use condition.

The study could investigate whether benefits concentrate in conditions targeted by synthetic generation. Such analyses should be planned or labeled exploratory, with suitable attention to multiple comparisons. An overall improvement can hide a decline for a smaller class or visibility state. The application determines which tradeoffs matter. Reporting separate outcomes permits that judgment without forcing an arbitrary weighted score to stand for all users. Additional real-data testing would be needed before a limited benchmark gain becomes a broad deployment or safety claim.

Finally, transport should be tied to the target environment. A synthetic collection that helps on one benchmark may fail under different sensors, geography, or annotation conventions. Reweighting observed assessment cases can change emphasis but cannot supply missing target conditions. A useful result would state exactly where the observed benefit holds and which stages of the pipeline are necessary to reproduce it. That bounded claim can guide practical data generation while leaving broader generalization to subsequent direct evidence rather than an optimistic extrapolation from visual quality.

A validation split defines which form of transfer has been observed. Related observations can share subjects, scenes, prompts, or acquisition conditions, so randomly splitting rows may leave important information on both sides of the boundary. Structured cross-validation addresses these dependencies (Roberts et al., 2017). Model selection introduces another boundary: choosing a representation, threshold, or tuning parameter using the final assessment data makes the resulting error estimate optimistic (Varma and Simon, 2006). The appropriate held-out unit follows the intended future use. Generalization to a new frame, a new trajectory, and a new subject are not equivalent claims. A study should record those units before fitting or choosing a scoring rule, and preserve an assessment set that has not become part of iterative development.

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.

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

Synthetic-data utility requires a complete, independently assessed training comparison. The proposed design preserves budgets, provenance, label validity, scene separation, and uncertainty. It reports no new downstream gain. Its contribution is a testable path from generated driving scenes to real perception outcomes, allowing favorable intermediate quality to remain useful evidence without being mistaken for the task benefit or broad transport capability that only a downstream experiment can establish.

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.