Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

Thirty Weather Records, Five Camera Paths: Counting Independent Geometry in Virtual KITTI 2

Patrick Bailey[1] · Mason Hughes[2] · Matthew Fisher[3]

[1] Malta · [2] Canada · [3] Germany

Original publication: October 6, 2026 | Article 684

Abstract

Thirty benchmark conditions need not provide thirty independent trajectories. Across five Virtual KITTI 2 scenes and six non-yaw weather or lighting variants, the supplied camera extrinsic files are identical within each scene, leaving five distinct files. Mean left-camera translation is 0.606 metres per adjacent frame under either counting scheme. Its standard error rises from 0.044 under an illustrative independent-condition calculation to 0.119 when dependence is grouped by scene. The weather conditions may still produce meaningfully different visual evaluations; the duplication applies to this geometric endpoint. The audit shows why the unit supporting an uncertainty claim should be checked separately for each measured outcome.

Source-identity and covariance audit of public synthetic metadata. The study makes no claim that different weather renderings look identical.

Thirty Weather Records, Five Camera Paths: Counting Independent Geometry in Virtual KITTI 2

Source-identity and covariance audit of public synthetic metadata. The study makes no claim that different weather renderings look identical.

Abstract

Thirty benchmark conditions need not provide thirty independent trajectories. Across five Virtual KITTI 2 scenes and six non-yaw weather or lighting variants, the supplied camera extrinsic files are identical within each scene, leaving five distinct files. Mean left-camera translation is 0.606 metres per adjacent frame under either counting scheme. Its standard error rises from 0.044 under an illustrative independent-condition calculation to 0.119 when dependence is grouped by scene. The weather conditions may still produce meaningfully different visual evaluations; the duplication applies to this geometric endpoint. The audit shows why the unit supporting an uncertainty claim should be checked separately for each measured outcome.

1. A benchmark condition is not automatically an independent unit

A driving benchmark can contain multiple weather, lighting, and camera conditions for one underlying scene. These variants can be valuable because they change what a model sees. The same variants can nevertheless share other properties, such as the camera trajectory. Counting condition names without checking the endpoint's source information can therefore exaggerate how many independent motion examples support a numerical summary.

We ask how many distinct camera trajectories are represented by six non-yaw variants across five public virtual scenes. We then calculate one common motion statistic under two covariance assumptions, keeping its point estimate fixed. This separates the observed data-identity question from the inferential consequence of treating repeated geometry as independent. It does not assess whether a video model responds differently to the weather conditions.

The multi-view driving-generation setting in Meng et al., 2026 motivates careful accounting of evaluation conditions. We do not run that model or analyze its generated videos. Our study examines a source resource that could inform evaluation design. A camera-file identity check can clarify a motion endpoint's effective units without establishing anything about the fidelity, consistency, or controllability of a particular world-model architecture.

2. Multiple conditions can support different legitimate questions

The virtual-world approach in Gaidon et al., 2016 and its extension Yohann et al., 2020 provide annotated alternatives to exclusively real-data evaluation. Weather and lighting variants can support a controlled visual comparison when other scene elements are held fixed. That is a strength of the design. The issue arises only if a reader treats unchanged geometry under those variants as additional independent realizations of geometry, rather than as a shared component of the controlled comparison.

Real-driving resources such as Geiger et al., 2012, Liao et al., 2021, Caesar et al., 2019, Sun et al., 2019, and Yu et al., 2018 organize scenes and observations under their own collection designs. Simulation settings such as Dosovitskiy et al., 2017 create other kinds of controlled variation. Planning-oriented resources in Caesar et al., 2021 and Wilson et al., 2023 concern additional task structures. We do not

combine their scene counts or assume that a condition label has a universal independence meaning across these resources.

Driving-generation systems including Gao et al., 2023, Wang et al., 2023, and Wen et al., 2023 address different synthesis or conditioning problems. Interactive world-model work such as Bruce et al., 2024 and Alonso et al., 2024 adds other settings in which trajectories and rendered observations can vary separately. These sources locate the broader need to define an evaluation endpoint. Their published performance is not used as additional observations, and no inference about their handling of weather follows from our source-file audit.

3. Restrict the question to non-yaw camera geometry

The official metadata archive supplies five scene identifiers: 01, 02, 06, 18, and 20. For each, we select clone, fog, morning, overcast, rain, and sunset. These six variants yield thirty scene-condition records. The four yaw variants are excluded because camera orientation is intentionally changed there, which would answer a different identity question. We retain only the left camera when calculating the motion statistic.

We read the complete extrinsic file for every selected condition. A file hash supplies an exact byte-identity check, and a separate numeric comparison checks the matrices after parsing. The numeric comparison aligns frame and camera identifiers rather than assuming that row counts alone establish correspondence. This dual check avoids treating a matching numerical summary as proof that the underlying camera trajectories are identical.

All six extrinsic files within each scene have the same SHA-256 hash. Across all thirty conditions, there are five distinct hashes. After selecting the left camera and aligning frames, the maximum elementwise matrix difference from that scene's clone file is zero. These are direct facts about the captured metadata. They do not require a learned duplicate detector or a visual similarity judgment over the rendered images.

The file identity also includes the additional camera records present in each full file, although our motion endpoint uses only camera zero. We do not count the other stereo camera as another independently sampled environment. A stereo view can add a legitimate observation condition while sharing the same scene dynamics. That distinction parallels the weather comparison and reinforces why independence must be tied to the quantity being estimated.

4. Calculate a motion statistic with a declared weighting rule

For each selected left-camera trajectory, we derive its center from the released rotation and translation coefficients. We calculate Euclidean distance between centers at adjacent frame indices and average those distances within the trajectory. The resulting endpoint is mean camera translation in metres per adjacent frame. No frame rate is assumed, so the quantity is not converted to metres per second.

The calculation uses the source matrices with their captured numerical precision. Identical files consequently produce identical per-condition motion summaries within a scene. This is expected from the identity check and should not be presented as a surprising correlation between weather and motion. The empirical source finding is the identity itself; the repeated motion value follows from applying the same deterministic calculation to the same data.

Each of the thirty condition records receives equal weight in the illustrative overall mean. Since every scene contributes the same six identical summaries, that mean equals the mean of the five scene summaries with equal scene weights. It does not equal a frame-weighted mean unless the scene lengths

and step values happen to support that equality. We keep the weighting rule explicit rather than allowing a longer trajectory to dominate unnoticed.

The common mean is 0.606326 metres per adjacent frame. This value describes the five selected virtual trajectories under the equal-scene construction. It is not an average driving speed for a country, a measure of camera-control quality, or a model-generated trajectory error. The source scenarios are fixed and few, so even a technically correct covariance formula cannot make them representative of an arbitrary road population.

5. Hold the mean fixed while changing the covariance assumption

We fit an intercept-only ordinary-least-squares model to the thirty condition summaries. The first table entry uses ordinary independent-row covariance with twenty-nine residual degrees of freedom. It is shown as an illustration of what condition-level independence would imply. The exact within-scene file repetition contradicts interpreting those thirty rows as thirty independently sampled camera paths, so this first interval is not endorsed as an appropriate independent-trajectory inference.

The second entry uses the same outcome rows and identical mean coefficient, but clusters covariance by scene. Following the dependence concerns in Cameron and Miller, 2015, it uses five scene groups, the recorded correction, and a t reference with four degrees of freedom. The difference between the entries is therefore not caused by a new sample or another geometric calculation. It comes from the declared grouping and reference distribution used for uncertainty.

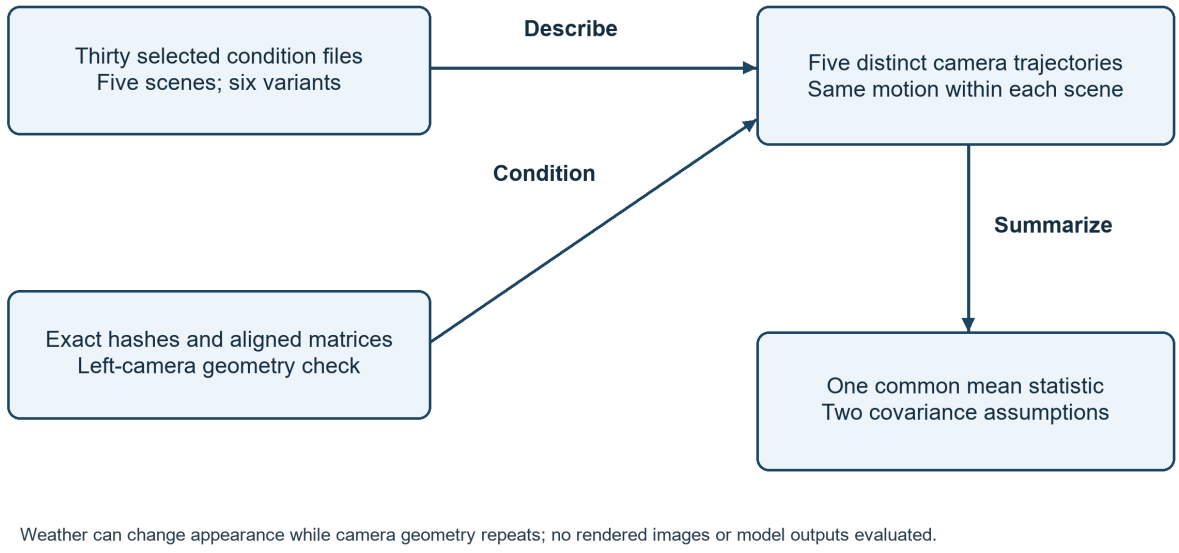

Figure 1. Condition-level camera files are compared for identity before one mean translation statistic is summarized under two dependence assumptions.

Table 1. The same mean translation per adjacent frame under independent-condition and scene-grouped covariance assumptions.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.606326 (0.044339)	[0.515642, 0.697010]	30
M1: fit	R-squared = 0.0000	-	-	30
M2	Mean outcome/contrast	0.606326 (0.119387)	[0.274855, 0.937797]	30
M2: fit	R-squared = 0.0000	-	-	30

Model labels: M1 = Condition-row independence illustration; M2 = Scene-level dependence accounting.

Heteroskedasticity correction in White, 1980 does not by itself encode shared trajectory information across condition rows. Repeated-data modeling in Laird and Ware, 1982 provides another framework for dependence, but no mixed model is fitted here. The scene-clustered entry remains a small-group approximation, not an exact statement of uncertainty over all possible virtual environments. Its purpose is to make the consequence of the repeated unit visible.

6. Identical point estimates conceal materially different intervals

The independent-condition entry reports standard error 0.044339 and a 95 percent interval from 0.515642 to 0.697010 metres per frame. The scene-grouped entry reports standard error 0.119387 and interval 0.274855 to 0.937797. Both have the same coefficient, 0.606326. The clustered standard error is 2.6926 times the illustrative independent-row value; the interval also changes its t reference because only five scene groups support the chosen sampling interpretation.

These values are not competing estimates of how much the camera moved in the captured files. That finite-data motion summary is already determined by the supplied trajectories. The intervals have meaning only if the scenes or conditions are interpreted as observations from a broader sampling process. The audit shows that the unit used for such an interpretation cannot be justified by counting filenames alone.

Both intercept models have effectively zero R-squared. That is appropriate for a mean summary and provides no evidence about a driving model's predictive fit. Scoring principles in Gneiting and Raftery, 2007 and accuracy-scale discussion in Hyndman and Koehler, 2006 reinforce that the endpoint must be named. Here it is a trajectory statistic in metres per adjacent frame, rather than a forecast error, a generation score, or a quality judgment.

7. Shared geometry does not make weather evaluation redundant

Different atmospheric or lighting conditions can change image appearance even when the camera matrices are identical. A vision model may therefore behave differently under fog, rain, or sunset. Our metadata equality does not establish equality of those rendered pixels or model responses. It would be incorrect to discard all weather evaluations simply because this particular geometric endpoint repeats.

For a within-scene weather-effect experiment, shared geometry can be advantageous because it helps hold another factor constant. The relevant analysis would then preserve the pairing rather than pretend that the weather conditions are unrelated scene draws. Our source check supports that distinction. It does not estimate a weather effect, because neither images nor model outputs enter the current outcome.

Likewise, intentionally changed yaw variants could supply meaningful geometry variation, but should be described in relation to their common underlying scene. We excluded them to keep the current equality question focused. Their existence does not turn the thirty selected non-yaw records into independent trajectories. A broader benchmark report could distinguish scene identities, appearance conditions, viewpoint conditions, and the specific endpoint measured in each.

8. Generalization depends on the unit and the endpoint

Validation guidance in Roberts et al., 2017 is relevant when train and test partitions share structured observations. If geometry is repeated across conditions, a random condition split can carry the same camera path into both sides. Whether that is problematic depends on the intended task. A visual robustness experiment with fixed geometry can deliberately use such pairing, while a claim of transfer to new camera trajectories requires a different separation.

Selection concerns in Varma and Simon, 2006 would arise if scene groups or variants were chosen after inspecting a favored model's output. Here the condition set is declared and the file comparison precedes any model-performance analysis. Influence ideas in Cook, 1977 could identify which of the five scenes contributes most to mean variation, but no scene is removed. The small scene population remains visible in the covariance calculation and the interpretation.

Missing-data principles in Little and Rubin, 2002 caution against assuming that absent environments resemble the supplied five scenes. Measurement-error reasoning in Carroll et al., 2006 also applies to camera matrices and any future predicted trajectories. Exact equality of captured files is strong evidence of source identity, but not of perfect geometric truth. A copied rounded matrix remains a copied rounded matrix; repeating it cannot reduce the uncertainty in its original acquisition or construction.

9. A practical reporting rule from a narrow source audit

Statistical interpretation under Wasserstein and Lazar, 2016 supports communicating the fixed mean, the identity evidence, and the covariance assumptions together. Multiplicity considerations in Benjamini and Hochberg, 1995 become relevant if many benchmark endpoints are searched for dramatic independence effects. We report one predeclared camera statistic and the exact source identity that motivates its grouping. The result does not accuse the dataset creators of making an invalid claim; it clarifies what a secondary user's counting convention can support.

A useful evaluation report could state both the number of conditions and the number of independent scene-level units for its endpoint. It could also identify which variations are paired controls, which introduce new geometry, and which introduce only appearance changes. Such reporting would preserve the value of controlled virtual-world variants while preventing a large condition count from silently substituting for independent evidence about a motion population.

Documentation under Gebru et al., 2021 records all selected paths, full file hashes, aligned numeric equality, camera choice, motion calculation, and weighting. Wilkinson et al., 2016 motivates preservation of those checks with the official metadata and source attribution. The contribution is a concrete example of endpoint-specific independence: thirty non-yaw weather records contain five distinct camera paths, and uncertainty about a scene-level motion summary changes substantially when that dependence is acknowledged.

The exact camera-path identity does not make the appearance variants redundant for every purpose. A visual predictor could respond differently to weather while the metadata geometry remains identical. What repeats is the endpoint used in this audit, not every potentially informative property of the resource.

Reporting independent units should therefore follow the analyzed outcome and dependence structure instead of assigning one universal effective sample size to an entire multimodal dataset.

References

1. Bruce, Jake, et al. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.
2. Caesar, Holger, et al. "nuScenes: A multimodal dataset for autonomous driving" *arXiv preprint arXiv:1903.11027*, 2019. https://arxiv.org/abs/1903.11027.
3. 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.
4. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
5. Alonso, Eloi, et al. "Diffusion for World Modeling: Visual Details Matter in Atari" *arXiv preprint arXiv:2405.12399*, 2024. https://arxiv.org/abs/2405.12399.
6. Wilson, Benjamin; Qi, William; Agarwal, Tanmay; Lambert, John; Singh, Jagjeet; Khandelwal, Siddhesh; Pan, Bowen; Kumar, Ratnesh; Hartnett, Andrew; Pontes, Jhony Kaesemodel; Ramanan, Deva; Carr, Peter; Hays, James. (2023). Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. arXiv:2301.00493. https://arxiv.org/abs/2301.00493.
7. Gaidon, Adrien; Wang, Qiao; Cabon, Yohann; Vig, Eleonora. (2016). Virtual Worlds as Proxy for Multi-Object Tracking Analysis. CVPR 2016, 4340-4349. https://openaccess.thecvf.com/content_cvpr_2016/html/Gaidon_Virtual_Worlds_as_CVPR_2016_paper.html.
8. 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.
9. Dosovitskiy, Alexey; Ros, German; Codevilla, Felipe; Lopez, Antonio; Koltun, Vladlen. (2017). CARLA: An Open Urban Driving Simulator. arXiv:1711.03938. https://arxiv.org/abs/1711.03938.
10. Gao, Ruiyuan; Chen, Kai; Xie, Enze; Hong, Lanqing; Li, Zhenguo; Yeung, Dit-Yan; Xu, Qiang. (2023). MagicDrive: Street View Generation with Diverse 3D Geometry Control. arXiv:2310.02601. https://arxiv.org/abs/2310.02601.
11. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.
12. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
13. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
14. Yohann, Cabon,; Naila, Murray,; Martin, Humenberger,. (2020). Virtual KITTI 2. https://arxiv.org/abs/2001.10773.
15. Laird, Nan M. and James H. Ware. “Random-Effects Models for Longitudinal Data.” *Biometrics*, vol. 38, no. 4, 1982, pp. 963. https://doi.org/10.2307/2529876.
16. Liao, Yiyi; Xie, Jun; Geiger, Andreas. (2021). KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. arXiv:2109.13410. https://arxiv.org/abs/2109.13410.
17. Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.
18. 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.
19. Caesar, Holger, et al. "NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles." *arXiv preprint arXiv:2106.11810*, 2021. https://arxiv.org/abs/2106.11810.

20. 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.

21. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

22. Wang, Xiaofeng; Zhu, Zheng; Huang, Guan; Chen, Xinze; Zhu, Jiagang; Lu, Jiwen. (2023). DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving. arXiv:2309.09777. https://arxiv.org/abs/2309.09777.

23. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

24. 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.

25. Wen, Yuqing; Zhao, Yucheng; Liu, Yingfei; Jia, Fan; Wang, Yanhui; Luo, Chong; Zhang, Chi; Wang, Tiancai; Sun, Xiaoyan; Zhang, Xiangyu. (2023). Panacea: Panoramic and Controllable Video Generation for Autonomous Driving. arXiv:2311.16813. https://arxiv.org/abs/2311.16813.

26. 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.

27. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

28. Geiger, A.; Lenz, P.; Urtasun, R.. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2012.6248074.

29. Meng, Zijie; Liu, Yufei; Ma, Chengqian; Li, Zhiyu; Liu, Jiyuan; Nie, Wenhua; Wei, Bingcai; Chen, Shuqin; Xu, Weichen; Yuan, Jiquan; Zhang, Miao. (2026). OmniDrive: An LLM-Choreographed Multi-Agent World Model with Unified Latent Co-Compression for Multi-View Driving Video Generation. arXiv:2606.17536. https://arxiv.org/abs/2606.17536.

30. 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.