Journal of Audit and Accountability
Volume 1 · Number 1 · 2026

Does the Visible Leader Help? A Purged One-Second Forecast Comparison in NGSIM

Henry Peterson[1] · Tyler Coleman[2] · Paul Lewis[3]

[1] Serbia · [2] Montenegro · [3] Estonia

Original publication: November 7, 2026 | Article 592

Abstract

Nearby vehicles should matter to motion prediction, but that intuition needs a comparison against the focal vehicle's own state. In a capped I-80 NGSIM window, we fitted an own-speed-and-acceleration baseline and a model adding current lead-relative speed and headway. Training targets ended before the chronological split. Across 60,740 held-out vehicle-time rows, root mean squared error fell from 1.002 to 0.932 m/s. Equal-time paired squared loss declined by 0.135, with HAC interval -0.164 to -0.105. The holdout spans only 47.8 seconds, so this is local predictive evidence rather than independent-drive validation or evidence about generated driving videos.

Short-window public-trajectory research draft. No driving policy or generative world model is evaluated; human review is pending.

Does the Visible Leader Help? A Purged One-Second Forecast Comparison in NGSIM

Short-window public-trajectory research draft. No driving policy or generative world model is evaluated; human review is pending.

Abstract

Nearby vehicles should matter to motion prediction, but that intuition needs a comparison against the focal vehicle's own state. In a capped I-80 NGSIM window, we fitted an own-speed-and-acceleration baseline and a model adding current lead-relative speed and headway. Training targets ended before the chronological split. Across 60,740 held-out vehicle-time rows, root mean squared error fell from 1.002 to 0.932 m/s. Equal-time paired squared loss declined by 0.135, with HAC interval -0.164 to -0.105. The holdout spans only 47.8 seconds, so this is local predictive evidence rather than independent-drive validation or evidence about generated driving videos.

1. Interaction information needs a baseline

A traffic trajectory is embedded in a stream of other vehicles. A driver may accelerate toward a faster leader or slow when spacing becomes constrained. Yet an apparent interaction feature can merely restate information already present in the focal vehicle's speed and acceleration. We ask whether current observations of the recorded preceding vehicle add predictive information for a one-second velocity change after a transparent own-state baseline is included.

The NGSIM vehicle-trajectory resource supplies vehicle identifiers, time and frame fields, velocities, accelerations, and leader-related fields (Federal Highway Administration, 2026). It supports a direct test of that incremental-information question without generating video or training a large neural model. Our data acquisition is capped to the first 150,000 ordered I-80 rows, making the experiment computationally modest and reproducible. That cap also creates a short observation window whose limits must remain central to interpretation.

Multi-agent driving world models such as Meng et al., 2026 motivate attention to interactions among vehicles. The present analysis does not evaluate their latent representations, compression, view synthesis, or planning capabilities. It tests a simple observable interaction channel in numeric trajectories. A benefit from leader information is consistent with its potential usefulness, but cannot establish that any particular world-model architecture captures it or generates physically consistent scenes.

2. Data scope and related prediction settings

Autonomous-driving resources such as KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2019), Waymo Open Dataset (Sun et al., 2019), and BDD100K (Yu et al., 2018) support different sensor and task configurations. Argoverse 2 (Wilson et al., 2023) and nuPlan (Caesar et al., 2021) add other forecasting or planning contexts. We do not combine them with NGSIM or claim cross-dataset validation. Their relevance is to emphasize that numeric freeway trajectories represent only one narrow setting within driving research.

Simulation in CARLA (Dosovitskiy et al., 2017) and driving-generation work including Gao et al., 2023, Wang et al., 2023, Li et al., 2023, and Wen et al., 2023 concern additional objectives. Photorealistic video, controllability, physical motion, and forecasting loss require separate measurements. Our dependent

variable is a one-second velocity change in the released numeric record. It cannot be used as a surrogate score for visual realism or a guarantee that a generated multi-view scene preserves interactions correctly.

Traffic-network prediction methods in Li et al., 2017 and Yu et al., 2017 illustrate the broader value of structured dependence. Graph-signal perspectives in Ortega et al., 2017 and Shuman et al., 2012 also motivate treating neighboring observations as potentially useful information. Our relation is simpler: each focal row uses only its currently recorded preceding vehicle, when that vehicle is present in the same captured clip and timestamp. It is not an inferred road graph or a full learned interaction network.

3. Capture integrity and physical units

The acquired slice contains 150,000 rows, 297 vehicle identifiers, and 1,627 distinct timestamps over 162.6 seconds. The combination of clip key, timestamp, and vehicle identifier is unique. We define a clip key as global_time minus one hundred times frame_id; it is constant throughout this slice. This check guards against joining identifiers across unrelated recording windows that could reuse vehicle numbers.

Source velocities are in feet per second, accelerations in feet per second squared, and headway in feet. We multiply by 0.3048 to express the modeled quantities in metric units. We use the timestamp fields for exact elapsed-time joins rather than inventing a local calendar interpretation. The target row is found by matching the same focal vehicle exactly 1,000 milliseconds later within the same clip, not by taking the tenth next available row after a possible gap.

For leader features, the recorded preceding identifier must be positive and match an observed vehicle at the same clip and time. Headway must also be positive. A missing leader match is excluded rather than assigned zero relative speed. After these restrictions and the future-target match, 138,600 eligible rows remain. This is a selected following-vehicle sample, not all vehicles in the acquired slice and not a sample of situations with an unobserved leader.

Missing-data principles in Little and Rubin, 2002 help clarify that exclusion can alter the target population. The first and last parts of trajectories, unmatched leaders, and positive-headway restrictions can all change which situations contribute. We do not infer how a model would perform when neighboring vehicles are occluded or absent. Such a test would require a declared missing-neighbor policy and an evaluation outcome defined for those cases.

4. A chronological split with purging

We sort the captured unique timestamps and use the timestamp at seventy percent of their count as the split. Training rows must have their one-second target strictly before that boundary. Test rows begin at or after the boundary. Rows whose current time precedes the split but whose target reaches or crosses it are purged. This removes 1,352 otherwise eligible rows and prevents training labels from extending into the evaluation period.

The final training set has 76,508 rows from 193 vehicles and 109 one-second time blocks. The test set has 60,740 rows at 479 retained timestamps, spanning 47.8 seconds. Some vehicles can appear in both periods because the intended question is local chronological continuation, not unseen-driver transfer. We do not describe this as a vehicle-disjoint evaluation or an independent recording. The short horizon and continuing traffic conditions materially limit generalization.

Structured-validation guidance in Roberts et al., 2017 and selection-bias concerns in Varma and Simon, 2006 motivate preserving temporal order. A random row split would scatter adjacent observations from the same trajectories across fitting and evaluation. Purging addresses the direct target-overlap problem, but not every form of temporal dependence or source preprocessing. Released trajectory velocities and

accelerations may themselves reflect tracking or smoothing procedures, so the predictors should not be assumed equivalent to unprocessed live sensor measurements.

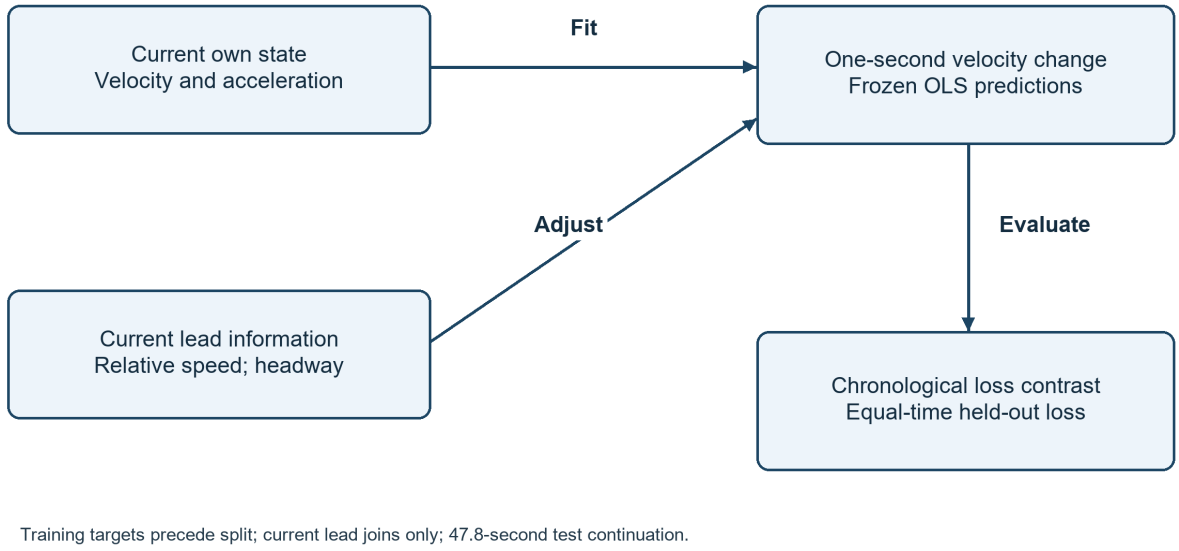

Figure 1. Current own speed and acceleration form the baseline; current lead-relative speed and headway augment it. Frozen training fits feed a purged chronological test and a paired equal-time loss comparison.

5. Models and comparison units

Both models predict future velocity minus current velocity in meters per second. The baseline includes current focal velocity and acceleration. The augmented model adds leader velocity minus focal velocity and log one plus metric headway. These features use current observations only. No future leader speed, future gap, future lane assignment, or target-derived predictor enters either fit. Model coefficients are estimated by ordinary least squares on the same training rows.

Training covariance clusters in two dimensions: focal vehicle and one-second time block. Cameron and Miller, 2015 motivates allowing both repeated-vehicle and common-time dependence, while White, 1980 supplies the general heteroskedastic framework. The reported coefficient intervals use the smaller cluster dimension for t inference. This is an approximation for the acquired short window, not proof that all residual traffic dependence has been captured by those two grouping schemes.

The primary test comparison first calculates augmented minus baseline squared error for every held-out row. It then averages those differences within timestamp and fits an intercept-only regression across 479 timestamps. This gives equal weight to observed test times rather than more weight to times with more recorded vehicles. HAC covariance uses ten retained timestamp lags, with correction and t-based intervals. The lag is fixed before result interpretation.

Predictive-accuracy work by Diebold and Mariano, 1995 and West, 1996 motivates separating an incremental training association from a held-out loss benefit. Newey and West, 1987 provides the HAC covariance basis. Forecast-metric discussions in Hyndman and Koehler, 2006 and scoring principles in Gneiting and Raftery, 2007 explain why the chosen loss must be explicit. We retain row-weighted root mean squared error as a familiar descriptive measure, while the inferential comparison uses equal-time mean squared-loss differences.

Table 1. Selected augmented-model coefficients and the separate held-out paired squared-loss contrast. Training coefficients use two-way clustering; test loss uses HAC ten retained timestamps.

Model	Term	Estimate (SE)	95% CI	N
M1: fit	R-squared = 0.0668	-	-	76508
M2	Lead minus own velocity	0.2683 (0.0172)	[0.2341, 0.3025]	76508
M2	Log(1 + headway metres)	0.2410 (0.0501)	[0.1416, 0.3404]	76508
M2: fit	R-squared = 0.2550	-	-	76508
M3	Mean contrast	-0.1347 (0.0150)	[-0.1642, -0.1053]	479
M3: fit	R-squared = 0.0000	-	-	479

Model labels: M1 = Own-state fit; M2 = Own-plus-leader fit; M3 = Held-out equal-time leader minus own squared loss.

6. Incremental information in this window

The augmented training model has a lead-relative velocity coefficient of 0.2683, with standard error 0.0172 and interval 0.2341 to 0.3025. Its log-headway coefficient is 0.2410, with interval 0.1416 to 0.3404. The model R-squared rises from 0.0668 in the own-state fit to 0.2550 with leader information. These are conditional predictive associations, not estimates of the causal response to externally changing another vehicle's speed.

On held-out rows, baseline root mean squared error is 1.0017 m/s and augmented error is 0.9319 m/s. The equal-time mean squared-loss difference is -0.1347, with standard error 0.0150 and interval -0.1642 to -0.1053. Negative values favor the augmented model. The two summaries use different weighting conventions, but both show a predictive improvement in the captured continuation window.

The held-out mean-difference regression has zero R-squared by construction. Its purpose is to summarize a paired loss contrast, not explain the full variation in vehicle behavior. The training fit statistics refer to velocity-change regressions with covariates and should not be compared directly with that zero. Keeping the training and evaluation rows distinct also prevents a significant leader coefficient from being mistaken for independent evidence of a test improvement.

7. Why this is local evidence

Less than one minute of held-out continuation cannot represent variation across road geometry, congestion regimes, weather, tracking systems, or human driving populations. Thousands of correlated row observations do not create many independent traffic episodes. The HAC interval addresses short-lag dependence under its assumptions, but it cannot manufacture new episodes or measure between-recording uncertainty. A strong local result should therefore remain local in the manuscript's claims.

Influence considerations from Cook, 1977 could motivate checking whether an unusual trajectory dominates an association. We do not remove vehicles based on their errors or select a better split after seeing the result. A further study should specify multiple independent windows, a missing-leader policy, and a comparison against more demanding baselines. It could then test whether the incremental gain persists when own-trajectory history, lane context, and nonlinear car-following behavior are represented.

There is also no closed-loop driving intervention. Forecasting another second from observed trajectories differs from choosing acceleration actions and observing the resulting traffic response. The data do not estimate safety, comfort, collision frequency, or planning efficiency. They likewise contain no generated camera views to inspect. The appropriate implication for multi-agent modeling is simply that a currently visible relationship contains useful local information beyond this modest own-state baseline.

8. Reusable trajectory evidence

The headway variable is a recorded front-center spacing measure, not a verified collision margin or stopping-distance calculation. Its logarithm reduces the influence of large gaps in a linear model but does not impose a physical safety law. Likewise, the relative-speed coefficient can reflect traffic conditions and driver responses jointly. Neither coefficient should be inserted into a controller as if the regression had identified a stable behavioral rule. The experiment evaluates a prediction from observed current states, with no intervention on either vehicle.

The row cap also affects which vehicles have complete future targets. Near the capture boundary, a vehicle may leave the recorded slice before its one-second target becomes available. Those rows are excluded by the exact join rather than extrapolated. The common held-out sample ensures that the two fitted models are evaluated on identical eligible observations, but it does not guarantee that excluded boundary cases have the same error distribution. A broader extraction would be needed to investigate that selection mechanism rather than assuming it is harmless.

Equal-time averaging is a deliberate choice because the number of visible vehicles varies across the short clip. A row-weighted loss would emphasize busier timestamps, whereas the primary contrast gives each retained time one contribution. Neither weighting is universally correct. An application that predicts each vehicle independently might prefer the row-weighted objective, while a traffic-state monitor might care about time-balanced performance. We retain both the row predictions and the aggregation rule so that this distinction remains inspectable without pretending that the two summaries are independent replications.

The next validation unit should be a complete independent traffic episode. Repeating a row split within the same 162.6 seconds would add little evidence about transfer to another congestion state. A more persuasive study would specify windows before fitting, preserve all temporal joins, and estimate variation across episodes. This local result is valuable because it makes the incremental comparison concrete, but its confidence interval remains narrower in scope than the range of conditions faced by a real driving system.

Documentation follows Gebru et al., 2021 by recording the source query, cap, clip check, joins, unit conversions, temporal boundary, and purged rows. Wilkinson et al., 2016 motivates retaining source provenance, analysis tables, predictions, and numerical outputs. Public trajectory availability does not imply unrestricted use outside the source conditions. No person-level identity is inferred from the numeric vehicle identifiers, and no claim of a newly collected traffic experiment accompanies the paper.

The study's contribution is a transparent incremental-information test with a meaningful temporal safeguard. Current leader features improve one-second velocity-change prediction within one short I-80 continuation. That finding motivates broader validation, but it does not establish the quality of a generative driving model or a safe controller. The distinction between observable predictive information and a validated multi-agent simulation remains essential to the interpretation.

References

1. 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.
2. Li, Yaguang; Yu, Rose; Shahabi, Cyrus; Liu, Yan. (2017). Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. arXiv:1707.01926. https://arxiv.org/abs/1707.01926.
3. 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.
4. Diebold, Francis X.; Mariano, Roberto S.. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599.
5. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
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. 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.
8. West, Kenneth D.. (1996). Asymptotic Inference about Predictive Ability. Econometrica. https://doi.org/10.2307/2171956.
9. 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.
10. Federal Highway Administration. (2026). Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. Official resource, accessed 6 November 2026. https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Vehicle-Trajectories/8ect-6jqj.
11. Li, Xiaofan; Zhang, Yifu; Ye, Xiaoqing. (2023). DrivingDiffusion: Layout-Guided multi-view driving scene video generation with latent diffusion model. arXiv:2310.07771. https://arxiv.org/abs/2310.07771.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. Ortega, Antonio; Frossard, Pascal; Kovačević, Jelena; Moura, José M. F.; Vandergheynst, Pierre. (2017). Graph Signal Processing: Overview, Challenges and Applications. arXiv:1712.00468. https://arxiv.org/abs/1712.00468.
18. 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.
19. 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.
20. Caesar, Holger, et al. "nuScenes: A multimodal dataset for autonomous driving" *arXiv preprint arXiv:1903.11027*, 2019. https://arxiv.org/abs/1903.11027.
21. 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.
22. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
23. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

24. 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.

25. Yu, Bing; Yin, Haoteng; Zhu, Zhanxing. (2017). Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. arXiv:1709.04875. https://arxiv.org/abs/1709.04875.

26. Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

27. Shuman, David I; Narang, Sunil K.; Frossard, Pascal; Ortega, Antonio; Vandergheynst, Pierre. (2012). The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. arXiv:1211.0053. https://arxiv.org/abs/1211.0053.

28. 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.

29. 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.

30. 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.