**Temporal Systems and Forecasting**
Volume 1 · Number 1 · 2026

# Stable Error Magnitude, Shifting Bias: A Frozen Traffic Nowcast in Two Later Calendar Windows

Luke Cooper[1] · Liam Powell[2] · Jason Reed[3]

[1] United States · [2] Australia · [3] Ireland

Original publication: November 4, 2026 | Article 596

## Abstract

Mean absolute error changed little when a frozen traffic model moved from January-September 2017 to the same months of 2018. Its average signed error changed more clearly. Across 519 complete dates at one I-94 station, the calendar-adjusted 2018 contrast was 1.4 vehicles per hour for absolute error and 66.7 for predicted-minus-observed error. The distinction matters for temporal monitoring: an apparently stable error magnitude can coexist with a directional shift. All predictions use simultaneous weather and a model fitted through 2016, so the exercise is retrospective nowcasting.

*Exploratory public-data draft. Concurrent-weather nowcasting at one station; no causal year effect, future-weather forecast, or generated-driving-video evaluation.*

# Stable Error Magnitude, Shifting Bias: A Frozen Traffic Nowcast in Two Later Calendar Windows

> Exploratory public-data draft. Concurrent-weather nowcasting at one station; no causal year effect, future-weather forecast, or generated-driving-video evaluation.


## Abstract

Mean absolute error changed little when a frozen traffic model moved from January-September 2017 to the same months of 2018. Its average signed error changed more clearly. Across 519 complete dates at one I-94 station, the calendar-adjusted 2018 contrast was 1.4 vehicles per hour for absolute error and 66.7 for predicted-minus-observed error. The distinction matters for temporal monitoring: an apparently stable error magnitude can coexist with a directional shift. All predictions use simultaneous weather and a model fitted through 2016, so the exercise is retrospective nowcasting.


## 1. Temporal transport has more than one error summary

A model used in a later period can retain a similar average error magnitude while changing the direction of its mistakes. Underprediction can move toward overprediction without a large change in absolute error. A monitoring system that records only one unsigned loss can miss that movement. For traffic applications, this distinction concerns what the numerical prediction says about a road state, rather than whether a single aggregate metric has crossed a significance threshold.

The research question asks how a frozen historical traffic-volume nowcast behaves in two later, matched-calendar windows at the same station. We compare January through September of 2017 and 2018 using both daily mean absolute error and daily mean signed error. The public Metro Interstate Traffic Volume dataset Hogue, 2019 supplies hourly westbound I-94 counts and simultaneous weather features. Earlier observations fit the model; later observations evaluate it. No information about later traffic outcomes is used to refit the predictor.

This is relevant background for systems that organize driving environments over time. Meng et al., 2026 propose an LLM-choreographed multi-agent world model for multi-view driving-video generation. Its setting motivates attention to the transport of environment-related representations, but the present evidence is a station-level numerical series. We do not run its world model, generate camera views, or infer video fidelity from traffic-count error. The study addresses a simpler diagnostic: whether two temporal error summaries tell the same story for a frozen road-state model.

Our analytical proposition is that stable absolute loss need not imply stable directional bias. The proposition is tested descriptively with two outcome models on the same complete dates. A difference between years is not interpreted as the causal effect of time, climate, road management, or a particular intervention. The observations represent two later periods at one location. Their utility is to make the monitoring distinction concrete while showing how timestamp construction, calendar composition, and missing dates affect the analysis.

## 2. A road count is a different environment measurement

Driving datasets measure different aspects of an environment. KITTI Geiger et al., 2012, nuScenes Caesar et al., 2019, and Waymo Sun et al., 2019 provide settings for perception and related driving tasks, while BDD100K Yu et al., 2018 organizes diverse driving observations. Their camera and scene information cannot be reconstructed from the hourly count analyzed here. Conversely, a point traffic series can reveal temporal demand patterns that an individual short video does not directly identify. We retain the station-level scope rather than treating all driving data as one interchangeable evidence object.

Simulation and planning sources also define distinct tasks. CARLA Dosovitskiy et al., 2017 supplies a simulation environment; nuPlan Caesar et al., 2021 and Argoverse 2 Wilson et al., 2023 organize other autonomous-driving evidence. None is executed or added to the evaluation sample. The I-94 dataset contains no per-vehicle trajectories, collision outcome, or multi-view correspondence. Its hourly volume is an aggregate count. A useful temporal diagnostic for that count does not validate a policy's safety or its ability to model individual interacting road users.

Learned environment models, including World Models Ha and Schmidhuber, 2018 and DreamerV3 Hafner et al., 2023, motivate studying how a representation behaves outside its fitting experience. BEVFormer Li et al., 2022 and unified perception-planning work Hu et al., 2022 provide additional context for structured driving representations. Our fitted model is ordinary least squares on weather and calendar variables. Its interpretability makes error construction inspectable, but no result below ranks it against those architectures or claims that their temporal behavior can be inferred from this baseline.

Temporal evaluation itself has a methodological literature. Rolling-origin and out-of-sample forecasting discussion Tashman, 2000 and work on cross-validation for autoregressive prediction Bergmeir et al., 2018 distinguish time-respecting evaluation from random mixing of observations. The present model is frozen after an earlier fitting period, and later years supply outcomes. Because current weather enters the same-hour prediction, however, this is a retrospective nowcast. It is not a future forecast that assumes the realized weather would have been known at an earlier decision time.

## 3. Constructing physical-hour observations

The raw file contains 48,204 rows but only 40,575 unique timestamps. Multiple weather descriptions can accompany the same traffic hour. Counting every row as an independent volume observation would therefore overrepresent some hours. We inspect the selected numeric predictors and traffic count within each timestamp. Groups with conflicting numeric values are removed as complete groups; groups whose selected values agree are represented once. Categorical weather descriptions and holiday labels do not enter the predictor, avoiding an arbitrary choice among overlapping text labels.

Seventy-nine timestamp groups contain a conflict in at least one selected numeric field. After removing those groups, collapsing identical selected records removes 7,522 duplicate rows. A declared numeric screen then removes eleven unique hours: temperature must be between 200 and 330 kelvin, hourly rain and snow between zero and one hundred millimeters, cloud cover between zero and one hundred percent, and traffic nonnegative. These bounds are a reproducible screen, not a claim that all remaining measurements are error-free. Missing or conflicting values are not imputed.

All retained observations through 2016 form the fitting data, totaling 25,240 hours. OLS predicts same-hour traffic volume using hour-of-day, weekday, and month indicators, temperature in Celsius, log-one-plus rain, log-one-plus snow, and cloud percentage. The design has forty-five parameters and full numerical rank. Its fitting R-squared is 0.8344880. This is a description of the earlier fit, not a held prediction score or evidence that the weather coefficients identify causal effects.

For the later comparison, both years are restricted to January through September because the file ends in September 2018. The candidate set contains 13,045 hours across 546 dates. We retain only dates with twenty-four unique valid hours, leaving 519 dates: 258 in 2017 and 261 in 2018. Twenty-seven partial dates are excluded. The main scored data therefore contain 12,456 hours. Each complete date receives equal weight in the outcome regressions, rather than weighting dates by the original number of duplicated weather rows.

## 4. The research model and error definitions

For every later hour, signed error equals predicted traffic minus observed traffic. Absolute error removes its sign. Daily outcomes average each quantity across the same twenty-four valid hours. A negative daily signed value indicates net underprediction; a positive one indicates net overprediction. The daily absolute outcome measures error magnitude in vehicles per hour. These constructions prevent a change in the mix of observed hours within partial dates from silently becoming the main comparison.

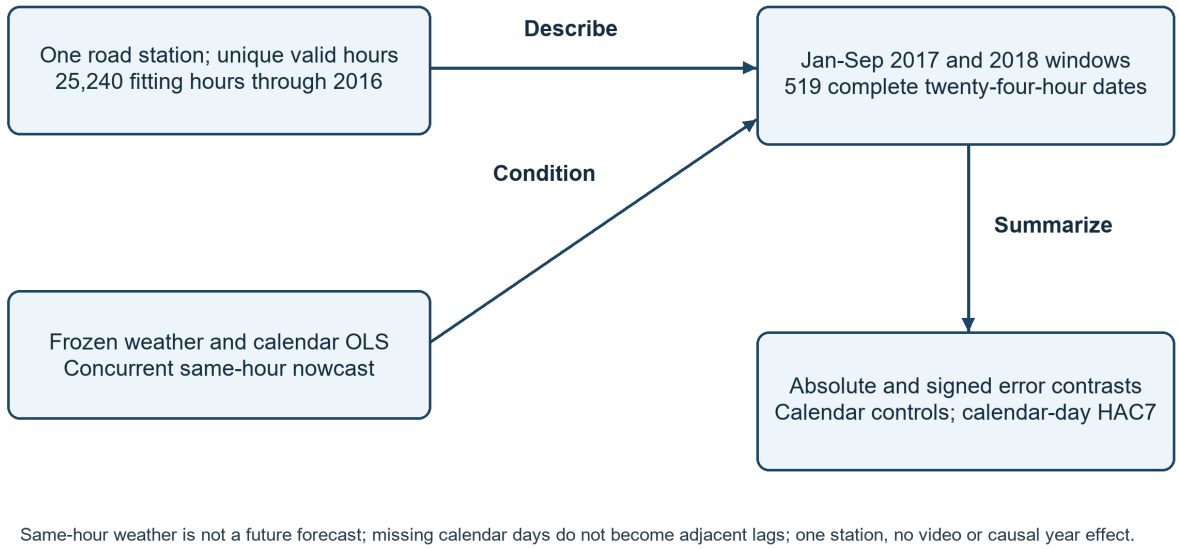



*Figure 1. Earlier-only fitting, matched later windows, and two daily error outcomes.*

*Figure 1 separates the frozen numerical predictor from the temporal comparison. Each daily outcome is regressed on a 2018 indicator plus month and weekday indicators. The year coefficient compares the two observed windows after those calendar adjustments. It is not a treatment effect and does not identify an omitted road event. Structured validation principles Roberts et al., 2017 motivate retaining this temporal separation while acknowledging that the site, sensor, model specification, and source collection remain shared.*

The loss choice determines what stability means. Forecast-error measurement work Hyndman and Koehler, 2006 and point-forecast evaluation Gneiting, 2011 emphasize matching the summary to the predictive object. Discussions by Willmott and Matsuura, 2005, Chai and Draxler, 2014, and Hodson, 2022 compare properties of absolute and squared-error summaries. We report absolute magnitude and signed direction because they answer the stated question. No percentage error is used, and no root-mean-square result is inferred from the available absolute averages. Signed bias is also not a probability-calibration statistic.

The two outcomes are intentionally not collapsed into a composite. A small average signed error can result from cancellation of substantial positive and negative hourly errors. A stable MAE can likewise coexist with a movement in directional imbalance. Comparing their coefficients makes those possibilities

visible. The regression does not establish which traffic conditions contribute most to the signed movement. Such a subgroup or mechanism analysis would require its own specified variables and uncertainty assessment.

## 5. Calendar-aware uncertainty with gaps

Daily errors can be serially dependent. We therefore use a Bartlett-weighted heteroskedasticity-and-autocorrelation covariance with seven calendar-day lags, following the covariance-estimation motivation of Newey and West, 1987. The date grid includes missing and excluded dates with zero score contributions. This avoids treating the last retained date in September 2017 and the first in January 2018 as adjacent daily observations. Their ninety-three-day gap is not compressed into one lag merely because the outcome table omits the intervening months.

The implementation first confirms that the same score-based covariance construction, without calendar gaps, agrees numerically with the ordinary statsmodels HAC calculation. It then uses the complete calendar grid for the reported covariance. A finite-sample multiplier uses the ratio of N to N minus the parameter count. Intervals use the residual t reference with 503 degrees of freedom. Robust covariance reasoning White, 1980 and dependence guidance Cameron and Miller, 2015 support stating these details; they do not guarantee correct inference under every possible long-range dependence process.

Zero-padding the score grid is not an imputation of zero traffic or zero prediction error. Excluded dates contribute no fitted observation and no score to the covariance sum. The sample size remains 519, even though the calendar grid spans 638 dates. This construction preserves actual day separation when applying lag weights, rather than creating fictitious outcomes for the missing interval. Dependence lasting beyond seven days can still matter, and no claim is made that the chosen lag cutoff is universally sufficient. The covariance is a stated working summary for this analysis, not a replacement for additional independent time periods or stations.

The fitting model remains fixed when estimating outcome uncertainty. Its training-sample uncertainty is not propagated. Only one earlier fit and two later windows are examined. Forecast-comparison tests associated with Diebold and Mariano, 1995 and subsequent methodological discussion Diebold, 2015 concern other comparison settings; our two outcome regressions are temporal diagnostic contrasts for one model, not a competition between independently trained forecasting systems. Their intervals should not be repurposed as an equivalence test or a guarantee about another road site.

## 6. Absolute error is similar while signed error moves

Daily mean absolute error averages 569.4509325 vehicles per hour in the 2017 window and 566.6352225 in 2018. These unadjusted means are close. After month and weekday adjustment, the 2018 coefficient is positive 1.4243933, with standard error 13.3088287 and interval from negative 24.7233481 to positive 27.5721347. The sign differs from the small raw mean difference because the regression adjusts calendar composition. The interval does not establish equivalence; it describes limited precision around this adjusted contrast.

Signed error has a different pattern. Its unadjusted daily mean is negative 68.1977820 in 2017 and positive 0.2201078 in 2018. The adjusted 2018 coefficient is 66.7158196, with standard error 28.5832459 and interval from 10.5585620 to 122.8730772. Thus directional underprediction diminishes in the observed later window while average absolute error remains similar. That finding answers the monitoring question without requiring the overall error magnitude to rise.

*Table 1. Calendar-adjusted 2018 contrasts for complete-day nowcast errors*

| Model | Term | Estimate (SE) | 95% CI | N |
|---|---|---|---|---|
| M1 | 2018 minus 2017 window | 1.4244 (13.3088) | [-24.7233, 27.5721] | 519 |
| M1: fit | R-squared = 0.6191 | - | - | 519 |
| M2 | 2018 minus 2017 window | 66.7158 (28.5832) | [10.5586, 122.8731] | 519 |
| M2: fit | R-squared = 0.1105 | - | - | 519 |

Model labels: M1 = 2018 versus 2017 daily absolute error, calendar adjusted; M2 = 2018 versus 2017 daily signed prediction error, calendar adjusted.

*Table 1 reports OLS models on the same 519 complete dates, controlling for month and weekday indicators. M1 uses daily mean absolute error; M2 uses daily mean predicted-minus-observed error. Covariance is calendar-day Bartlett HAC7, with zero score on unobserved dates, finite-sample correction, and t503 intervals. The displayed coefficient is the 2018 indicator. R-squared is 0.6191 for M1 and 0.1105 for M2. No road fixed effects, intervention variables, or causal identification design are included.*

The baseline also has an important practical limitation. It produces 443 negative hourly predictions among the 12,456 hours on complete scored dates. Predictions are not clipped after inspecting this issue. An unrestricted linear mean model is not guaranteed to respect the nonnegative traffic support, and the negative values help explain why its numerical diagnostic should not be mistaken for a deployment recommendation. Altering the predictor to enforce support would constitute a new model comparison rather than a cosmetic correction to these frozen outputs.

## 7. What temporal monitoring can conclude

The result suggests tracking error direction alongside error magnitude. It does not show that calendar time itself improved the model or that a particular environmental mechanism changed. Holiday patterns, road activity, sensor behavior, model misspecification, and unobserved conditions remain possible contributors. Month and weekday indicators adjust two measured compositional features, not all differences between years. Stable absolute error therefore supports only a narrow statement about this metric on the declared complete-day cohort.

Complete-day selection is another part of that boundary. It ensures a common twenty-four-hour support for the daily summaries, but can exclude dates with unusual recording failures or conflicting measurements. The study does not observe what the missing hours would have contributed. Comparing a partial-day rule, inverse-observation weighting, or an explicit missingness model would require additional assumptions and another executed sensitivity analysis. The present result concerns observed complete dates and should retain that qualification when describing temporal transport.

Practical forecasting guidance Goodwin and Lawton, 1999 and broad benchmark experience such as the M3 study Makridakis and Hibon, 2000 provide context for evaluating more than one temporal summary. They do not add replications to this dataset. A fuller transport study could include repeated fitting origins, alternative nonnegative predictors, additional stations, and predeclared drift alarms. It should also assess the effect of excluding partial dates, which may select relatively complete conditions. None of those extensions is described here as already executed.

The source and analysis record preserve the UCI identifier, original ZIP checksum, duplicate and conflict decisions, bounds, fitting dates, hourly predictions, complete-date membership, and the covariance

construction. The dataset is documented under CC BY 4.0 with attribution to John Hogue. Dataset documentation Gebru et al., 2021 and FAIR stewardship Wilkinson et al., 2016 motivate keeping the physical-hour construction visible. A model-monitoring conclusion is more reviewable when a reader can see which observations represent distinct traffic hours and which have been excluded.

The central finding is a difference between two notions of temporal stability. A frozen same-hour traffic predictor has a small, uncertain absolute-error contrast between the matched later windows, while its signed error shifts toward zero. Neither summary should stand in for the other. Keeping their definitions, calendar adjustment, and observed-data boundaries together yields a clearer account of what the model carried forward from its earlier fitting period.

## References


1. 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.
2. 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.
3. 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.
4. Goodwin, Paul; Lawton, Richard. (1999). On the asymmetry of the symmetric MAPE. International Journal of Forecasting. https://doi.org/10.1016/s0169-2070(99)00007-2.
5. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
6. 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.
7. Makridakis, Spyros; Hibon, Michèle. (2000). The M3-Competition: results, conclusions and implications. International Journal of Forecasting. https://doi.org/10.1016/s0169-2070(00)00057-1.
8. Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.
9. Willmott, CJ; Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. https://doi.org/10.3354/cr030079.
10. Hafner, Danijar, et al. "Mastering Diverse Domains through World Models." *arXiv preprint arXiv:2301.04104*, 2023. https://arxiv.org/abs/2301.04104.
11. 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.
12. 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.
13. Li, Zhiqi; Wang, Wenhai; Li, Hongyang; Xie, Enze; Sima, Chonghao; Lu, Tong; Yu, Qiao; Dai, Jifeng. (2022). BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. arXiv:2203.17270. https://arxiv.org/abs/2203.17270.
14. 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.
15. Chai, T.; Draxler, R. R.. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geoscientific Model Development. https://doi.org/10.5194/gmd-7-1247-2014.

16. Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis. https://doi.org/10.1016/j.csda.2017.11.003.

17. Tashman, Leonard J.. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting. https://doi.org/10.1016/s0169-2070(00)00065-0.

18. Hodson, Timothy O.. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development. https://doi.org/10.5194/gmd-15-5481-2022.

19. Hu, Yihan; Yang, Jiazhi; Chen, Li; Li, Keyu; Sima, Chonghao; Zhu, Xizhou; Chai, Siqi; Du, Senyao; Lin, Tianwei; Wang, Wenhai; Lu, Lewei; Jia, Xiaosong; Liu, Qiang; Dai, Jifeng; Qiao, Yu; Li, Hongyang. (2022). Planning-oriented Autonomous Driving. arXiv:2212.10156. https://arxiv.org/abs/2212.10156.

20. Gneiting, Tilmann. (2011). Making and Evaluating Point Forecasts. Journal of the American Statistical Association. https://doi.org/10.1198/jasa.2011.r10138.

21. 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.

22. 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.

23. 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.

24. 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.

25. Diebold, Francis X.. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.2014.983236.

26. Hogue, John. (2019). Metro Interstate Traffic Volume [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5X60B.

27. 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.

28. 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.

29. Caesar, Holger, et al. "nuScenes: A multimodal dataset for autonomous driving" *arXiv preprint arXiv:1903.11027*, 2019. https://arxiv.org/abs/1903.11027.

30. Diebold, Francis X.; Mariano, Roberto S.. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599.