Benchmarking Chronos-2 as a Foundation Model for Traffic Forecasting

Pablo Moreno Moreu[1⋆][0009−0007−8099−5464], Arnau Neches-Vilá[1][0009−0007−5727−4562], María Martínez-Ballesteros[1,2][0000−0003−3160−7414], and José C. Riquelme[1,2][0000−0002−8243−2186]

[1] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Av. Reina Mercedes S/N, 41012 Sevilla, Spain
{pmoreno7,aneches,mariamartinez,riquelme}@us.es
[2] SCORE Lab, I3US Institute, Universidad de Sevilla, Sevilla, Spain

Abstract. Foundation models for time series forecasting, such as Chronos-2, offer a compelling zero-shot alternative to specialized forecasting architectures. Their ability to perform multivariate inference without dataset-specific training makes them particularly attractive for traffic forecasting, where sensor networks are large, dynamic, and expensive to retrain on. In this work, we systematically evaluate Chronos-2 on established traffic benchmarks and compare it against classical methods, recurrent neural networks, and state-of-the-art Graph Neural Network–based models across the standard forecasting horizons of 3, 6, and 12 steps. Chronos-2 is evaluated in a direct whole-matrix setting, where all sensors are provided simultaneously as a single multivariate context, allowing the model to exploit cross-sensor dependencies without explicit graph construction or fine-tuning. Our results show that Chronos-2 achieves competitive yet generally non-leading performance: it substantially improves over classical baselines and, in several cases, approaches or matches older graph-based models, while recent supervised Graph Neural Networks remain stronger in terms of Mean Absolute Error and Mean Absolute Percentage Error. Interestingly, Chronos-2 obtains highly competitive Root Mean Squared Error values on several datasets, suggesting that foundation models can capture useful global temporal structure in traffic networks even without task-specific optimization. These findings position Chronos-2 as a strong zero-shot baseline for traffic forecasting, while also highlighting the limitations of direct whole-matrix inference for larger sensor networks. We therefore identify scalable inference strategies, such as graph-informed node batching, as a promising direction for future work.

Keywords: Traffic forecasting · Foundation models · Chronos-2 · Zero-shot forecasting · Graph neural networks · Multivariate time series · PEMS.

⋆ Corresponding author.

1 Introduction

Traffic-flow prediction is a canonical spatio-temporal forecasting task with direct real-world impact on congestion management, urban planning, and intelligent transportation systems. Over the last several years, the dominant paradigm for this task has been based on Graph Neural Networks (GNNs), which explicitly model the road network and learn to propagate information across spatially correlated sensors [12,13,18,20]. Models such as DCRNN, STGCN, Graph WaveNet, and DGCRN have achieved strong performance on widely used benchmarks such as PEMS-BAY, METR-LA, PEMS03, PEMS04, PEMS07, and PEMS08 by jointly learning spatial dependencies and temporal dynamics. However, these models are typically trained separately for each dataset, often require graph-specific design choices or adjacency matrices, and may need to be retrained when the sensor network changes.

At the same time, large pre-trained foundation models for time series forecasting have emerged as a promising zero-shot alternative. These models are trained on diverse collections of time series and can be applied to new forecasting tasks without dataset-specific training or fine-tuning. Chronos-2 [1] is one such model: a transformer-based architecture pre-trained on a large collection of real and synthetic time series, with native support for multivariate forecasting through its variable dimension and group attention mechanism. This makes Chronos-2 especially attractive for traffic forecasting, where sensor networks are multivariate by nature and where retraining specialized models can be costly. Nevertheless, its performance on standard large-scale traffic benchmarks remains insufficiently characterized.

This study addresses the following question:

> *How competitive is Chronos-2, used as a zero-shot multivariate forecaster, against classical, recurrent, and state-of-the-art graph-based traffic forecasting models on standard traffic benchmarks?*

To answer this question, we evaluate Chronos-2 in a direct *whole-matrix* setting, where all sensors in a traffic network are provided simultaneously as a single multivariate context. In this setup, Chronos-2 receives the full historical sensor matrix and produces multi-step forecasts without task-specific training, fine-tuning, explicit graph construction, or hyperparameter tuning. We compare its performance against a broad set of baselines across six widely used traffic datasets and the standard forecasting horizons of 3, 6, and 12 steps.

The main contribution of this work is a systematic benchmark of Chronos-2 for zero-shot traffic forecasting. Our results show that Chronos-2 is competitive with several established baselines and, in some cases, approaches or matches older graph-based models, particularly in terms of RMSE. However, recent supervised GNN-based architectures remain stronger in terms of Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) on most datasets. These findings suggest that Chronos-2 should be understood not as a replacement for specialized traffic models, but as a strong zero-shot baseline that can be deployed without dataset-specific training.

The remainder of this paper is organized as follows. Section 2 reviews relevant spatio-temporal GNNs and foundation models. Section 3 presents our Chronos benchmarking framework, including four inference modes. Section 4 describes the experimental setup (datasets, evaluation protocol, and baselines) and reports the comparison of Chronos-2 against state-of-the-art GNN models across all benchmarks and forecast horizons. Finally, Section 5 summarizes our findings and outlines directions for future work.

2 Background

This section reviews the two families of models relevant to our study: spatio-temporal graph neural networks and pre-trained foundation models for time series.

2.1 GNN-Based Traffic Forecasting

Spatio-temporal GNNs have become the standard approach for multi-step traffic forecasting, combining both spatial and temporal relations into their predictions. Graph WaveNet [18] learns an adaptive adjacency matrix, reducing dependence on pre-specified road graphs. DGCRN [12] combines diffusion convolution with a dynamic temporal graph for modeling spatial relationships. STGCN [20] replaces recurrence with gated temporal convolutions stacked between graph convolution layers. All of these approaches require dataset-specific training and often dataset-specific adjacency matrix tuning, introducing another set of hyperparameters and long training times. Moreover, their training is specific to a graph, and thus struggles to generalize when new nodes are introduced to the network, requiring expensive retraining on datasets which might not be yet available.

2.2 Foundation Models for Time Series

Time Series Foundation Models (TSFMs) have rapidly emerged as a competitive zero-shot alternative to dataset-specific deep learning models. Several families have been proposed in the last two years: TimeGPT [8], TimesFM [5], Chronos and its successor Chronos-2 [1, 2], Moirai and Moirai-2 [14, 17], MOMENT [9] and Toto [4], among others. These models differ in their tokenization strategy (patch-based versus token-based), training corpora and support for multivariate or covariate-informed forecasting, but they share the same operating principle: a single forward pass on a previously unseen series produces a forecast without any task-specific training. Among them, Chronos-2 is one of the most recent and the only model in this family that natively supports multivariate forecasting through a dedicated *group attention* mechanism, which makes it a natural candidate for evaluation on multi-sensor traffic networks. Recent work has begun to benchmark TSFMs systematically, both on general-purpose forecasting suites

and in specialized domains, including zero-shot benchmarks of the Chronos family against classical baselines [16]; their application to road-traffic forecasting, however, remains comparatively under-explored.

Chronos [2] is a family of pretrained time series forecasting models. Chronos-2 [1] adopts an encoder-only transformer architecture where time series are first normalized using a robust scaling scheme and then divided into non-overlapping patches, which are mapped to continuous embeddings. Furthermore, instead of predicting single values, Chronos-2 produces direct multi-step quantile forecasts via a quantile regression head, offering a richer representation of predictive uncertainty and better coverage of extreme events. A major advancement in Chronos-2 is its native support for multivariate and covariate-informed forecasting. This is achieved through a group attention mechanism that alternates with standard time attention layers. Group attention facilitates in-context learning by efficiently sharing information across multiple time series within a dynamically defined group. By sharing information across the batch axis rather than concatenating sequences, it scales gracefully with high-dimensional data, achieving $\mathcal{O}(V)$ memory scaling, where V is the total number of variates (both prediction targets and covariates). Crucially, Chronos-2 functions entirely in a zero-shot manner at inference time. It generates accurate predictions for new, unseen datasets without requiring any task-specific training, adaptation or fine-tuning, establishing it as a highly versatile, general-purpose forecaster capable of being deployed "as is" in real-world production pipelines.

3 Methodology: Chronos evaluation framework

This section describes the framework we developed to apply Chronos-2 to traffic sensor networks. We implement four distinct inference modes that explore different ways of presenting the multivariate context to the model, ranging from fully independent univariate forecasts to the full network as a single multivariate batch. Each mode trades off spatial context, memory footprint and inference time differently. The specific mode used for the benchmark reported in this paper is indicated in Section 4.2.

3.1 Inference Modes

We implement four distinct inference modes for Chronos-2 on traffic graphs:

`single_node` Each of the N sensors is predicted independently as a univariate series. No spatial information is used. Chronos-2 receives a context tensor of shape $[1, 1, L]$ per node, making N separate forward passes. Processing each node individually dismisses the multivariate capability of Chronos, resulting in a loss of information that yields worse results. Moreover, Chronos-2 scales better across the variable dimension (number of covariates it takes as input) than across the time dimension, meaning this prediction paradigm is actually slower.

`whole_matrix` All N sensors are presented simultaneously as variates of one multivariate series. Chronos-2 receives a single context tensor of shape $[1, N, L]$. This maximizes spatial context but becomes computationally intractable for large N, due both to memory costs and rising model complexity. In larger datasets, this also offloads all spatial relationship computation to the model, which hurts the performance.

`adj_neighbor` For each target node i, its context includes the node itself plus all its immediate neighbors under the chosen adjacency matrix $\mathbf{A}$, used as informative covariates. The prediction for node i is extracted from the first variate of the output. This recovers local spatial context, although it is limited to the immediate neighborhood. As with single node prediction, Chronos-2's memory scaling means this is slower than predicting the entire network.

`node_batches` Nodes are partitioned into disjoint batches $\mathcal{B} = \{B_1, B_2, \ldots, B_K\}$ and each batch is forecast jointly. Chronos-2 receives $[1, |B_k|, L]$ for batch B_k, requiring $K \ll N$ forward passes while retaining intra-batch spatial context. It attempts to replicate Chronos-2's own capacity for grouping together series for better prediction, offloading this task from the model into pre-computed statistical measures. It maintains similar results as utilizing the entire network while reducing the memory footprint.

4 Experiments and Results

This section presents the empirical evaluation of Chronos-2 on six standard traffic benchmarks. We first describe the datasets and evaluation protocol, and then report the comparison against state-of-the-art GNN baselines across all forecast horizons and inference modes.

4.1 Datasets

We evaluate on six widely used traffic speed/flow datasets summarized in Table 1. Data is collected by the Caltrans Performance Measurement System (PeMS), a network of traffic sensors in the state of California. They are divided by geographic area and by feature (traffic speed, flow, occupancy). Sensors take measurements every 30 seconds and each data point is the average over 5 minutes.

Each dataset includes a pre-computed road-network adjacency matrix obtained by applying a thresholded Gaussian Kernel to pairwise sensor distance [13] (`adj_mx`), which is used by the supervised GNN baselines included in our comparison.

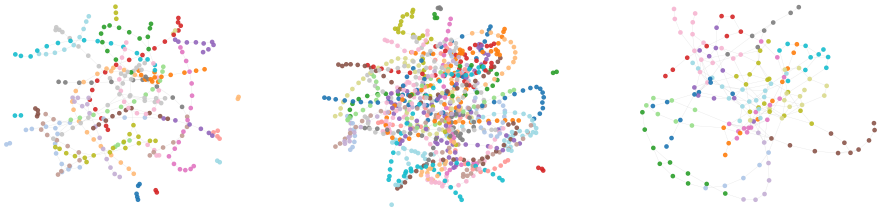
Applying community detection methods such as Louvain or spectral clustering to the purely distance-based adjacency matrix recovers the underlying road-network structure, as illustrated in Figure 1. This reveals densely intercommunicated node groups, suggesting a natural decomposition of the sensor network into spatially coherent batches that could be exploited by scalable inference strategies, as discussed in Section 5.

Table 1: Characteristics of Traffic Datasets.

Dataset	Nodes (N)	Timesteps (T)	Interval	Feature
PEMS-BAY	325	∼52,116	5 min	Speed (mph)
METR-LA	207	∼34,272	5 min	Speed (mph)
PEMS03	358	∼26,208	5 min	Flow, occupancy
PEMS04	307	∼16,992	5 min	Flow, occupancy
PEMS07	883	∼28,224	5 min	Flow, occupancy
PEMS08	170	∼17,856	5 min	Flow, occupancy



(a) PEMS-BAY. (b) METR-LA. (c) PEMS03.



(d) PEMS04. (e) PEMS07. (f) PEMS08.

Fig. 1: Spatial distribution of traffic sensor nodes for the six benchmark datasets.

4.2 Evaluation Protocol

Following the standard protocol for these benchmarks, we use a rolling-window evaluation with a stride of 1 time step over the test split. All datasets follow a 6:2:2 split save for PEMS-BAY and METR-LA, which follow 7:1:2, as done in all comparisons utilised. For each evaluation window:

1. A context of $L = 2016$ (1 week) at 5-minute resolution is presented to the model.
2. The model produces predictions for horizons $h \in \{3, 6, 12\}$ steps (15, 30, 60 minutes).
3. Predictions are compared against ground truth using MAE, RMSE, and MAPE. As Chronos-2 produces quantiles, we use the median as value.

Results are averaged over all evaluation windows in the test split as is common for this task, and we also compare across the key horizons of 3, 6 and 12

(15, 30 and 60 minutes) for the most common datasets. It is standard to use averages on the PEMS03, PEMS04, PEMS07 and PEMS08 datasets, whereas PEMS-BAY and METR-LA are often compared across each horizon. We use the whole_matrix prediction method for a fairer comparison. We consider this choice to be fairer given that other networks are trained and evaluated on the entire dataset/matrix. Employing a non-standard inference method on Chronos-2 would reduce the usefulness of the comparison of the zero-shot, non-modified Chronos-2 against other models. To complement the accuracy evaluation, we also report the total inference time required to process each benchmark.

4.3 Hardware and software environment.

All experiments were conducted on a workstation equipped with an Intel Core i9-10900x CPU running at 3.70 GHz, 64 GB of RAM and an NVIDIA GeForce RTX 2080-ti GPU. Chronos-2 inference is executed in `bfloat16` precision on CUDA using the publicly released Chronos-2 checkpoint; no parameter of the model is modified at any point. The complete code and experimental scripts are available at `https://github.com/PabloMMRuiz/minerva`.

4.4 Results and Discussion

The results show that Chronos-2 is consistently competitive as a zero-shot forecaster, although it does not generally outperform specialized supervised traffic forecasting models. This is expected, since models such as DGCRN, STD-MAE, GAMMA-Net, Graph WaveNet or DCRNN are trained directly on each target dataset and often exploit graph-structured inductive biases, whereas Chronos-2 is applied without any dataset-specific training or fine-tuning, using the full sensor matrix as a multivariate input. The computational cost of this whole-matrix setting is reported in Table 5. Inference time increases with both the number of evaluation windows and the number of sensors: PEMS07 is the most expensive dataset, requiring 9762.07 seconds due to its 883 sensors, while PEMS08 is the fastest despite having a comparable number of evaluation windows to PEMS04, mainly because it contains only 170 nodes. These results indicate that whole-matrix inference is feasible for the evaluated benchmarks, but may become a practical bottleneck for larger traffic networks. On the PEMS03, PEMS04, PEMS07 and PEMS08 benchmarks, Chronos-2 obtains results close to recent state-of-the-art models, especially in terms of Root Mean Squared Error (RMSE), as shown in Table 2. On PEMS03, Chronos-2 achieves an RMSE of 21.67, improving over GAMMA-Net and STD-MAE, which report 22.11 and 24.43 respectively, while on PEMS04 it obtains the lowest RMSE among the reported methods, with 28.23. However, this advantage does not consistently translate to MAE or MAPE, where specialized models generally remain better. For instance, on PEMS03, Chronos-2 obtains 14.04 MAE compared with 13.72 for GAMMA-Net, and on PEMS08 it obtains 14.89 MAE compared with 13.17 for GAMMA-Net. This suggests that Chronos-2 is able to capture the global

temporal structure of traffic series and avoid very large squared errors in some datasets, but it may be less precise in terms of average absolute deviation.

Table 2: Performance comparison on PEMS03, PEMS04, PEMS07, and PEMS08 datasets.

Model	**PEMS03**			**PEMS04**			**PEMS07**			**PEMS08**		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
ARIMA	35.31	47.59	33.78	33.73	48.80	24.18	38.17	59.27	19.46	31.09	44.32	22.73
VAR	23.65	38.26	24.51	23.75	36.66	18.09	75.63	115.24	32.22	23.46	36.33	15.42
SVR	21.97	35.29	21.51	28.70	44.56	19.20	32.49	50.22	14.26	23.25	36.16	14.64
LSTM	21.33	35.11	23.33	27.14	41.59	18.20	29.98	45.84	13.20	22.20	34.06	14.20
GWNet [18]	19.85	32.94	19.31	25.45	39.70	17.29	26.85	42.78	12.12	19.13	31.05	12.68
DCRNN [13]	18.18	30.31	18.91	24.70	38.12	17.12	25.30	38.58	11.66	17.86	27.83	11.45
STGCN [20]	17.49	30.12	17.15	22.70	35.55	14.59	25.38	38.78	11.08	18.02	27.83	11.40
MSTA-GNN [19]	15.92	27.02	14.86	19.33	31.41	12.88	21.73	35.05	9.54	15.77	24.98	9.97
TGLRN [11]	15.15	26.49	14.47	18.84	31.03	12.39	20.58	34.49	8.66	15.28	24.58	9.78
DSTGA-MAMBA [3]	14.25	23.04	*13.57*	19.02	30.74	12.32	19.69	33.67	8.62	14.66	24.13	9.64
STD-MAE [7]	*13.80*	24.43	13.96	**17.80**	*29.25*	*11.97*	*18.65*	31.44	**7.84**	*13.44*	*22.47*	**8.76**
GAMMA-Net [10]	**13.72**	*22.11*	**13.39**	*17.90*	29.53	**11.62**	**18.49**	**30.79**	*8.08*	**13.17**	22.53	*9.04*
Chronos-2	14.04	**21.67**	13.70	18.85	**28.23**	13.13	21.22	*32.91*	9.00	14.89	**22.45**	9.52

The METR-LA results show a clearer gap between Chronos-2 and the strongest supervised baselines, particularly as the forecasting horizon increases, as reported in Table 3. At the 3-step horizon, Chronos-2 obtains 3.31 MAE, while GAMMA-Net reports 2.39. At 6 steps, Chronos-2 obtains 3.95 MAE compared with 2.60 for GAMMA-Net, and at 12 steps the gap increases to 4.72 versus 2.87. This indicates that, on METR-LA, Chronos-2 struggles to match specialized graph-based models at longer horizons. Nevertheless, it remains competitive with older classical and recurrent baselines such as ARIMA, VAR, SVR and LSTM, which is notable given that no training is performed on METR-LA.

On PEMS-BAY, Chronos-2 is closer to the supervised graph models, as shown in Table 4: at the 3-step horizon it obtains 1.46 MAE, compared with 1.12 for GAMMA-Net and 1.21 for CITRUS; at 6 steps it obtains 1.71 MAE, worse than the best supervised models but comparable to DCRNN and Graph WaveNet; and at 12 steps it reaches 1.95 MAE, matching Graph WaveNet and outperforming DCRNN and STGCN in MAE. Its RMSE is also competitive at longer horizons, with 3.78 at 12 steps, close to GAMMA-Net's 3.67 and better than several established baselines.

Overall, the results support three main observations. First, Chronos-2 should not be interpreted as a replacement for specialized supervised GNNs when maximum accuracy is required, since recent traffic-specific architectures still achieve the best MAE and MAPE in most cases. Second, Chronos-2 provides a strong zero-shot baseline that is often substantially better than classical statistical and generic neural baselines, and in some cases approaches or matches older

Table 3: Performance comparison on METR-LA across horizons.

Model	**3 steps**			**6 steps**			**12 steps**		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
ARIMA	3.99	8.21	9.60	5.15	10.45	12.70	6.90	13.23	17.40
VAR	4.42	7.89	10.20	5.41	9.13	12.70	6.52	10.11	15.80
SVR	3.99	8.45	9.30	5.05	10.87	12.10	6.72	13.76	16.70
LSTM	3.44	6.30	9.60	3.77	7.23	10.90	4.37	8.69	13.20
GWNet [18]	2.69	5.15	6.90	3.07	6.22	8.37	3.53	7.37	10.01
DCRNN [13]	2.77	5.38	7.30	3.15	6.45	8.80	3.60	7.60	10.50
STGCN [20]	2.88	5.74	7.62	3.47	7.24	9.57	4.59	9.40	12.70
DGCRN [12]	*2.62*	*5.01*	*6.63*	2.99	6.05	8.02	3.44	7.19	9.73
CITRUS [6]	2.70	5.14	6.74	*2.98*	*5.90*	**7.78**	3.44	*6.85*	*9.28*
STD-MAE [7]	*2.62*	5.02	6.70	2.99	6.07	8.04	*3.40*	7.07	9.59
GAMMA-Net [10]	**2.39**	**4.54**	**6.22**	**2.60**	**5.20**	*7.06*	**2.87**	**5.99**	**8.16**
Chronos-2	3.31	5.90	8.16	3.95	6.89	9.94	4.72	7.98	12.03

Table 4: Performance comparison on PEMS-BAY across horizons.

Model	**3 steps**			**6 steps**			**12 steps**		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
ARIMA	1.62	3.30	3.50	2.33	4.76	5.40	3.38	6.50	8.30
VAR	1.74	3.16	3.60	2.32	4.25	5.00	2.93	5.44	6.50
SVR	1.85	3.59	3.80	2.48	5.18	5.50	3.28	7.08	8.00
LSTM	2.05	4.19	4.80	2.20	4.55	5.20	2.37	4.96	5.70
GWNet [18]	1.30	2.74	2.73	1.63	3.70	3.67	1.95	4.52	4.63
DCRNN [13]	1.38	2.95	2.90	1.74	3.97	3.90	2.07	4.74	4.90
STGCN [20]	1.36	2.96	2.90	1.81	4.27	4.17	2.49	5.69	5.79
DGCRN [12]	1.28	2.69	2.66	1.59	3.63	3.55	1.89	4.42	4.43
CITRUS [6]	*1.21*	*2.61*	*2.51*	*1.48*	*3.28*	*3.23*	1.78	3.99	*4.08*
STD-MAE [7]	1.23	2.62	2.56	1.53	3.53	3.42	*1.77*	4.20	4.17
GAMMA-Net [10]	**1.12**	**2.28**	**2.31**	**1.34**	**2.97**	**2.91**	**1.59**	**3.67**	**3.61**
Chronos-2	1.46	2.81	3.20	1.71	3.34	3.93	1.95	*3.78*	4.71

Table 5: Inference time.

Dataset	Evaluation Windows	Time taken (s)
PEMS-BAY	10412	6620.18
METR-LA	6843	2638.62
PEMS03	5230	3633.73
PEMS04	3387	2028.60
PEMS07	5633	9762.07
PEMS08	3560	1111.94

graph-based models. Third, its competitive RMSE values suggest that foundation models may already capture useful global temporal patterns across large multivariate traffic systems, even without explicit access to the road graph or dataset-specific optimization. From a practical perspective, the main advantage of Chronos-2 is not state-of-the-art accuracy, but the ability to produce reasonable forecasts without training, hyperparameter search, graph construction, or dataset-specific model design, making it attractive when fast deployment, lack of historical training data, changing sensor networks, or limited training resources are more important than achieving the absolute lowest error.

However, the evaluation also highlights important limitations. Since the assessed datasets are public and widely used, possible exposure during pretraining cannot be ruled out. In addition, the whole-matrix setting may become challenging for larger sensor networks, as all nodes are passed simultaneously as multivariate context. Therefore, the present results should be interpreted as a benchmark of Chronos-2 under direct multivariate zero-shot inference, while scalability strategies such as graph-informed batching remain an open direction for future work.

5 Conclusions and Future Work

This paper presents a benchmarking study of Chronos-2, a pre-trained multivariate foundation model, on traffic forecasting. Across the six evaluated datasets, Chronos-2 produces competitive although not leading results: the performance gap to fully supervised GNNs is small in absolute terms (approximately 0.34 MAE on PEMS-BAY at the 15-minute horizon), and Chronos-2 attains consistently low RMSE values across benchmarks, suggesting that it is less sensitive to anomalous observations than several specialized models even when its average MAE is slightly worse. Beyond raw accuracy, the ease of deployment of foundation models, with zero-shot inference and no need for retraining when the sensor network changes, makes them an attractive option for real-world traffic-forecasting applications. We argue that Chronos-2 establishes itself as a credible reference point for new traffic-forecasting datasets, substantially more informa-

tive than the classical statistical baselines (ARIMA, VAR, SVR) traditionally used in this role.

As future work, and as with any benchmark of a foundation model, evaluation on datasets released before the model's training cut-off carries the risk of data saturation, since the target series may have been seen during pre-training. Chronos-2 is trained on data from diverse sources. While all multivariate datasets are synthetic, the univariate training data includes LargeST [15], a traffic dataset that partially overlaps with the evaluated datasets (PEMS-BAY, PEMS03, PEMS04, and PEMS07). We observe no evidence of data saturation, and performance on these datasets is consistent with that on previously unseen datasets such as METR-LA. For a cleaner comparison, we plan to evaluate Chronos-2 on traffic datasets released after October 2025, the release date of Chronos-2. Moreover, a wider survey on less standardized datasets from outside the US will also help mitigate the possibility of saturation. We are also interested in pushing the limits of Chronos-2's group attention mechanism by performing inference on larger sensor networks with additional covariates, where the memory footprint of the multivariate context becomes the main bottleneck. Even though Chronos-2 exhibits linear memory scaling with the number of variates, very large networks may still become intractable in practice. We therefore propose exploring *graph-informed node batching* as a scalable inference strategy that preserves multivariate context within spatially coherent groups of sensors while reducing the overall memory footprint and enabling straightforward Single Instruction, Multiple Data (SIMD) parallelisation.

References

1. Ansari, A.F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S.S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D.C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P.M., Wang, H., Rangwala, H., Karypis, G., Wang, Y., Bohlke-Schneider, M.: Chronos-2: From univariate to universal forecasting (2025), https://arxiv.org/abs/2510.15821
2. Ansari, A.F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D.C., Mahoney, M.W., Torkkola, K., Wilson, A.G., Bohlke-Schneider, M., Wang, Y.: Chronos: Learning the language of time series. Transactions on Machine Learning Research (TMLR) (2024), https://arxiv.org/abs/2403.07815
3. Chen, L., Wu, Q.: Dstga-mamba: a disentangled spatio-temporal graph attention mamba model for traffic flow prediction. Scientific Reports **16**, 2243 (2026). https://doi.org/10.1038/s41598-025-31990-3, https://doi.org/10.1038/s41598-025-31990-3
4. Cohen, B., et al.: This time is different: An observability perspective on time series foundation models (2025), https://arxiv.org/abs/2505.14766
5. Das, A., Kong, W., Sen, R., Zhou, Y.: A decoder-only foundation model for time-series forecasting. In: Proceedings of the 41st International Conference on Machine Learning (ICML) (2024), https://arxiv.org/abs/2310.10688
6. Einizade, A., Malliaros, F.D., Giraldo, J.H.: Continuous product graph neural networks (2024), https://arxiv.org/abs/2405.18877

7. Gao, H., Jiang, R., Dong, Z., Deng, J., Ma, Y., Song, X.: Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting (2024), https://arxiv.org/abs/2312.00516
8. Garza, A., Challu, C., Mergenthaler-Canseco, M.: Timegpt-1 (2023), https://arxiv.org/abs/2310.03589
9. Goswami, M., Küken, J., Choudhry, A., Challu, C., Mukherjee, S., Dubrawski, A.: Moment: A family of open time-series foundation models. In: Proceedings of the 41st International Conference on Machine Learning (ICML). PMLR, vol. 235, pp. 16115–16152 (2024)
10. He, D., Gao, Y., Jiang, B., Yan, H.: GAMMA-Net: Adaptive long-horizon traffic spatio-temporal forecasting model based on interleaved graph attention and multi-axis mamba (2026), https://arxiv.org/abs/2604.16859
11. Lee, S., Park, C.: Temporal graph learning recurrent neural network for traffic forecasting (2024), https://arxiv.org/abs/2406.02726
12. Li, F., Feng, J., Yan, H., Jin, G., Jin, D., Li, Y.: Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution (2021), https://arxiv.org/abs/2104.14917
13. Li, Y., Yu, R., Shahabi, C., Liu, Y.: Graph convolutional recurrent neural network: Data-driven traffic forecasting. CoRR **abs/1707.01926** (2017), http://arxiv.org/abs/1707.01926
14. Liu, C., et al.: Moirai 2.0: When less is more for time series forecasting (2025), https://arxiv.org/abs/2511.11698
15. Liu, X., Xia, Y., Liang, Y., Hu, J., Wang, Y., Bai, L., Huang, C., Liu, Z., Hooi, B., Zimmermann, R.: Largest: A benchmark dataset for large-scale traffic forecasting (2023), https://arxiv.org/abs/2306.08259
16. Meseguer-Gómez, J.I., Reina-Jiménez, P., Jiménez-Navarro, M.J., Martínez-Ballesteros, M., Riquelme-Santos, J.C.: Exploring zero-shot learning for time series forecasting using large language models. In: Proceedings of Soft Computing Models in Industrial and Environmental Applications, SOCO 2025. Communications in Computer and Information Science. vol. 2806, pp. 217–226 (2026). https://doi.org/10.1007/978-3-032-19763-4_20
17. Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., Sahoo, D.: Unified training of universal time series forecasting transformers. In: Proceedings of the 41st International Conference on Machine Learning (ICML) (2024), https://arxiv.org/abs/2402.02592
18. Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.: Graph wavenet for deep spatial-temporal graph modeling (2019), https://arxiv.org/abs/1906.00121
19. Xiong, R., Liao, L., Wang, Z., Wang, C., Xiao, H.: Msta-gnn: A graph neural network with multi-view spatio-temporal attention for traffic flow prediction. Journal of Network Intelligence **9**(4), 2471–2485 (2024)
20. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional neural network: A deep learning framework for traffic forecasting. CoRR **abs/1709.04875** (2017), http://arxiv.org/abs/1709.04875