RIDEBENCH: A LARGE-SCALE EXOGENOUS-AWARE BENCHMARK FOR RIDE-HAILING TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

We release **Ride-Hailing**, a large-scale ride-hailing time series dataset synthesized from marketplace data across 200 spatial areas. Ride-Hailing spans four consecutive years at half-hourly granularity and covers three representative exogenous scenarios: *weather disturbance*, *holiday effect*, and *large-scale event impact*. Built upon Ride-Hailing, we introduce **RideBench**, a comprehensive benchmark for exogenous-aware ride-hailing forecasting, covering both regular week-ahead forecasting and long-horizon 8-week-ahead forecasting with up to 2,688 prediction steps. RideBench evaluates over 30 representative forecasting methods, including endogenous-only models, exogenous-aware models, and time series foundation models. Our results show that future-known exogenous variables provide clear benefits in regular week-ahead forecasting, especially under weather, holiday, and large-scale event (e.g., major sporting events and concerts) scenarios. However, current exogenous-aware models still struggle to fully capture disturbance-induced pattern changes under complex external contexts. For long-horizon forecasting, existing models cannot simultaneously achieve low pointwise errors, accurate broad trends, and reliable near-term forecasts. These findings reveal a clear mismatch between existing forecasting models and real-world ride-hailing requirements, highlighting the need for models that can better exploit future-known exogenous information, scale across heterogeneous areas, and support long-horizon planning.

1 INTRODUCTION

Ride-hailing platforms have become an essential component of modern urban mobility systems (Geng et al., 2019; Chen et al., 2021). Their efficient operation relies heavily on accurate forecasting of marketplace time series, which supports a wide range of platform decisions, including supply-demand balancing, driver incentive planning, resource allocation, and event-specific operations (Hao et al., 2026; Yao et al., 2018). Over the past decade, time series forecasting has made substantial progress (Wu et al., 2022; Li et al., 2025; Zhou et al., 2021), driven by deep learning architectures (Vaswani et al., 2017; He et al., 2016; Gu & Dao, 2023), large-scale datasets (Liu et al., 2023; Mouatadid et al., 2023), and standardized benchmarking protocols (Qiu et al., 2024; Wang et al., 2024a). However, much of this progress has been driven by general-purpose forecasting settings, where benchmarks emphasize endogenous temporal patterns, long-range dependencies, and multivariate correlations more systematically than domain-specific operational requirements.

Ride-hailing forecasting differs from these settings in several important ways. Unlike many standard forecasting settings (Zhu et al., 2025; Lu et al., 2025; Yu et al., 2018), ride-hailing systems are marketplace-driven, spatially heterogeneous, and highly sensitive to external disturbances. These properties make forecasts depend not only on endogenous histories, but also on future-known exogenous signals (Li et al., 2026; Lim et al., 2021), cross-area variation, and longer-term planning needs. We summarize these requirements into three key challenges, illustrated in Figure 1, that a practical benchmark for this domain should capture.

i) Complex exogenous effects. Weather conditions, holidays, and large-scale events can alter both the level and temporal pattern of ride-hailing demand (Hao et al., 2026). However, exist-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ing benchmarks offer limited support for evaluating forecasts with future-known exogenous variables, especially categorical holiday and event indicators. ***ii) Heterogeneous multi-area forecasting.*** Real-world ride-hailing systems span diverse areas with distinct mobility patterns and external sensitivities. Unified models must learn shared temporal and exogenous patterns while preserving area-specific responses. However, existing benchmarks provide limited evaluation of the benefits of multi-area training under such heterogeneity (Wang et al., 2024a; Zhang et al., 2024). ***iii) Long-horizon operational planning.*** Beyond short-term forecasting, ride-hailing platforms require long-horizon predictions for strategic planning and event preparation. In such settings, forecasting is no longer only about minimizing pointwise errors. A useful model should also capture broad temporal trends, including whether the target series will rise or fall and whether the predicted trajectory follows the correct long-term shape. However, existing benchmarks largely evaluate short-horizon forecasting with pointwise metrics, providing limited evidence about model behavior under long-horizon forecasting (Xue et al., 2026; Qiao et al., 2026; Qiu et al., 2024).

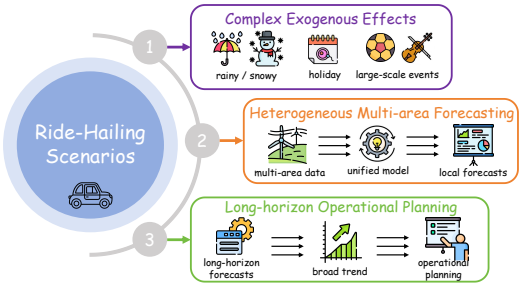


Figure 1: Three key challenges of real-world ride-hailing time series forecasting.

To address these limitations, we introduce **Ride-Hailing**, a large-scale ride-hailing time series dataset synthesized from marketplace data. It contains four years of half-hourly observations from 200 representative areas, totaling over 14 million records, with multiple target variables and future-known exogenous signals covering *Weather Disturbance*, *Holiday Effect*, and *Large-scale Event Impact*. Based on Ride-Hailing, we develop **RideBench**, a benchmark for exogenous-aware ride-hailing forecasting. RideBench includes two tasks: ***i)*** regular week-ahead forecasting with scenario-specific robustness evaluation, and ***ii)*** long-horizon 8-week-ahead forecasting with up to 2,688 prediction steps. Beyond pointwise accuracy, RideBench evaluates long-term trend modeling and near-term accuracy preservation through additional trend and first-week metrics. Compared with existing benchmarks (see Table 1), RideBench combines support for future-known continuous and categorical exogenous variables, scenario-specific evaluation, heterogeneous multi-area forecasting, and long-horizon trend assessment.

Table 1: Overview of existing benchmarks and our proposed RideBench.

Benchmark	Future Exo. Con./Dis.	Scenario Eval.	Multi-Area Modeling	Max Eval. Horizon
Monash	✗/✗	✗	✗	168
TSlib	✗/✗	✗	✗	720
BasicTS+	✗/✗	✗	✗	720
TFB	✓/✗	✗	✗	720
ProbTS	✗/✗	✗	✗	720
GIFT-Eval	✗/✗	✗	✗	900
fev-bench	✓/✓	✗	✗	168
QuitoBench	✗/✗	✗	✗	512
TIME	✗/✗	✗	✗	864
RideBench	✓/✓	✓	✓	2688

We conduct a comprehensive evaluation of RideBench with over 30 representative methods. Our results show that existing methods achieve strong performance on regular ride-hailing forecasting, yet remain limited in realistic deployment scenarios. First, while future-known exogenous variables improve short-term forecasting, current exogenous-aware models struggle to capture disturbance-induced distribution shifts, particularly for categorical events and area-specific responses. Second, long-horizon forecasting remains challenging, as existing models fail to simultaneously preserve pointwise accuracy, global trend consistency, and near-term prediction quality. These findings highlight the gap between current forecasting models and real-world ride-hailing requirements. RideBench provides a unified platform for advancing exogenous-aware and long-horizon time series forecasting research. In summary, our contributions are as follows:

- We release **Ride-Hailing**, a large-scale synthesized ride-hailing time series dataset, spanning four years at half-hourly granularity across 200 representative spatial areas, with rich future-known exogenous variables covering weather, holidays, and large-scale events.
- We introduce **RideBench**, an exogenous-aware ride-hailing forecasting benchmark covering both regular week-ahead forecasting and long-horizon 8-week-ahead forecasting, with overall, scenario-specific, multi-area, and trend-level evaluations.

- We systematically evaluate over 30 representative forecasting methods under unified protocols, providing a comprehensive empirical comparison for ride-hailing forecasting.
- We reveal key limitations of existing models in exploiting future-known exogenous signals and maintaining reliable long-horizon forecasts.

2 Related Work

Benchmarks are essential for time series forecasting because they provide unified datasets, data splits, evaluation metrics, and experimental protocols for comparing different models (Xue et al., 2026; Shchur et al., 2025). Early benchmarks mainly focused on dataset curation and standardized evaluation for common forecasting domains, such as electricity (Zhang et al., 2023; Fiot & Dinuzzo, 2016), traffic (Yu et al., 2018; Patil et al., 2025), weather (Liang et al., 2023; Wang et al., 2023), and retail demand (Fildes et al., 2022; Singh et al., 2013). Recent efforts further improve benchmarking quality by examining fairness factors in preprocessing, training pipelines, and metric computation, or by extending evaluation to multiple tasks such as forecasting, imputation, classification, and anomaly detection (Qiu et al., 2024; Wang et al., 2024a). With the rise of time series foundation models, benchmarks such as GIFT-Eval (Aksu et al., 2024) and TIME (Qiao et al., 2026) further emphasize zero-shot forecasting and scalable evaluation protocols.

Despite these advances, existing benchmarks still provide limited support for evaluating exogenous-aware forecasting. Most of them are not designed around marketplace-driven demand, heterogeneous spatial areas, future-known categorical exogenous variables, or long-horizon operational planning. RideBench complements existing benchmarks by focusing on these practical requirements in a large-scale ride-hailing setting that includes real-world exogenous factors. A more detailed review of related work is provided in Appendix A.1.

3 Dataset

3.1 Dataset Construction

Data sources and scope. As shown in Figure 2, Ride-Hailing is synthesized from ride-hailing marketplace data collected across 200 spatial areas. It spans four consecutive years with a 30-minute sampling interval, resulting in 14,025,600 observations. Each observation contains historical marketplace signals and future-known contextual features. To protect sensitive operational information, data processing includes anonymizing area and variable identifiers, transforming endogenous targets, aggregating weather disturbances, and encoding large-scale events as coarse-grained impact categories. These procedures aim to preserve temporal patterns, exogenous relationships, and cross-area structures for forecasting research.

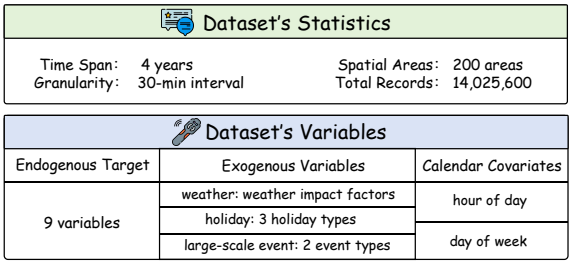


Figure 2: Overview of the Ride-Hailing dataset.

Endogenous and exogenous variables. Ride-Hailing includes nine anonymized platform-level endogenous variables (endo-1 to endo-9) as forecasting targets. These variables capture diverse marketplace dynamics and may exhibit interactions through disturbance, substitution, and co-movement effects. To enable exogenous-aware forecasting, Ride-Hailing provides future-known variables from three categories: weather disturbance, holiday effects (i.e., public holidays, traditional festivals, and Western festivals, each consisting of multiple representative holiday events), and large-scale events (e.g., major sporting events and concerts). Weather is represented by an impact factor, holidays by categorical indicators, and large-scale events by event-impact variables describing different mobility responses. Calendar covariates, including hour-of-day and day-of-week indicators, are also included to capture regular temporal patterns.

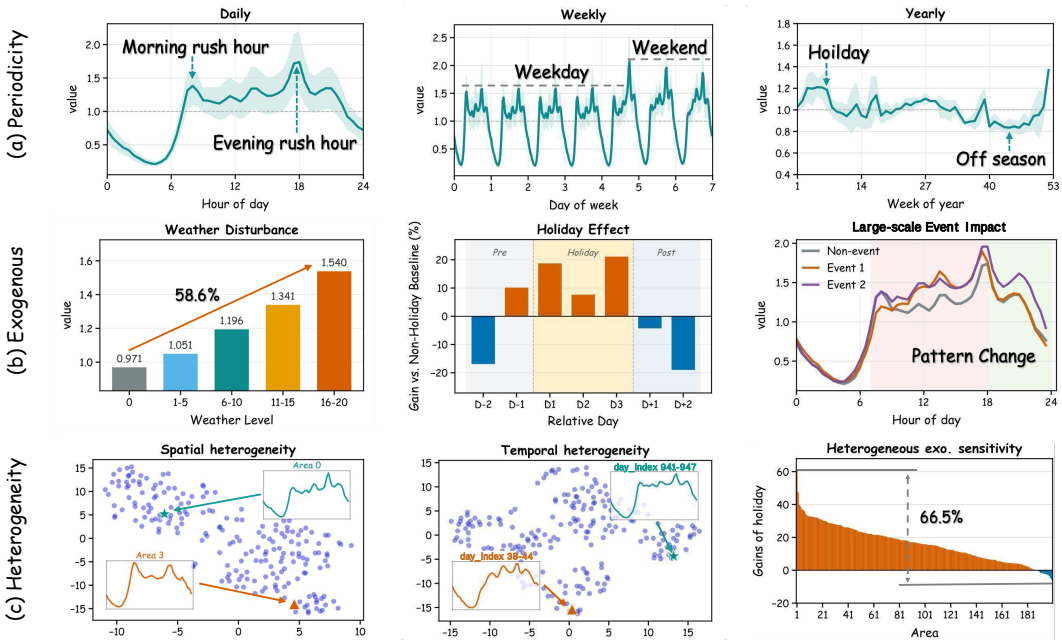


Figure 3: Key characteristics of Ride-Hailing. (a) Temporal patterns at daily, weekly, and yearly scales. (b) Variations associated with weather disturbances, holiday stages, and large-scale event types. (c) Heterogeneity across areas, over time, and in area-specific responses to holidays.

3.2 Dataset Characteristics

Multi-scale periodicity. Ride-Hailing exhibits temporal patterns at daily, weekly, and yearly scales, as shown in Figure 3(a). Daily demand falls to a low in the early morning and exhibits morning and evening peaks. Weekly patterns show recurring weekday profiles and distinct weekend shapes, indicating that a single daily template cannot capture all calendar-dependent variation. At the yearly scale, the weekly-aggregated series shows broader fluctuations, including holiday-related peaks and an off-season trough. These patterns involve variation in both the shape of daily demand profiles and the overall demand level over longer periods. Capturing daily peaks alone is therefore insufficient: models must also account for weekday–weekend differences and broader seasonal variation.

Exogenous effects. Ride-Hailing is influenced by diverse future-known exogenous factors, including weather, holidays, and large-scale events (see Figure 3(b)). The average endogenous value increases with weather disturbance intensity, showing that stronger disturbances are associated with larger deviations from regular conditions. Holiday effects are temporally structured: demand varies across pre-holiday, in-holiday, and post-holiday stages, with both positive and negative deviations from the non-holiday baseline. These patterns require models to capture phase-dependent responses rather than treating holidays as simple level changes. Large-scale events introduce localized and time-dependent variations. Ride-Hailing represents these events through two coarse-grained impact categories: one primarily affects daytime mobility, while the other affects both daytime and evening demand. Distinguishing these responses requires models to exploit event-type information and capture its interaction with temporal patterns.

Heterogeneity. Ride-Hailing exhibits substantial heterogeneity across regions, time periods, and exogenous conditions (see Figure 3(c)). Different areas possess distinct mobility structures due to urban functions and activity patterns, while temporal distributions evolve with changing behaviors and platform dynamics. Moreover, the same exogenous factor can induce different responses across areas. For example, holidays may increase demand in some areas but reduce it in others, reflecting differences in local mobility patterns. A unified forecasting model must therefore learn common temporal structures while preserving area-specific responses to external factors.

Overall, Ride-Hailing provides a challenging real-world forecasting setting that integrates multi-scale temporal dependencies, exogenous influences, and heterogeneous spatial-temporal patterns.

4 BENCHMARK

Built upon Ride-Hailing, we introduce **RideBench**, a unified benchmark for evaluating exogenous-aware ride-hailing time series forecasting. As shown in Figure 4, RideBench consists of three core components: data processing, forecasting methods, and comprehensive evaluation protocols.

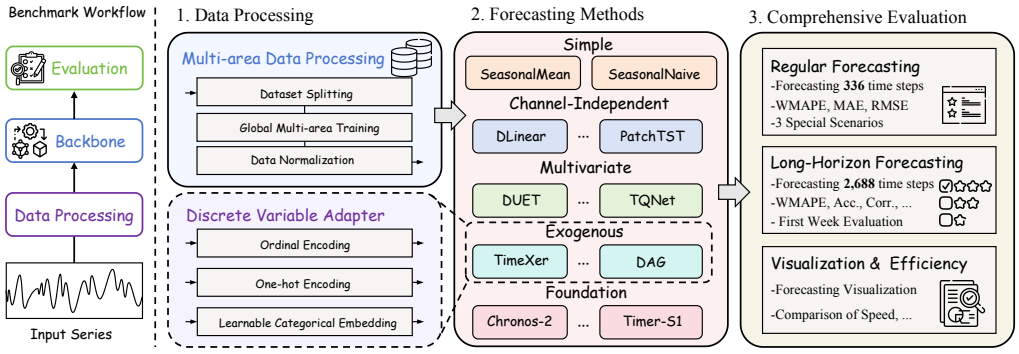


Figure 4: Overall architecture of RideBench.

4.1 DATA PROCESSING

RideBench uses a global multi-area training protocol. Each temporal window from an individual area forms a forecasting sample, and samples from all 200 areas are pooled to train a unified model for forecasting across all areas. This protocol allows models to learn common temporal and exogenous patterns while forecasting heterogeneous area-specific dynamics.

The data are split chronologically into three years for training, six months for validation, and six months for testing. Each sample is assigned to a split only when its prediction window falls entirely within that split; its input window may use preceding observations. Continuous variables are standardized using training-set statistics, and predictions are inverse-transformed before evaluation. Further processing details are provided in Appendix A.2.1.

4.2 FORECASTING METHODS

RideBench evaluates more than 30 representative methods across five groups: simple statistical baselines, channel-independent models, multivariate models, exogenous-aware models, and time series foundation models. These groups enable complementary comparisons of seasonal regularity, individual target dynamics, cross-variable dependencies, external covariate utilization, and zero-shot transferability.

To accommodate both continuous and categorical exogenous inputs, we adapt exogenous-aware methods using ordinal encoding, one-hot encoding, or learnable categorical embeddings. Where needed, their input interfaces are extended to incorporate future-known covariates over the prediction horizon. Time series foundation models are evaluated in a zero-shot setting. Detailed baseline descriptions and adaptations are provided in Appendix A.2.2.

4.3 FORECASTING TASKS AND EVALUATION

RideBench provides two half-hourly forecasting settings. The first is **week-ahead forecasting**, where models predict the next 336 steps (one week) using 1,344 historical steps (four weeks). This setting evaluates short-term forecasting accuracy. We evaluate point forecasting accuracy using three standard metrics: Weighted Mean Absolute Percentage Error (WMAPE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). WMAPE measures scale-normalized absolute error, MAE measures average absolute error, and RMSE emphasizes large deviations. Formally,

$$\text{WMAPE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i|}, \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{1}$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted values of the i-th evaluated point, and n is the total number of evaluated points.

In addition to overall accuracy, we evaluate week-ahead forecasting under three exogenous scenarios: *Weather Disturbance*, *Holiday Effect*, and *Large-scale Event Impact*. For each scenario, we use the corresponding future-known exogenous indicators to select affected periods in the test set and compute WMAPE, MAE, and RMSE on the resulting subset. This scenario-specific evaluation assesses forecasting performance under external disturbances that may be less apparent in aggregate results, complementing the overall evaluation dominated by regular temporal patterns.

The second is **8-week-ahead forecasting**, where models forecast the next 2,688 steps (eight weeks) using 4,704 historical steps (fourteen weeks). Beyond pointwise accuracy, this setting emphasizes the ability to capture long-term demand evolution. Therefore, we additionally evaluate daily-level trend consistency using Directional Accuracy of Day (Acc.) and Correlation of Day (Corr.). Specifically, we aggregate half-hourly predictions into daily values by averaging every 48 time steps. Given the ground-truth and predicted daily values Y_i and $\hat{Y}_i$, Acc. measures the consistency of day-to-day variation directions:

$$\text{Acc.} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathbb{I}\left[\text{sign}(Y_i - Y_{i-1}) = \text{sign}(\hat{Y}_i - \hat{Y}_{i-1})\right], \tag{2}$$

where $\Omega = \{i : |Y_i - Y_{i-1}| > 10^{-6}\}$ excludes days with negligible changes. Corr. measures the correlation between predicted and ground-truth daily trajectories:

$$\text{Corr.} = \frac{\sum_{i=1}^{m}(Y_i - \mu)(\hat{Y}_i - \hat{\mu})}{\sqrt{\sum_{i=1}^{m}(Y_i - \mu)^2}\sqrt{\sum_{i=1}^{m}(\hat{Y}_i - \hat{\mu})^2}}, \tag{3}$$

where m is the number of evaluated days, and μ and $\hat{\mu}$ are the means of the ground-truth and predicted daily sequences, respectively. Acc. measures whether a model predicts the correct direction of day-to-day demand changes, while Corr. measures whether its predicted daily trajectory follows the overall shape of the ground truth. Together, these metrics assess long-horizon trend quality beyond pointwise errors.

We additionally report first-week accuracy within the 8-week prediction horizon to assess whether models preserve near-term precision while forecasting farther into the future. The long-horizon evaluation thus jointly examines pointwise accuracy over the full horizon, daily trend quality, and near-term accuracy preservation. Implementation details are provided in Appendix A.2.3.

5 Experiments

5.1 Main Results

Regular forecasting. Table 2 reports the results of regular week-ahead forecasting. Overall, exogenous-aware models achieve the strongest performance across both overall evaluation and the three scenario-specific evaluations. Among them, TiDE (Das et al., 2023a) achieves strong performance across different metrics and exogenous scenarios, highlighting the importance of effectively using future-known covariates for modeling demand deviations caused by external disturbances.

Some endogenous-only models, including PETformer (Lin et al., 2024c) and PMDformer (Hu et al., 2026), remain competitive in overall WMAPE and MAE. This reflects the strong daily and weekly periodicity of regular ride-hailing demand, which can be captured from historical target values alone. However, without access to future-known exogenous variables, these models are less effective under exogenous scenarios and also show less favorable overall RMSE. This indicates that historical endogenous patterns alone are insufficient for modeling high-magnitude or abrupt demand changes.

Time series foundation models also show promising zero-shot transferability. Chronos-2 (Ansari et al., 2025) achieves performance comparable to several task-specific trained models, suggesting the potential of large-scale pretrained forecasting models. Nevertheless, they still lag behind specialized exogenous-aware models, especially under exogenous scenarios where future-known external information is important.

Table 2: **Regular week-ahead forecasting performance.** Darker shades indicate better performance. The average rank is computed as the arithmetic mean of ranks across all evaluation metrics.

	Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact			Rank
	Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	Average
Sim.	SeasonalMean	0.1285	4.916	14.37	0.1509	6.311	20.90	0.1701	7.052	20.85	0.1224	5.395	16.32	29
	SeasonalNaive	0.1490	5.693	17.75	0.1755	7.338	24.41	0.1981	8.203	24.57	0.1422	6.261	19.10	32
Channel-Independent	PatchTST	0.1079	4.132	13.38	0.1391	5.821	20.84	0.1507	6.249	20.37	0.1130	4.983	16.56	10
	PETformer	0.1078	4.124	13.35	0.1396	5.839	20.86	0.1520	6.303	20.42	0.1130	4.982	16.48	11
	TimeMixer	0.1113	4.261	13.47	0.1411	5.902	20.90	0.1548	6.420	20.42	0.1159	5.113	16.73	19
	STID	0.1122	4.299	13.48	0.1419	5.941	20.92	0.1542	6.397	20.22	0.1166	5.149	16.74	20
	DLinear	0.1171	4.483	13.48	0.1441	6.029	20.64	0.1545	6.403	19.78	0.1186	5.234	16.29	21
	TimeBase	0.1222	4.674	13.68	0.1437	6.008	20.38	0.1553	6.437	19.67	0.1209	5.342	16.19	22
	SegRNN	0.1116	4.275	13.60	0.1418	5.932	20.98	0.1533	6.360	20.48	0.1196	5.280	17.03	23
	RMLP	0.1135	4.344	13.51	0.1431	5.986	20.91	0.1562	6.476	20.39	0.1176	5.185	16.67	25
	SparseTSF	0.1134	4.342	13.60	0.1420	5.941	20.99	0.1553	6.442	20.53	0.1176	5.193	16.75	27
	PhaseFormer	0.1151	4.411	13.69	0.1432	5.993	21.03	0.1529	6.339	20.26	0.1226	5.423	17.07	28
Multivariate	PMDformer	0.1077	4.122	13.28	0.1384	5.790	20.70	0.1486	6.157	20.01	0.1125	4.962	16.39	4
	TimeFilter	0.1099	4.206	13.30	0.1390	5.812	20.67	0.1494	6.187	20.00	0.1137	5.012	16.42	7
	Crossformer	0.1096	4.199	13.42	0.1390	5.816	20.76	0.1533	6.355	20.46	0.1141	5.039	16.59	12
	SOFTS	0.1111	4.252	13.45	0.1408	5.888	20.84	0.1533	6.354	20.29	0.1159	5.112	16.76	13
	ModernTCN	0.1117	4.279	13.42	0.1416	5.925	20.86	0.1512	6.269	20.11	0.1171	5.171	16.78	14
	TQNet	0.1117	4.277	13.44	0.1413	5.910	20.83	0.1537	6.373	20.25	0.1161	5.121	16.66	14
	iTransformer	0.1117	4.276	13.47	0.1410	5.899	20.88	0.1518	6.290	20.16	0.1169	5.159	16.79	16
	Leddam	0.1117	4.280	13.47	0.1417	5.928	20.87	0.1534	6.361	20.30	0.1163	5.133	16.68	18
	CrossGNN	0.1125	4.309	13.59	0.1417	5.926	21.00	0.1542	6.392	20.45	0.1170	5.162	16.81	24
	DUET	0.1147	4.397	13.57	0.1437	6.011	20.97	0.1539	6.379	20.23	0.1193	5.267	16.86	26
Exogenous	TiDE	0.1043	3.996	12.43	0.1293	5.397	18.96	0.1282	5.305	17.29	0.1104	4.866	15.81	1
	CrossLinear	0.1051	4.032	12.65	0.1269	5.296	18.74	0.1328	5.505	18.43	0.1100	4.854	15.87	2
	XLinear	0.1075	4.124	12.86	0.1300	5.431	19.42	0.1326	5.493	18.52	0.1144	5.055	16.30	3
	CATS	0.1103	4.225	12.95	0.1321	5.510	19.53	0.1379	5.703	18.76	0.1159	5.106	16.13	5
	DAG	0.1090	4.174	13.30	0.1388	5.803	20.64	0.1476	6.115	19.91	0.1141	5.033	16.55	6
	TimeXer	0.1093	4.207	12.93	0.1388	5.804	20.14	0.1371	5.687	18.42	0.1172	5.177	16.40	8
Foundation	Chronos-2	0.1089	4.174	13.24	0.1363	5.703	20.08	0.1523	6.319	20.32	0.1155	5.101	16.34	9
	Timer-S1	0.1121	4.295	13.48	0.1417	5.927	20.86	0.1539	6.384	20.37	0.1153	5.090	16.54	17
	TimesFM 2.5	0.1160	4.457	13.69	0.1448	6.062	21.04	0.1585	6.581	20.58	0.1208	5.341	16.88	30
	Moirai 2.0	0.1218	4.678	13.80	0.1466	6.130	20.92	0.1589	6.593	20.38	0.1260	5.572	17.06	31

In summary, these results show that future-known exogenous variables are valuable for regular week-ahead forecasting, improving both overall accuracy and scenario-specific performance.

Long-horizon forecasting. Table 3 presents the results of long-horizon 8-week-ahead forecasting. Different from week-ahead forecasting, this setting evaluates long-horizon pointwise accuracy, day-level trend quality, and first-week accuracy, reflecting fine-grained numerical precision, broad trajectory modeling, and near-term accuracy preservation, respectively.

The results show that exogenous-aware models do not maintain the clear advantage observed in regular week-ahead forecasting. Instead, several endogenous-only models achieve stronger overall accuracy. This suggests that long-horizon ride-hailing forecasting relies heavily on stable temporal structures, such as weekly seasonality and long-term demand evolution. It also indicates that simply adding exogenous inputs does not guarantee better long-horizon performance, and that current exogenous-aware architectures may not fully exploit such information over extended horizons.

Another important observation is that pointwise accuracy and trend quality are not strictly aligned. Models with lower WMAPE, MAE, or RMSE do not necessarily achieve the best Acc. or Corr. This indicates that local numerical precision and broad trajectory modeling capture distinct forecasting capabilities, and that existing models still struggle to achieve both simultaneously. For long-horizon operational planning, an effective model should go beyond minimizing pointwise errors and reliably capture the direction and overall shape of future demand changes.

Moreover, several channel-independent models and foundation models achieve strong first-week accuracy. This suggests that models with strong temporal pattern modeling can preserve near-term precision even when evaluated under an 8-week horizon. By contrast, the advantages of modeling cross-variable dependencies or exogenous interactions in week-ahead forecasting do not necessarily transfer to the long-horizon setting.

In summary, no existing model dominates all dimensions of long-horizon forecasting. This reveals an open challenge: future long-horizon ride-hailing forecasting models should balance long-horizon pointwise accuracy, trend quality, and first-week accuracy. We also analyze the computational efficiency of these models, with detailed results presented in Appendix A.3.

Table 3: **Eight-week-ahead forecasting performance.** It reports long-horizon point-wise errors, day-level trend quality (i.e., directional accuracy of day and correlation of day), and first-week accuracy. Rank is calculated as the arithmetic mean of metric ranks across all experimental settings.

	Setting	Overall				Trend			First Week			
	Metrics	WMAPE	MAE	RMSE	Rank	Acc.(↑)	Corr.(↑)	Rank	WMAPE	MAE	RMSE	Rank
Sim.	SeasonalMean	0.1485	5.661	14.94	28	0.5875	0.2059	28	0.1391	5.349	14.86	31
	SeasonalNaive	0.1731	6.582	19.63	31	0.5518	0.1246	30	0.1613	6.182	19.36	32
Channel-Independent	PETformer	0.1278	4.879	14.58	2	0.6046	0.2549	12	0.1162	4.463	14.33	2
	TimeMixer	0.1294	4.940	14.57	3	0.6061	0.2679	9	0.1182	4.542	14.43	3
	SegRNN	0.1273	4.862	14.66	4	0.6059	0.2470	13	0.1173	4.508	14.48	6
	PatchTST	0.1299	4.957	14.58	7	0.6001	0.2457	14	0.1188	4.562	14.42	3
	PhaseFormer	0.1309	4.998	14.72	14	0.5983	0.2370	19	0.1190	4.568	14.50	10
	SparseTSF	0.1338	5.106	14.92	19	0.5956	0.2121	24	0.1191	4.572	14.54	11
	STID	0.1370	5.233	14.97	23	0.6026	0.2343	17	0.1254	4.818	14.73	27
	TimeBase	0.1385	5.279	14.74	22	0.5895	0.1949	26	0.1239	4.752	14.54	25
	RMLP	0.1398	5.322	14.99	28	0.5919	0.2346	21	0.1266	4.856	14.65	27
	DLinear	0.1396	5.318	14.84	25	0.5889	0.2282	25	0.1272	4.876	14.67	29
Multivariate	ModernTCN	0.1262	4.816	14.34	1	0.6089	0.3273	3	0.1193	4.587	14.28	8
	Leddam	0.1295	4.945	14.54	4	0.6110	0.2837	2	0.1194	4.586	14.38	9
	SOFTS	0.1280	4.886	14.64	6	0.6078	0.2783	4	0.1190	4.572	14.59	13
	iTransformer	0.1302	4.972	14.70	11	0.6094	0.2527	5	0.1207	4.637	14.56	20
	PMDformer	0.1316	5.022	14.63	11	0.5959	0.2402	20	0.1202	4.613	14.40	11
	CrossGNN	0.1323	5.054	14.70	15	0.5952	0.2222	23	0.1200	4.610	14.46	15
	TQNet	0.1328	5.075	14.74	17	0.6019	0.2367	16	0.1213	4.659	14.53	20
	TimeFilter	0.1340	5.112	14.66	17	0.5992	0.2439	15	0.1227	4.708	14.51	22
	DUET	0.1338	5.122	14.89	20	0.6106	0.2476	6	0.1235	4.756	14.69	26
	Crossformer	0.1348	5.187	15.21	24	0.6094	0.2476	7	0.1210	4.663	14.75	23
Exogenous	TiDE	0.1328	5.069	14.32	10	0.6137	0.3715	1	0.1194	4.582	13.99	7
	CrossLinear	0.1293	4.941	14.66	7	0.6077	0.2671	7	0.1194	4.590	14.49	14
	TimeXer	0.1317	5.040	14.56	9	0.6058	0.2707	10	0.1218	4.690	14.42	17
	XLinear	0.1303	4.975	14.73	13	0.6072	0.2492	11	0.1211	4.658	14.53	19
	DAG	0.1335	5.099	14.63	15	0.5975	0.2428	17	0.1222	4.695	14.45	18
	CATS	0.1349	5.155	14.85	21	0.5961	0.2200	21	0.1230	4.726	14.56	24
Foundation	Chronos-2	0.1374	5.270	15.07	26	0.5905	0.1702	26	0.1140	4.382	14.27	1
	Timer-S1	0.1379	5.288	15.11	27	0.5583	0.0910	31	0.1168	4.487	14.49	3
	TimesFM 2.5	0.1564	6.026	15.66	30	0.5737	0.1128	29	0.1193	4.589	14.59	16
	Moirai 2.0	0.2039	7.925	18.48	32	0.5405	0.0582	32	0.1263	4.864	14.79	30

5.2 Ablation and Analysis

We conduct additional analyses under the regular week-ahead forecasting setting to understand the impact of key design factors.

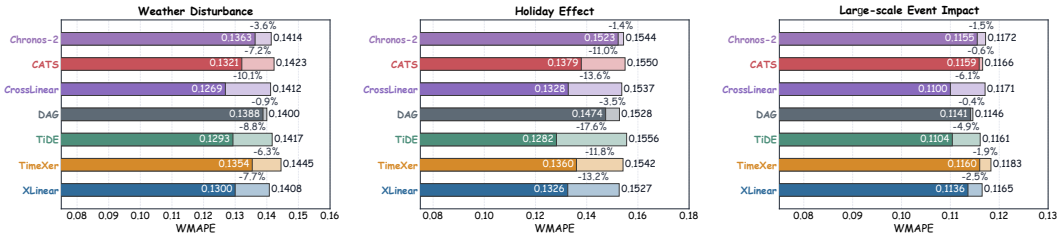


Figure 5: Exogenous variable utilization capabilities of different models under scenario evaluation. Larger improvements demonstrate the stronger capability in leveraging exogenous information.

Exogenous variable utilization. We evaluate the contribution of future-known external variables by comparing models with and without exogenous inputs (Figure 5). Although Chronos-2 (Ansari et al., 2025) is a zero-shot foundation model, we include it since it supports exogenous inputs.

Most exogenous-aware models benefit from additional variables, confirming that external signals provide useful predictive information. TiDE (Das et al., 2023a) achieves the largest average improvement, with a relative gain of 10.4%. Chronos-2 also improves under exogenous inputs, showing that foundation models can exploit covariates, although the gains are smaller for categorical factors (e.g., *Holiday Effect*), possibly reflecting difficulties in learning semantic mappings from discrete encodings.

Compared with weather and holiday scenarios, *Large-scale Event Impact* yields smaller improvements, likely because event effects are sparse and heterogeneous, making reliable pattern learning

challenging. Overall, exogenous variables consistently improve ride-hailing forecasting, while their effectiveness depends on signal type and scenario complexity.

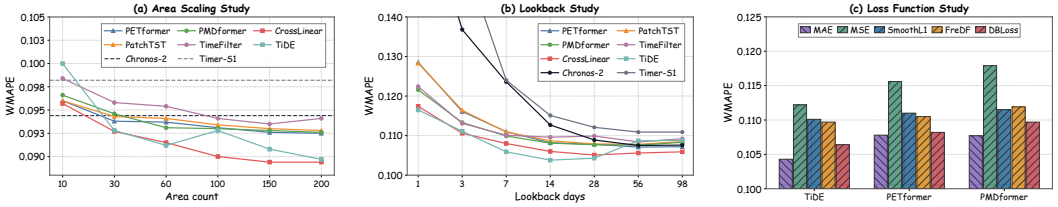


Figure 6: Studies on key benchmark design factors.

Area scaling study. Ride-Hailing provides parallel time series across 200 spatial areas, enabling us to study the effect of multi-area training. We fix 10 target areas for evaluation and increase the number of training areas from 10 to 200, starting with the target areas alone and progressively adding other areas. As shown in Figure 6(a), Chronos-2 (Ansari et al., 2025) performs strongly in the limited-data setting, demonstrating the advantage of pretrained models. However, task-specific models continue to improve with more training areas, indicating that additional areas provide transferable information rather than merely introducing heterogeneity.

This result supports the global training design of RideBench: although areas differ in demand patterns, they share common temporal dynamics and responses to external factors. Large-scale domain data remain valuable for developing specialized ride-hailing forecasting models.

Lookback study. We vary the input lookback length to examine the effect of historical context (Figure 6(b)). Most models improve with longer inputs, confirming that ride-hailing demand benefits from capturing multiple temporal patterns, including daily and weekly cycles. However, overly long contexts do not always help. When extending the lookback window to 98 days, several task-specific models slightly degrade, likely due to irrelevant historical information and fewer training samples. Therefore, RideBench adopts a 28-day lookback window for week-ahead forecasting, balancing temporal coverage and training efficiency. Foundation models are more sensitive to short contexts because they rely directly on input observations to infer task-specific patterns. Sufficient historical context is thus particularly important for zero-shot forecasting.

Loss function study. We compare MSE, MAE, Smooth L1 (Girshick, 2015), FreDF (Wang et al., 2025), and DBloss (Qiu et al., 2026a) as training objectives (Figure 6(c)). FreDF emphasizes frequency-domain discrepancies, while DBloss considers decomposed temporal components such as periodicity and trend. Despite the strong periodicity of ride-hailing demand, MAE achieves the best overall performance among the evaluated objectives. This result suggests that explicitly targeting temporal structures alone may be insufficient; MAE's lower sensitivity to large errors than MSE may also be beneficial under exogenous disturbances. Based on this empirical comparison, we adopt MAE as the default training objective in RideBench.

6 Conclusion

In this paper, we introduced **Ride-Hailing**, a synthesized ride-hailing time series dataset, and developed **RideBench**, a unified benchmark for exogenous-aware forecasting under regular week-ahead and long-horizon settings. RideBench provides standardized processing, global multi-area training, scenario-specific evaluation, and long-horizon trend assessment. Through extensive evaluation of over 30 representative methods, we show that future-known exogenous variables generally improve week-ahead forecasting, particularly under external disturbance scenarios. Our analyses (see Section 5.2) further demonstrate the importance of scalable multi-area training, sufficient historical context, and robust optimization objectives. Meanwhile, RideBench reveals two key challenges: existing models remain limited in capturing disturbance-induced pattern shifts, and long-horizon forecasting requires better trade-offs between accuracy, trend consistency, and short-term reliability. By providing a privacy-preserving dataset and unified evaluation framework, RideBench aims to advance research on scalable, exogenous-aware, and deployment-oriented forecasting for ride-hailing systems.

AI USE STATEMENT

In this work, we used a generative AI tool (i.e., ChatGPT) solely for assisting with manuscript writing refinement, including checking spelling errors, correcting grammar, and improving language clarity. We did not use generative AI tools for research ideation, methodology development, data analysis, experimental design, implementation, or generation of scientific results. All AI-assisted revisions were carefully reviewed and verified by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The benchmark code is publicly available at Anonymous GitHub.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

George EP Box and David A Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332):1509–1526, 1970.

Long Chen, Piyushimita Thakuriah, and Konstantinos Ampountolas. Short-term prediction of demand for ride-hailing services: A deep learning approach. *Journal of Big Data Analytics in Transportation*, 3(2):175–195, 2021.

Xinyang Chen, Huidong Jin, Yu Huang, and Zaiwen Feng. Xlinear: A lightweight and accurate mlp-based model for long-term time series forecasting with exogenous inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023a.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023b.

Robert Fildes, Shaohui Ma, and Stephan Kolassa. Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4):1283–1318, 2022.

Jean-Baptiste Fiot and Francesco Dinuzzo. Electricity demand forecasting by multi-task learning. *IEEE Transactions on Smart Grid*, 9(2):544–551, 2016.

Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3656–3663, 2019.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. Softs: Efficient multivariate time series forecasting with series-core fusion. *Advances in Neural Information Processing Systems*, 37: 64145–64175, 2024.

Xixuan Hao, Guicheng Li, Daiqiang Wu, Xusen Guo, Yumeng Zhu, Zhichao Zou, Peng Zhen, Yao Yao, and Yuxuan Liang. Enhancing ride-hailing forecasting at didi with multi-view geospatial representation learning from the web. In *Proceedings of the ACM Web Conference 2026*, pp. 8200–8211, 2026.

Andrew C Harvey and Andrew Harvey. Forecasting, structural time series models and the kalman filter. 1990.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending tabpfn-v2 to time series forecasting. *arXiv preprint arXiv:2501.02945*, 2025.

Ao Hu, Liangjian Wen, Jiang Duan, Yong Dai, YAN HE, Dongkai Wang, Jun Wang, Yukun Zhang, Ruoxi Jiang, and Zenglin Xu. Pmdformer: Patch-mean decoupling information transformer for long-term forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *International Conference on Machine Learning*, pp. 24893–24911. PMLR, 2025.

Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. *Advances in Neural Information Processing Systems*, 36:46885–46902, 2023.

Qihe Huang, Zhengyang Zhou, Kuo Yang, Zhongchao Yi, Xu Wang, and Yang Wang. TimeBase: The power of minimalism in efficient long-term time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 26227–26246, 2025.

Rob Hyndman, Anne Koehler, Keith Ord, and Ralph Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer, 2008.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

Yue Jiang, Yile Chen, Xiucheng Li, Qin Chao, Shuai Liu, and Gao Cong. Fstllm: Spatio-temporal llm for few shot time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International conference on learning representations*, volume 2024, pp. 23857–23880, 2024.

Boyuan Li, Yicheng Luo, Zhen Liu, Junhao Zheng, Jianming Lv, and Qianli Ma. Hyperimts: Hypergraph neural network for irregular multivariate time series forecasting. *arXiv preprint arXiv:2505.17431*, 2025.

Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on linear mapping. *arXiv preprint arXiv:2305.10721*, 2023.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Chenjuan Guo, and Bin Yang. Gcgnet: Graph-consistent generative network for time series forecasting with exogenous variables. *arXiv preprint arXiv:2603.08032*, 2026.

Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 14329–14337, 2023.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024a.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. Sparsetsf: modeling long-term time series forecasting with 1k parameters. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 30211–30226, 2024b.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Songbo Wang, and Yongxiang Wang. Petformer: Long-term time series forecasting via placeholder-enhanced transformer. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024c.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025a.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Feiyu Zhao, Ruichao Mo, and Haotong Zhang. Segrnn: Segment recurrent neural network for long-term time series forecasting. *IEEE Internet of Things Journal*, 2025b.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18780–18788, 2025b.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in Neural Information Processing Systems*, 36:75354–75371, 2023.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Yong Liu, Xingjian Su, Shiyu Wang, Haoran Zhang, Haixuan Liu, Yuxuan Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Timer-s1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jiecheng Lu, Xu Han, Yan Sun, and Shihao Yang. Cats: Enhancing multivariate time series forecasting by constructing auxiliary time series as exogenous variables. In *Forty-first International Conference on Machine Learning*, 2024.

Kuan Lu, Menghao Huo, Yuxiao Li, Qiang Zhu, and Zhenrui Chen. Ct-patchtst: Channel-time patch time-series transformer for long-term renewable energy forecasting. In *2025 10th International Conference on Computer and Information Processing Technology (ISCIPT)*, pp. 86–95. IEEE, 2025.

Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *International Conference on Machine Learning*, 2024.

Francisco Martínez, María Pilar Frías, María Dolores Pérez, and Antonio Jesús Rivera. A methodology for applying k-nearest neighbor to time series forecasting. *Artificial Intelligence Review*, 52(3):2019–2037, 2019.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Knight, Maria Geogdzhayeva, Sam Levang, Ernest Fraenkel, et al. Subseasonalclimateusa: A dataset for subseasonal forecasting and benchmarking. *Advances in Neural Information Processing Systems*, 36:7960–7992, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Yiming Niu, Jinliang Deng, and Yongxin Tong. Phaseformer: From patches to phases for efficient and effective time series forecasting. In *International Conference on Learning Representations*, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Mayur Patil, Qadeer Ahmed, and Shawn Midlam-Mohler. Travel time and weather-aware traffic forecasting in a conformal graph neural network framework. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's time: Towards the next generation of time series forecasting benchmarks. *arXiv preprint arXiv:2602.12147*, 2026.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17(9): 2363–2377, 2024.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1185–1196, 2025.

Xiangfei Qiu, Xingjian Wu, Hanyin Cheng, Xvyuan Liu, Chenjuan Guo, Jilin Hu, and Bin Yang. Dbloss: Decomposition-based loss function for time series forecasting. *Advances in Neural Information Processing Systems*, 38:27741–27768, 2026a.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. Dag: A dual correlation network for time series forecasting with exogenous variables. In *International Conference on Machine Learning*, 2026b.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 4454–4458, 2022.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

Arunesh Kumar Singh, S Khatoon Ibraheem, Md Muazzam, and DK Chaturvedi. An overview of electricity demand forecasting techniques. *Network and complex systems*, 3(3):38–48, 2013.

Zhong Siru, Ruan Weilin, Ming Jin, Li Huan, Wen Qingsong, and Liang Yuxuan. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *Forty-Second International Conference on Machine Learning (ICML 2025)*. Proceedings of Machine Learning Research, 2025.

Hiro Y Toda and Peter CB Phillips. Vector autoregression and causality: a theoretical overview and simulation study. *Econometric reviews*, 13(2):259–285, 1994.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hao Wang, Lichen Pan, Yuan Shen, Zhichao Chen, Degui Yang, Yifei Yang, Sen Zhang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. Fredf: Learning to forecast in the frequency domain. In *International Conference on Learning Representations*, volume 2025, pp. 6893–6922, 2025.

S Wang, H Wu, X Shi, T Hu, H Luo, L Ma, JY Zhang, and J Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*.

Xing Wang, Zhendong Wang, Kexin Yang, Junlan Feng, Zhiyan Song, Chao Deng, et al. Mppn: multi-resolution periodic pattern network for long-term time series forecasting. *arXiv preprint arXiv:2306.06895*, 2023.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024a.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37:469–498, 2024b.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Siqiao Xue, Zhaoyang Zhu, Wei Zhang, Rongyao Cai, Rui Wang, Yixiang Mu, Fan Zhou, Jianguo Li, Peng Di, and Hang Yu. Quitobench: A high-quality open time series forecasting benchmark. *arXiv preprint arXiv:2603.26017*, 2026.

Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3634–3640, 2018.

Guoqi Yu, Jing Zou, Xiaowei Hu, Angelica I Aviles-Rivero, Jing Qin, and Shujun Wang. Revitalizing multivariate time series forecasting: Learnable decomposition with inter-series dependencies and intra-series variations modeling. In *International Conference on Machine Learning*, pp. 57818–57841. PMLR, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Jiawen Zhang, Xumeng Wen, Zhenwei Zhang, Shun Zheng, Jia Li, and Jiang Bian. Probts: Benchmarking point and distributional forecasting across diverse prediction horizons. *Advances in Neural Information Processing Systems*, 37:48045–48082, 2024.

Yang Zhang, Lingbo Liu, Xinyu Xiong, Guanbin Li, Guoli Wang, and Liang Lin. Long-term wind power forecasting with hierarchical spatial-temporal transformer. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 6308–6316, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4120–4131, 2025.

Zhuohang Zhu, Haodong Chen, Qiang Qu, and Vera Chung. Fincast: A foundation model for financial time-series forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4539–4549, 2025.

A APPENDIX

This appendix provides more related work, additional details, and discussion. It is organized as follows:

- Appendix A.1 provides more related work.
- Appendix A.2 presents the data processing, baseline methods, and implementation details.
- Appendix A.3 presents additional experimental results on model efficiency.
- Appendix A.4 presents further discussion on the limitations of existing methods, including challenges in exploiting exogenous variables and performing long-horizon forecasting.

A.1 FULL RELATED WORK

Time series forecasting models. These benchmarks have driven rapid progress in time series forecasting, where methods have evolved from traditional statistical models (Box & Pierce, 1970; Hyndman et al., 2008; Harvey & Harvey, 1990; Toda & Phillips, 1994; Martínez et al., 2019) to deep learning architectures (Zhou et al., 2021; Lin et al., 2025b; Zeng et al., 2023; Lin et al., 2025a; Huang et al., 2023; Nie et al., 2023) and, more recently, foundation models (Jiang et al., 2025; Jin et al., 2024; Liu et al., 2025b; Hoo et al., 2025). Deep forecasting models can be broadly grouped into channel-independent models (Nie et al., 2023; Lin et al., 2024c; Zeng et al., 2023; Shao et al., 2022; Huang et al., 2025), multivariate models (Hu et al., 2026; 2025; Zhang & Yan, 2023; Lin et al., 2025a; Huang et al., 2023), and exogenous-aware models (Wang et al., 2024b; Chen et al., 2026; Das et al., 2023a; Lu et al., 2024; Qiu et al., 2026b). Channel-independent models improve robustness by modeling each target variable separately, multivariate models capture cross-variable dependencies, and exogenous-aware models incorporate external covariates for context-aware forecasting. Time series foundation models are pretrained on large-scale cross-domain datasets to improve transferability, generalization, and zero-shot forecasting (Ansari et al., 2025; Siru et al., 2025; Liu et al., 2026).

However, existing forecasting models are still not fully aligned with the requirements of ride-hailing forecasting. Many exogenous-aware models mainly focus on historical or continuous covariates (Zhou et al., 2025; Wang et al., 2024b; Chen et al., 2026; Qiu et al., 2026b), whereas ride-hailing forecasting requires effective use of future-known categorical signals, such as holiday types and special event indicators. Moreover, models need to capture area-dependent exogenous responses under heterogeneous urban contexts, where the same external factor may induce different demand patterns across areas. Existing models also tend to emphasize pointwise accuracy (Zhou et al., 2025; Lin et al., 2024a; Wang et al., 2024b; Li et al., 2026), while long-horizon operational planning further requires trend quality and near-term reliability. These limitations motivate RideBench to jointly evaluate exogenous utilization, multi-area generalization, and long-horizon planning capability.

A.2 BENCHMARK

A.2.1 DATA PROCESSING

Dataset splitting. RideBench constructs forecasting samples using a sliding-window strategy and splits them chronologically. The first three years are used for training, the following six months for validation, and the final six months for testing. A sample is assigned based on whether its prediction window falls entirely within the corresponding split, while its input window can access previous observations to preserve realistic historical context without future leakage.

Global multi-area training. Each temporal window from the 200 areas is treated as an independent sample, and all samples are jointly used to train a shared global model. This setting improves scalability and requires models to capture transferable temporal patterns while adapting to spatial heterogeneity.

Normalization. Endogenous variables and continuous exogenous variables are standardized using training-set statistics only, while discrete exogenous variables remain as integer identifiers. Predictions are inverse-transformed before evaluation to compute metrics on the original scale.

A.2.2 BASELINES

Simple Statistical Baselines include SeasonalNaive (Hyndman & Athanasopoulos, 2018) and SeasonalMean. SeasonalNaive directly repeats values from the previous seasonal cycle, while SeasonalMean forecasts each future step using the average of historical observations at the same seasonal position within the input window. These two methods capture recurring daily and weekly patterns without model training, providing simple references for assessing the gains of learned forecasting models.

Channel-Independent Forecasting Models include DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), PETformer (Lin et al., 2024c), STID (Shao et al., 2022), RMLP (Li et al., 2023), SegRNN (Lin et al., 2025b), SparseTSF (Lin et al., 2024b), PhaseFormer (Niu et al., 2026), TimeBase (Huang et al., 2025), and TimeMixer (Wang et al.). These models are evaluated on all endogenous forecasting targets, but adopt a channel-independent modeling strategy, where each endogenous variable is modeled based on its own temporal evolution. This group allows us to examine how much predictive power can be obtained from individual temporal dynamics alone, without explicitly modeling dependencies among endogenous variables in the multi-target ride-hailing forecasting task.

Multivariate Forecasting Models include Crossformer (Zhang & Yan, 2023), CrossGNN (Huang et al., 2023), DUET (Qiu et al., 2025), Leddam (Yu et al., 2024), ModernTCN (Luo & Wang, 2024), PMDformer (Hu et al., 2026), SOFTS (Han et al., 2024), TimeFilter (Hu et al., 2025), TQNet (Lin et al., 2025a), and iTransformer (Liu et al., 2024). These models also forecast all endogenous targets, but further capture joint structures among variables through cross-channel dependencies or shared multivariate representations. This group allows us to evaluate whether interactions among different marketplace signals provide additional benefits beyond independent temporal modeling.

Exogenous-Aware Forecasting Models include CATS (Lu et al., 2024), CrossLinear (Zhou et al., 2025), DAG (Qiu et al., 2026b), TiDE (Das et al., 2023a), TimeXer (Wang et al., 2024b), and XLinear (Chen et al., 2026). These models are closely aligned with the ride-hailing forecasting scenario considered in RideBench, where future-known external factors such as weather, holidays, and special events may substantially affect marketplace dynamics. However, most existing exogenous-aware models are primarily designed for continuous covariates, whereas RideBench contains both continuous and discrete exogenous variables. To make these methods applicable to our setting, we adapt discrete exogenous variables using three encoding strategies: ordinal encoding, one-hot encoding, and learnable categorical embeddings. For models that do not originally expose future covariates to the prediction horizon, we further modify their input interfaces so that future-known exogenous variables can be incorporated during forecasting.

Time Series Foundation Models include Chronos-2 (Ansari et al., 2025), Moirai 2.0 (Liu et al., 2025a), TimesFM 2.5 (Das et al., 2023b), and Timer-S1 (Liu et al., 2026). These models can be directly applied to downstream forecasting tasks in a zero-shot manner without task-specific training. Most of them follow a univariate forecasting paradigm, while Chronos-2 is the only one that explicitly supports multivariate endogenous interactions and future-known covariates, including both continuous and discrete exogenous variables. We include these models in RideBench to examine how current time series foundation models perform in realistic, fine-grained, and exogenous-driven ride-hailing scenarios.

A.2.3 IMPLEMENTATION DETAILS

All experiments are implemented in Python 3.11 with PyTorch (Paszke et al., 2019) and conducted on a cluster of eight NVIDIA RTX A6000 GPUs. Forecasting samples are generated with task-specific sliding-window strides: 48 steps, i.e., one day, for week-ahead forecasting, and 336 steps, i.e., one week, for 8-week-ahead forecasting.

Models are trained using the MAE loss and optimized with AdamW (Loshchilov & Hutter, 2019) under a unified hyperparameter search protocol. For each method, we search the learning rate from $\{3\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}, 1\times10^{-3}, 3\times10^{-3}\}$ and report the results under the best configuration. Unless otherwise specified, we use a batch size of 128 and reduce it when out-of-memory issues occur. Early stopping is applied if the validation performance does not improve for three consecutive epochs. For models that support exogenous inputs, future-known exogenous variables are provided

Table 4: Feature descriptions of the Ride-Hailing dataset.

Feature	Type	Description
Temporal Features		
day_index	Integer	Days since the earliest observation (0–1460).
time	Time	Time of day.
half hour of day	Integer	Half-hour interval index (0–47).
day of week	Integer	Day-of-week index (0–6).
Identifier		
area id	Integer	Anonymous area identifier.
Endogenous Targets		
endo 1	Float	Anonymous endogenous target 1.
endo 2	Integer	Anonymous endogenous target 2.
endo 3	Float	Anonymous endogenous target 3.
endo 4	Float	Anonymous endogenous target 4.
endo 5	Integer	Anonymous endogenous target 5.
endo 6	Float	Anonymous endogenous target 6.
endo 7	Float	Anonymous endogenous target 7.
endo 8	Integer	Anonymous endogenous target 8.
endo 9	Float	Anonymous endogenous target 9.
Exogenous Variables		
weather factor	Float	Weather disturbance factor.
large scale event 1	Binary	Indicator for large-scale event type 1.
large scale event 2	Binary	Indicator for large-scale event type 2.
public holiday	Integer	Public holiday category code (0–9).
traditional festival	Integer	Traditional festival category code (0–9).
western festival	Integer	Western festival category code (0–14).

through their model-specific interfaces; otherwise, only historical endogenous variables are used. For completeness, Table 4 provides a more detailed description of the Ride-Hailing dataset features.

A.3 Accuracy–Efficiency Trade-off

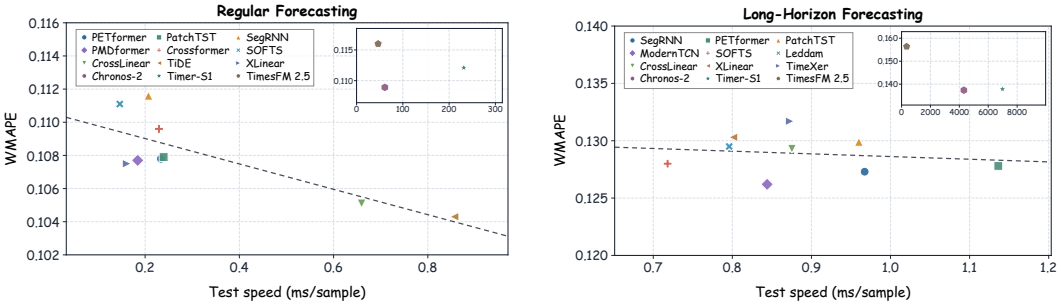


Figure 7: Efficiency of representative forecasting models under regular and long-horizon settings. Increasing model complexity improves regular forecasting; however, such improvements diminish in long-horizon forecasting.

While Tables 2 and 3 focus on overall forecasting performance, computational efficiency is also critical for practical deployment. Ride-hailing forecasting is typically performed over large-scale areas, multiple target variables, and repeated rolling prediction windows. Thus, models with excessive inference latency may be difficult to deploy even if they achieve competitive accuracy. To examine this trade-off, Figure 7 compares the forecasting accuracy and inference cost of representative models.

Figure 7 shows that task-specific deep models generally provide a more favorable accuracy–efficiency balance. Their architectures are usually lightweight enough for efficient batched inference under the same benchmark protocol. In contrast, time series foundation models tend to incur

substantially higher inference cost due to their larger model scale and, in many cases, prediction paradigms that decompose the original multi-target task into multiple sub-tasks, such as independent univariate forecasts.

The trade-off also differs between regular and long-horizon forecasting. In regular week-ahead forecasting, higher computational cost is more likely to bring accuracy gains, suggesting that larger capacity and better exogenous feature utilization can still be beneficial. However, in 8-week-ahead forecasting, this relationship becomes much weaker. This suggests that long-horizon performance is not simply constrained by model capacity, but also by the difficulty of capturing stable long-term trajectories.

These results suggest different research priorities for the two forecasting settings. For regular week-ahead forecasting, improving pointwise accuracy remains valuable, as larger capacity and stronger exogenous-aware architectures may further improve the use of future-known covariates. For long-horizon forecasting, however, simply increasing model complexity to reduce pointwise errors may be less effective. Future work should place more emphasis on trend-aware modeling, including the accurate prediction of demand directions, trajectory shapes, and stable long-term temporal structures.

A.4 Discussion

By benchmarking representative deep learning forecasting methods developed in recent years, RideBench demonstrates the encouraging progress achieved by current models. However, it also shows that existing methods still fall short of the requirements of real-world ride-hailing systems. We summarize two key gaps below.

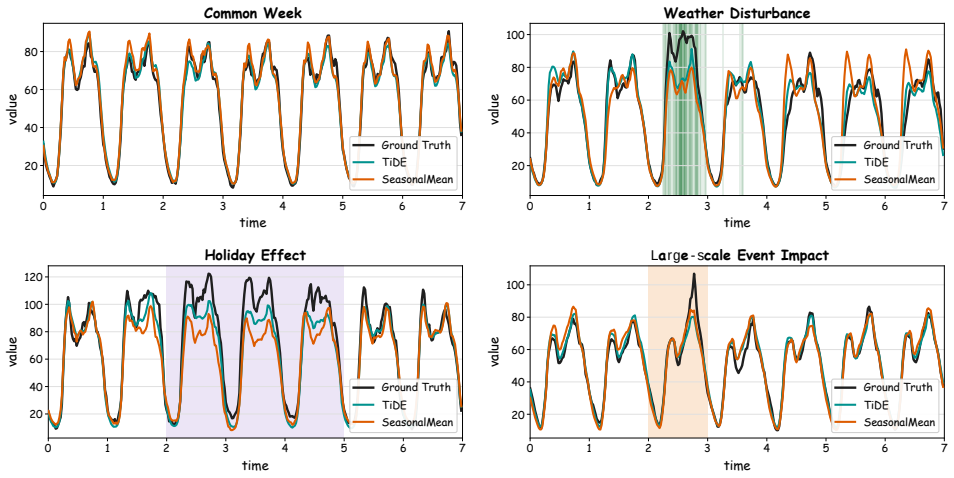


Figure 8: Forecasting visualization under different evaluation scenarios. The shaded regions denote scenario-specific effects. In the weather disturbance scenario, darker shades indicate stronger weather factors.

Exogenous-driven pattern modeling remains insufficient. Although exogenous-aware models generally benefit from future-known external variables, they still cannot fully capture how external factors reshape ride-hailing dynamics. As shown in Figure 8, even the best-performing exogenous-aware model, TiDE (Das et al., 2023a), only partially follows the ground-truth trajectory under strong external perturbations. This suggests that current models can use exogenous variables as auxiliary signals, but still lack sufficient capability to model disturbance-induced pattern changes. This limitation arises from both data side and model side.

On the data side, exogenous effects in ride-hailing systems are highly heterogeneous and temporally structured. Weather impacts can be nonlinear or delayed; holidays involve pre-holiday, in-holiday, and post-holiday phases; and special events are sparse, irregular, and temporally diverse. Moreover, the same external factor may trigger different demand responses across areas due to differences in land use, commuting structure, supply-demand conditions, and event exposure.

On the model side, many existing exogenous-aware methods are not natively designed for the type of future-known categorical covariates considered in RideBench. Models such as CrossLinear (Zhou et al., 2025) and TimeXer (Wang et al., 2024b) mainly focus on numerical covariates or historical contextual variables, whereas ride-hailing forecasting requires effective use of categorical holiday semantics, event types, and other future-known external indicators. To make these methods applicable to RideBench, we extend their input interfaces to the prediction horizon and introduce categorical adapters such as one-hot encoding or learnable embeddings. However, these adaptations are pragmatic compromises rather than principled solutions. They allow models to access categorical variables, but do not necessarily enable them to understand their semantic structure or heterogeneous effects across areas.

Long-horizon forecasting requires better multi-objective balance. RideBench further shows that no existing model dominates all aspects of 8-week-ahead forecasting, highlighting the difficulty of balancing multiple forecasting objectives. As shown in Table 3, models with strong pointwise accuracy do not necessarily achieve the best trend quality, while models that preserve near-term accuracy may still fail to capture long-term trajectory changes.

However, this trade-off is important for practical ride-hailing deployment. A desirable forecasting model should support both short-term operations and long-term planning: it should provide accurate near-term forecasts for operational decisions, reliable long-term trends for supply and resource planning, and stable performance over extended horizons. Such a model would reduce the need to maintain separate forecasting systems for short-term dispatching and long-term planning. However, existing benchmarks rarely evaluate these requirements jointly, which may partly explain why current models pay limited attention to this multi-objective balance.

Overall, RideBench suggests that real-world ride-hailing forecasting requires models that can jointly handle future-known exogenous variables, categorical event semantics, spatial heterogeneity, and long-horizon planning requirements. We hope Ride-Hailing and RideBench can help the community develop more exogenous-aware, scalable, and deployment-oriented forecasting models.