ExoRAF: Exogenous Residual Retrieval for Time Series Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Effectively leveraging exogenous variables remains a longstanding challenge in real-world time series forecasting. The key insight of this paper is that exogenous variables are more effectively utilized by modeling forecasting residuals rather than directly predicting the target series. While an endogenous-only backbone captures the intrinsic temporal dynamics of the target series, exogenous variables mainly explain the scenario-dependent residual variations that cannot be captured by historical observations alone. Moreover, similar exogenous contexts often induce similar residual patterns. Motivated by the insight, we propose **ExoRAF**, an exogenous residual retrieval method for time series forecasting. ExoRAF builds a residual memory from historical exogenous contexts, retrieves relevant residual information using the current exogenous context, and leverages the retrieved residual information to generate correction terms that refine the backbone forecasts. The resulting decoupling of endogenous temporal modeling and exogenous residual explanation makes ExoRAF a plug-and-play module for existing forecasting models. Extensive experiments on diverse real-world datasets demonstrate that ExoRAF consistently improves strong forecasting backbones across a wide range of application scenarios.

1 Introduction

Time series forecasting plays a critical role in many real-world domains, including traffic flow (Yu et al., 2018; Zheng et al., 2020), energy management (Fiot & Dinuzzo, 2016), and weather modeling (Liang et al., 2023). Recent advances in deep forecasting models (Wang et al., 2024a; Qiu et al., 2024; Shchur et al., 2025) have achieved remarkable progress by leveraging powerful architectures, including convolutional networks (He et al., 2016), transformers (Vaswani et al., 2017; Zhou et al., 2021), and state-space models (Gu & Dao, 2023). Despite these advances, many existing forecasting methods (Liu et al., 2024; Niu et al., 2026; Wang et al.) primarily rely on historical observations of the target series, i.e., endogenous variables. However, real-world forecasting is often affected by various external factors. For example, ride-hailing demand (Geng et al., 2019; Hao et al., 2026) can be significantly influenced by weather, holidays, and large-scale events (e.g., major sporting events and concerts). Incorporating such exogenous variables provides valuable contextual information and has the potential to improve forecasting accuracy and support more effective decision-making.

Existing exogenous-aware forecasting methods (Zhou et al., 2025; Chen et al., 2026) typically leverage exogenous variables through specifically designed interaction mechanisms. For example, TimeXer (Wang et al., 2024b) treats exogenous variables as variate tokens and employs cross-attention to facilitate their interaction with endogenous variables. DAG (Qiu et al., 2026) captures exogenous and endogenous dependencies across both temporal and channel dimensions, while GCGNet (Li et al., 2026) constructs relational graphs to align the temporal and channel-wise dependencies of predictions with ground truth. These approaches tightly couple exogenous modeling with architectures or training objectives, limiting their flexibility and applicability to diverse forecasting backbones. Therefore, this raises a question: *How can exogenous information be effectively leveraged in a plug-and-play manner while preserving the original forecasting architecture?*

In this work, we revisit the role of exogenous variables from a different perspective. Rather than directly modeling exogenous variables as additional predictors, we find that they can serve as contextual signals for explaining forecasting residuals. An effective endogenous-only forecasting

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.

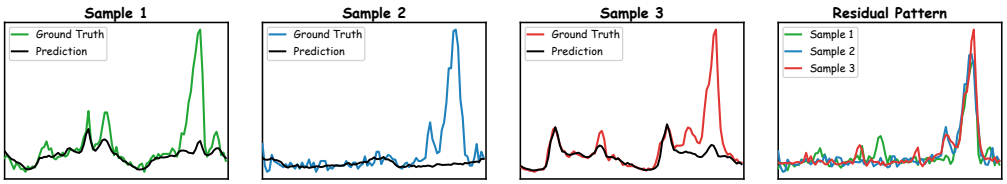

Figure 1: Samples under the holiday scenario exhibit **similar residual patterns** on the Ride-Hailing dataset. This observation suggests that, rather than directly modeling the target series with exogenous variables, we can exploit their contextual similarity to retrieve and model recurring residual patterns from historical data.

model (Nie et al., 2023; Liu et al., 2024) can often capture the intrinsic temporal dynamics of the target series, while exogenous factors mainly introduce scenario-specific deviations. Our empirical observation reveals that these deviations are highly structured: samples with similar exogenous contexts tend to exhibit similar residual patterns. As shown in Figure 1, samples under the same holiday scenario exhibit consistent residual structures on the Ride-Hailing dataset, suggesting that exogenous variables can effectively guide residual retrieval from historical scenarios.

Motivated by this insight, we propose **ExoRAF**, an exogenous residual retrieval framework for time series forecasting. Technically, ExoRAF constructs a residual memory from historical exogenous contexts and forecasting residuals, and uses the current exogenous context as a query to retrieve relevant residual patterns for generating correction terms. By decoupling endogenous temporal modeling from exogenous residual explanation, ExoRAF provides a lightweight and model-agnostic approach that can be seamlessly integrated with diverse forecasting backbones. We evaluate ExoRAF on the large-scale dataset across diverse application scenarios. Experiments with multiple forecasting backbones demonstrate that ExoRAF improves forecasting performance. Moreover, ExoRAF effectively enhances the utilization of exogenous variables in time series foundation models.

In summary, our contributions are as follows:

- We introduce a new perspective for utilizing exogenous variables in time series forecasting, where exogenous contexts serve as retrieval signals for identifying and correcting scenario-specific residual patterns rather than directly predicting future values.
- We propose ExoRAF, an exogenous residual retrieval framework that builds a residual memory and generates adaptive correction terms, enabling seamless integration with forecasting backbones.
- Extensive experiments demonstrate that ExoRAF consistently improves diverse forecasting models across various scenarios, highlighting its effectiveness and generality.

2 Related Work

Time Series Forecasting. Time series forecasting has evolved from classical statistical methods (Box & Pierce, 1970) to deep forecasting models (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022) that leverage attention mechanisms (Lin et al., 2025a), temporal decomposition (Zeng et al., 2023; Lin et al., 2024a), multivariate dependency learning (Luo & Wang, 2024; Yu et al., 2024; Lu et al., 2024; Huang et al., 2023), and large-scale pretrained models (Jin et al., 2024; Siru et al., 2025; Hoo et al., 2025; Auer et al., 2026). Despite these advances, effectively utilizing exogenous variables remains challenging. Existing exogenous-aware methods typically incorporate covariates through input concatenation (Zhou et al., 2025) or architecture-specific interaction modules (Li et al., 2026), with representative models including TimeXer (Wang et al., 2024b) and DAG (Qiu et al., 2026). However, these approaches often tightly couple exogenous modeling with specific architectures and struggle to capture sparse, heterogeneous, and context-dependent exogenous effects. In contrast, ExoRAF leverages exogenous contexts as retrieval signals to identify and correct forecasting residuals, providing a plug-and-play framework compatible with diverse forecasting backbones.

Retrieval-augmented forecasting. Retrieval-augmented methods enhance forecasting models by incorporating external knowledge or historical examples. Existing approaches can be broadly cat-

egorized into two groups. The first group (Zhou et al., 2026; Han et al., 2025) retrieves similar historical patterns as auxiliary information for forecasting. The second group (Lyu et al., 2026; Zhang et al., 2024) integrates retrieval mechanisms into time series foundation models, augmenting inputs with temporal or textual information to improve representation learning and forecasting. Different from these approaches, ExoRAF retrieves residual patterns rather than historical patterns or input examples. By exploiting the relationship between exogenous contexts and forecasting errors, ExoRAF provides targeted residual guidance for forecasting.

3 METHODOLOGY

In an exogenous-aware time series forecasting task, we define historical endogenous variables $X^{\text{endo}} \in \mathbb{R}^{D\times T}$, historical exogenous variables $X^{\text{exo}} = [X_c^{\text{exo}}; X_d^{\text{exo}}] \in \mathbb{R}^{K\times T}$, and known future exogenous variables $Y^{\text{exo}} = [Y_c^{\text{exo}}; Y_d^{\text{exo}}] \in \mathbb{R}^{K\times H}$. Given these inputs, the model aims to forecast the future endogenous series $\hat{Y}^{\text{endo}} \in \mathbb{R}^{D\times H}$, where D is the number of endogenous variables, K is the number of exogenous variables, T is the lookback length, and H is the forecasting horizon. X_c^{exo} and Y_c^{exo} denote continuous exogenous variables, while X_d^{exo} and Y_d^{exo} denote discrete exogenous variables. Let $\mathcal{F}_\theta$ denote the endogenous-only forecasting model. Given X^{endo}, the model produces the forecast $\hat{Y}^{\text{endo}} = \mathcal{F}_\theta(X^{\text{endo}})$, while the corresponding ground-truth future series is denoted by Y^{endo}.

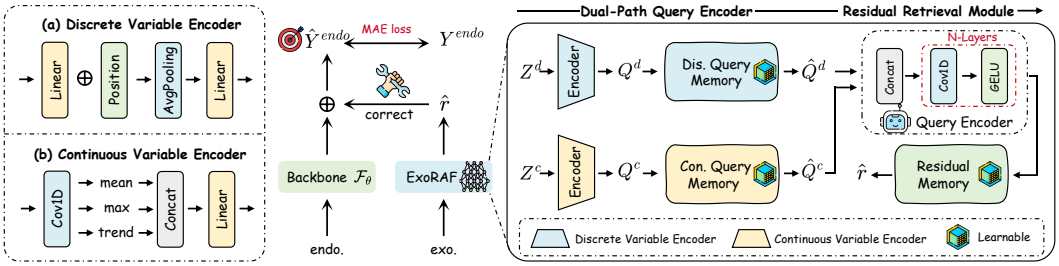

Figure 2: Model architecture of the proposed ExoRAF.

The key idea of ExoRAF is to exploit exogenous variables by modeling residuals rather than the target series. An endogenous-only forecasting backbone captures the dominant endogenous temporal patterns, while exogenous variables explain the remaining scenario-specific residuals. Accordingly, ExoRAF constructs a residual memory that stores residual patterns under historical exogenous contexts and retrieves relevant residual information for the current context. As illustrated in Figure 2, ExoRAF consists of two components: a dual-path query encoder and a residual retrieval module. the dual-path query encoder encodes discrete and continuous exogenous variables into query representations, while the residual retrieval module retrieves and aggregates relevant residual from the residual memory to produce residual corrections.

3.1 DUAL-PATH QUERY ENCODER

Discrete Query Encoder. For discrete exogenous variables, such as holidays and social events, their impacts often exhibit pronounced lead–lag effects. To capture these relative temporal dependencies, we adopt a relative positional encoding method. Specifically, given the discrete exogenous variables $Z^{dis} = (X_d^{\text{exo}}; Y_d^{\text{exo}})$, the relative positional embedding encodes the temporal distance between each forecasting step and its nearest active event, enabling the model to distinguish the heterogeneous influences before, during, and after the event. Formally, the encoding process is defined as

$$H_{i,t}^{dis} = \text{Linear}(Z_{i,t}^{dis}) + \text{PosEmbedding}(p_t), \tag{1}$$

where i indexes individual variables and p_t denotes the clipped relative distance to the event. The discrete query is obtained as

$$Q^{dis} = \text{LayerNorm}\left(\tanh\left(W_\xi \cdot \text{AvgPooling}(H^{dis})\right)\right) \in \mathbb{R}^d, \tag{2}$$

where d denotes the hidden dimension of the query. The sample-level query produced by the query encoder is inherently instance-specific. Although effective in characterizing individual exogenous

contexts, relying solely on instance-specific queries may limit the model's ability to capture global regularities shared across samples, particularly in large-scale datasets. To alleviate this limitation, we introduce a discrete query memory that refines instance-specific queries by retrieving representative query prototypes shared across samples, thereby providing a stable global prior for subsequent residual retrieval. Specifically, the query memory consists of a set of randomly initialized and learnable keys $K^{dis} \in \mathbb{R}^{M\times d}$ and corresponding values $V^{dis} \in \mathbb{R}^{M\times H\times D}$ which encode representative query prototypes in the latent query space. The refined query is obtained through memory retrieval as follows:

$$\hat{Q}^{dis} = \sum_{m=1}^{M} \text{softmax}\left(\frac{Q^{dis}(K^{dis})^\top}{\sqrt{d}}\right)_m V_m^{dis} \in \mathbb{R}^{H\times D}. \tag{3}$$

Continuous Query Encoder. For continuous exogenous variables $Z^{con} = [X_c^{\text{exo}}; Y_c^{\text{exo}}]$, we first normalize them and the series is encoded by a temporal convolutional layer. It is formulated as:

$$H^{con} = \text{Conv1D}(Z^{con}). \tag{4}$$

To construct more informative queries for continuous exogenous variables, we extract multiple statistics from their hidden representations, including the mean, maximum, and temporal variation trend Δ. These statistics jointly characterize the overall level, salient peaks, and dynamic evolution patterns of continuous covariates. The process is formally defined as follows:

$$Q^{con} = \text{LayerNorm}\left(\tanh\left(W_\psi \cdot [\text{Mean}(H^{con}); \text{Max}(H^{con}); \Delta(H^{con})]\right)\right) \in \mathbb{R}^d. \tag{5}$$

Similar to the discrete exogenous variables, we construct a learnable memory to capture global regularities, with keys $K^{con} \in \mathbb{R}^{M\times d}$ and values $V^{con} \in \mathbb{R}^{M\times H\times D}$. The process can be formulated as follows:

$$\hat{Q}^{con} = \sum_{m=1}^{M} \text{softmax}\left(\frac{Q^{con}(K^{con})^\top}{\sqrt{d}}\right)_m V_m^{con} \in \mathbb{R}^{H\times D}. \tag{6}$$

3.2 Residual Retrieval Module

Exogenous variables often affect the target series through complex interactions rather than independent additive effects. For instance, the joint occurrence of rainfall and holidays may induce demand deviations that cannot be explained by either factor alone. Therefore, ExoRAF models their interactions through a residual retrieval process. Specifically, the continuous and discrete exogenous variables are first encoded into query representations and fused by a lightweight query encoder:

$$Q = \text{QE}(\hat{Q}^{con}, \hat{Q}^{dis}) \in \mathbb{R}^{H\times D}. \tag{7}$$

The interaction query retrieves residual prototypes from the memory bank via attention-based retrieval. Since Q is a step-wise query representation, we first obtain a retrieval vector $q = \text{Proj}(\text{Pool}(Q)) \in \mathbb{R}^d$. The residual memory contains keys $K \in \mathbb{R}^{M\times d}$ and values $V \in \mathbb{R}^{M\times H\times D}$, where each key $K_m \in \mathbb{R}^d$ is used for similarity computation and each value $V_m \in \mathbb{R}^{H\times D}$ stores a residual prototype over all forecasting steps and target variables. The retrieval is then defined as:

$$\alpha = \text{softmax}\left(\frac{qK^\top}{\sqrt{d}}\right), \qquad r = \sum_{m=1}^{M} \alpha_m V_m \in \mathbb{R}^{H\times D}. \tag{8}$$

However, the retrieved residual should not be uniformly injected into every future step or target dimension. To enable adaptive residual correction, ExoRAF introduces two complementary gates. The step-wise gate $\mathcal{G}_s \in \mathbb{R}^{H\times D}$ controls the temporal allocation of residual corrections, while the channel-wise gate $\mathcal{G}_c \in \mathbb{R}^{1\times D}$ adjusts their magnitude across target variables. The residual correction is computed as:

$$\hat{r} = r \odot \mathcal{G}_s \odot \mathcal{G}_c, \tag{9}$$

and is used to refine the backbone prediction:

$$\tilde{Y}^{\text{endo}} = \hat{Y}^{\text{endo}} + \hat{r}. \tag{10}$$

Through this design, ExoRAF retrieves scenario-dependent residual patterns from exogenous contexts and selectively applies them to correct the backbone forecast. ExoRAF is optimized using the backbone loss function, and a theoretical analysis is provided in Appendix A.1.

4 EXPERIMENTS

Dataset. We conduct evaluations on one large-scale dataset and four small-scale datasets, all of which incorporate discrete and continuous exogenous variables. Specifically, we use the large-scale Ride-Hailing[1] dataset as our primary validation benchmark. The dataset contains ride-hailing records from 200 representative areas over a continuous four-year period, with observations recorded at 30-minute intervals, resulting in a total of 14,025,600 time-stamped observations. It further incorporates three types of exogenous scenarios: *weather disturbance*, *holiday effect*, and *large-scale event impact*. More details are provided in Appendix A.2.

Baseline. We conduct adaptations for four categories of models, including channel-independent forecasting models (DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), SparseTSF (Lin et al., 2024b), SegRNN (Lin et al., 2025b), PETformer (Lin et al., 2024c), and TimeBase (Huang et al., 2025)), multivariate forecasting models (Crossformer (Zhang & Yan, 2023), iTransformer (Liu et al., 2024), PMDformer (Hu et al., 2026), DUET (Qiu et al., 2025), TimeFilter (Hu et al., 2025), and TQNet (Lin et al., 2025a)), exogenous-aware forecasting models (CrossLinear (Zhou et al., 2025), DAG (Qiu et al., 2026), TiDE (Das et al., 2023a), TimeXer (Wang et al., 2024b), and XLinear (Chen et al., 2026)), and time series foundation models (Timer-S1 (Liu et al., 2026), TimesFM 2.5 (Das et al., 2023b), and Moirai 2.0 (Liu et al., 2025)).

Setup. All experiments are implemented in Python 3.11 with PyTorch and conducted on a cluster of eight NVIDIA RTX A6000 GPUs. We utilize the RideBench[2] for unified evaluation. For the large-scale Ride-Hailing dataset, RideBench defines two forecasting settings. The first is regular forecasting, where models predict the next 336 steps, using 1,344 historical steps. The second is long-horizon forecasting, where models predict the next 2,688 steps, using 4,704 historical steps. For the small-scale datasets, we follow the experimental protocol of EventTSF (Ge et al., 2025), where models use the historical 168 steps as input to forecast the next 92 steps. To comprehensively evaluate ExoRAF's performance, the Weighted Mean Absolute Percentage Error (WMAPE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) metrics are employed. For additional details, please refer to the Appendix A.2.

4.1 MAIN RESULT

Results on Small-Scale Datasets. As shown in Table 1, exogenous-aware models outperform endogenous-only models, demonstrating the effectiveness of incorporating exogenous information for forecasting. Moreover, integrating ExoRAF consistently improves the performance of all backbones across the four datasets, highlighting the robustness and broad applicability of our approach. Notably, several backbones with ExoRAF surpass existing exogenous-aware methods, further validating the effectiveness of ExoRAF in leveraging exogenous information.

Table 1: Results of exogenous-aware forecasting methods and ExoRAF on the four small-scale datasets. Average rank of exogenous-aware models across datasets and metrics (lower is better).

Dataset	New York Taxi			Synthetic			Houston's Temperature			New York's Temperature			Rank
Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	Avg.
XLinear	0.4321	5.604	9.75	0.0667	0.0451	0.1330	0.1155	8.16	10.68	0.1688	10.89	14.08	4
TimeXer	0.4479	5.809	9.77	0.0703	0.0476	0.0892	0.1165	8.23	10.69	0.1593	10.28	13.60	6
CrossLinear	0.4263	5.528	9.75	0.0684	0.0463	0.1362	0.1375	9.72	13.33	0.1787	11.53	14.84	8
TiDE	0.4899	6.353	10.32	0.0673	0.0455	0.1099	0.1487	10.51	13.90	0.1750	11.29	15.16	9
DAG	0.4461	5.785	9.81	0.0918	0.0622	0.1363	0.1545	10.92	14.61	0.1848	11.92	15.95	10
iTransformer	0.4398	5.704	10.08	0.0768	0.0520	0.1383	0.1550	10.95	14.80	0.1874	12.09	16.16	–
+ExoRAF	**0.4070**	**5.278**	**8.57**	**0.0496**	**0.0336**	**0.0891**	**0.1203**	**8.51**	**11.33**	**0.1758**	**11.34**	**14.29**	1
PatchTST	0.4544	5.893	10.07	0.0803	0.0544	0.1377	0.1526	10.78	14.59	0.1853	11.95	15.78	–
+ExoRAF	**0.4000**	**5.188**	**9.07**	**0.0513**	**0.0347**	**0.0945**	**0.1224**	**8.65**	**11.61**	**0.1730**	**11.16**	**14.05**	2
TimeBase	0.4584	5.945	10.15	0.1223	0.0828	0.1681	0.1595	11.27	15.01	0.1970	12.71	16.83	–
+ExoRAF	**0.4166**	**5.403**	**9.23**	**0.0641**	**0.0434**	**0.0989**	**0.1182**	**8.36**	**11.41**	**0.1650**	**10.65**	**13.42**	2
DLinear	0.4817	6.247	10.25	0.0991	0.0671	0.1450	0.1493	10.55	14.10	0.1894	12.22	15.84	–
+ExoRAF	**0.4389**	**5.692**	**9.39**	**0.0578**	**0.0392**	**0.0983**	**0.1168**	**8.26**	**11.02**	**0.1714**	**11.06**	**14.13**	5
DUET	0.4281	5.552	9.78	0.0828	0.0561	0.1361	0.1470	10.39	13.98	0.1773	11.44	15.21	–
+ExoRAF	**0.4164**	**5.400**	**9.54**	**0.0546**	**0.0369**	**0.0900**	**0.1207**	**8.53**	**11.40**	**0.1762**	**11.37**	**14.53**	7
Avg. Improve	-8.03%	-8.03%	-8.96%	-38.97%	-39.00%	-34.84%	-21.55%	-21.49%	-21.63%	-7.84%	-7.83%	-11.61%	–

[1] https://anonymous.4open.science/r/ridebench-338A/datasets

[2] https://anonymous.4open.science/r/ridebench-338A

Results on Large-Scale Dataset. As shown in Table 2, ExoRAF improves both types of backbones across all scenarios. Compared with weather disturbance and holiday effect, large-scale event impact scenarios introduce additional challenges. This is because large-scale events occur less frequently, making it difficult for the module to learn sufficient and stable residual prototypes, which consequently limits the performance improvements. Notably, PMDformer (Hu et al., 2026) augmented with ExoRAF outperforms existing exogenous-aware forecasting methods, demonstrating the effectiveness of ExoRAF in diverse exogenous scenarios. Detailed results are provided in Appendix A.6, with the full results reported in Table 11 and the long-horizon forecasting results reported in Table 7 of Appendix A.3.

Table 2: Regular week-ahead forecasting results of ExoRAF on the large-scale Ride-Hailing dataset. Average rank of exogenous-aware models across scenarios and metrics (lower is better).

Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact			Rank
Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	Avg.
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	2
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	4
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	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	9
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	10
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	–
+ExoRAF	**0.1027**	**3.939**	**12.52**	**0.1238**	**5.168**	**18.62**	**0.1327**	**5.505**	**18.40**	**0.1085**	**4.787**	**15.60**	1
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	–
+ExoRAF	**0.1038**	**3.981**	**12.64**	**0.1247**	**5.204**	**18.72**	**0.1346**	**5.584**	**18.75**	**0.1094**	**4.830**	**15.78**	3
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	–
+ExoRAF	**0.1060**	**4.064**	**12.78**	**0.1274**	**5.320**	**19.04**	**0.1344**	**5.569**	**18.61**	**0.1111**	**4.901**	**15.94**	5
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	–
+ExoRAF	**0.1084**	**4.158**	**12.87**	**0.1290**	**5.384**	**18.98**	**0.1361**	**5.637**	**18.80**	**0.1131**	**4.993**	**15.97**	7
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	–
+ExoRAF	**0.1095**	**4.207**	**12.82**	**0.1305**	**5.445**	**18.84**	**0.1395**	**5.787**	**18.67**	**0.1149**	**5.072**	**16.11**	8
Avg. Improve	-4.50%	-4.32%	-5.45%	-9.78%	-9.99%	-9.75%	-10.91%	-10.89%	-7.97%	-3.84%	-3.83%	-4.74%	–

4.2 Ablation Study and Analysis

Table 3: Ablation study of ExoRAF on the Ride-Hailing dataset under the regular forecasting setting.

Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact		
Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE
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
w/o *Dis. Query Mem.*	0.1057	4.048	13.00	0.1283	5.357	19.43	0.1478	6.130	20.02	0.1113	4.908	16.15
w/o *Con. Query Mem.*	0.1061	4.065	13.15	0.1375	5.752	20.62	0.1360	5.639	19.01	0.1119	4.935	16.36
RRM $\rightleftharpoons$ *MLP*	0.1067	4.093	12.83	0.1283	5.356	19.08	0.1389	5.755	18.99	0.1110	4.895	15.82
Full	**0.1038**	**3.981**	**12.64**	**0.1247**	**5.204**	**18.72**	**0.1346**	**5.584**	**18.75**	**0.1094**	**4.830**	**15.78**
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
w/o *Dis. Query Mem.*	0.1094	4.190	13.18	0.1309	5.469	19.53	0.1525	6.320	20.21	0.1144	5.050	16.40
w/o *Con. Query Mem.*	0.1089	4.170	13.17	0.1381	5.778	20.59	0.1373	5.684	18.87	0.1145	5.053	16.42
RRM $\rightleftharpoons$ *MLP*	0.1083	4.147	12.95	0.1301	5.431	19.36	0.1370	5.676	18.83	0.1136	5.013	16.23
Full	**0.1060**	**4.064**	**12.78**	**0.1274**	**5.320**	**19.04**	**0.1344**	**5.569**	**18.61**	**0.1111**	**4.901**	**15.94**

Ablation Study. To validate the effectiveness of each component in ExoRAF, we conduct ablation studies (see Table 3). Overall, the full ExoRAF achieves the best performance across all backbones and scenarios. Removing either query memory component degrades performance: the discrete memory is more critical for holiday effect scenario, while the continuous memory plays a larger role under weather disturbance scenario. This suggests that query memory enhances query representations by aggregating shared residual prototypes beyond individual samples.

Furthermore, we replace the Residual Retrieval Module (RRM) with an MLP-based residual module of comparable parameter size. The MLP variant consistently underperforms ExoRAF, indicating that direct parametric mappings struggle to capture localized and heterogeneous residual patterns. In contrast, RRM adaptively retrieves and combines residual prototypes conditioned on the query, demonstrating that its advantage comes from conditional residual pattern selection rather than increased model capacity.

Analysis of Residuals. To validate the motivation of ExoRAF, we investigate whether exogenous variables provide information for retrieving residual patterns. Specifically, we retrieve nearest neigh-

bors (KNN-selected) in the exogenous query space and compare their residual similarity against randomly selected neighbors. As shown in Figure 3(a), samples with similar exogenous contexts consistently exhibit lower residual MAE across different scenarios, indicating that exogenous variables capture meaningful contextual dependencies beyond historical observations.

Furthermore, Figure 3(b) shows that similar exogenous contexts correspond to highly consistent residual patterns, while the residual patterns retrieved by ExoRAF closely match the ground-truth residual dynamics. It suggests that forecasting residuals are not purely random but contain recurring patterns associated with external contexts, which can be effectively identified through exogenous-aware retrieval. These findings provide empirical evidence for the core design of ExoRAF: leveraging exogenous variables as retrieval signals enables models to capture context-dependent residual dynamics and improve forecasting robustness.

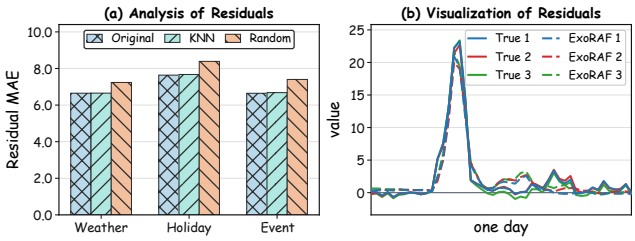

Figure 3: Analysis of residuals from PatchTST.

Table 4: Results of ExoRAF with backbone under exogenous variable Gaussian noise with magnitudes of 5%, 10%, and 20% on the Ride-Hailing dataset. And Std denotes the standard deviation.

	Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact		
	Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE
DLinear	0%	0.1120	4.291	12.88	0.1312	5.476	18.96	0.1398	5.788	18.64	0.1136	5.013	15.67
	5%	0.1120	4.290	12.88	0.1312	5.477	18.96	0.1398	5.787	18.63	0.1136	5.014	15.67
	10%	0.1120	4.290	12.87	0.1312	5.478	18.96	0.1399	5.790	18.62	0.1136	5.013	15.66
	20%	0.1121	4.297	12.86	0.1313	5.482	18.97	0.1402	5.804	18.59	0.1136	5.012	15.64
	Std	0.0000	0.003	0.01	0.0000	0.002	0.00	0.0002	0.007	0.02	0.0000	0.001	0.01
DUET	0%	0.1095	4.207	12.82	0.1305	5.445	18.84	0.1395	5.787	18.67	0.1149	5.072	16.11
	5%	0.1095	4.208	12.82	0.1305	5.446	18.85	0.1395	5.785	18.66	0.1150	5.075	16.10
	10%	0.1096	4.212	12.82	0.1305	5.447	18.86	0.1396	5.789	18.65	0.1151	5.078	16.10
	20%	0.1100	4.226	12.83	0.1305	5.447	18.89	0.1401	5.809	18.64	0.1153	5.090	16.10
	Std	0.0002	0.008	0.00	0.0000	0.001	0.02	0.0002	0.010	0.01	0.0001	0.007	0.00
PatchTST	0%	0.1038	3.981	12.64	0.1247	5.204	18.72	0.1346	5.584	18.75	0.1094	4.830	15.78
	5%	0.1038	3.984	12.64	0.1247	5.204	18.72	0.1347	5.588	18.74	0.1095	4.832	15.78
	10%	0.1040	3.988	12.64	0.1247	5.205	18.72	0.1349	5.593	18.73	0.1095	4.832	15.78
	20%	0.1043	4.000	12.65	0.1248	5.211	18.76	0.1352	5.607	18.71	0.1096	4.837	15.77
	Std	0.0002	0.007	0.00	0.0000	0.003	0.02	0.0002	0.009	0.01	0.0001	0.003	0.00
SparseTSF	0%	0.1084	4.158	12.87	0.1290	5.384	18.98	0.1361	5.637	18.80	0.1131	4.993	15.97
	5%	0.1085	4.159	12.87	0.1290	5.384	18.98	0.1361	5.639	18.79	0.1131	4.993	15.97
	10%	0.1085	4.161	12.87	0.1290	5.384	18.98	0.1362	5.642	18.78	0.1131	4.993	15.96
	20%	0.1087	4.168	12.86	0.1290	5.386	19.00	0.1363	5.644	18.74	0.1132	4.995	15.95
	Std	0.0001	0.004	0.00	0.0000	0.001	0.01	0.0001	0.003	0.02	0.0000	0.001	0.01
iTransformer	0%	0.1060	4.064	12.78	0.1274	5.320	19.04	0.1344	5.569	18.61	0.1111	4.901	15.94
	5%	0.1061	4.065	12.77	0.1274	5.320	19.03	0.1344	5.568	18.61	0.1111	4.902	15.94
	10%	0.1061	4.066	12.77	0.1274	5.320	19.02	0.1344	5.569	18.60	0.1111	4.902	15.93
	20%	0.1062	4.070	12.77	0.1274	5.321	19.02	0.1344	5.570	18.59	0.1111	4.904	15.93
	Std	0.0001	0.002	0.00	0.0000	0.000	0.01	0.0000	0.001	0.01	0.0000	0.001	0.01

Forecasting under Exogenous Variable Noise. Exogenous variables, such as future weather factors, are often difficult to obtain accurately in advance, whereas factors like holidays and large-scale events can usually be determined reliably. This raises an important question: *can ExoRAF still maintain effective forecasting performance under noisy exogenous factors?* To investigate this, we inject Gaussian noise with magnitudes of 5%, 10%, and 20% into the weather factor and report the corresponding results in Table 4.

The results demonstrate that ExoRAF maintains stable performance under various levels of noise perturbations. This robustness can be attributed to its residual retrieval mechanism. Specifically, instead of directly leveraging exogenous variables for prediction, ExoRAF retrieves historical residuals based on the similarity of exogenous scenarios. Consequently, moderate noise in exogenous variables has limited impact on the query matching process.

4.3 Application Analysis

Efficiency Analysis. In real-world deployment, forecasting models require a careful balance between accuracy and efficiency. To evaluate the deployment efficiency of ExoRAF, we compare three widely adopted metrics (Lin et al., 2024b): forecasting performance (i.e., the average WMAPE across four scenarios), parameter count, and inference time per sample. ExoRAF introduces only 0.24M additional parameters, highlighting its lightweight and plug-and-play nature. As shown in Figure 4, incorporating ExoRAF consistently improves the accuracy-efficiency trade-off across different backbones, with PMDformer (Hu et al., 2026) achieving the most significant improvement.

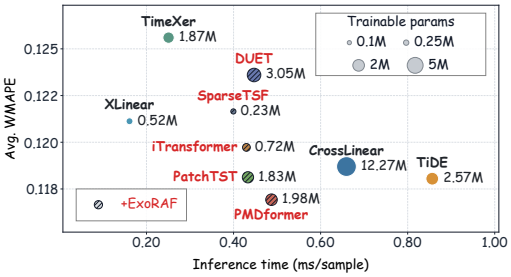

Figure 4: Efficiency comparison between ExoRAF and exogenous-aware forecasting methods.

Zero-Shot Area Forecasting. Ride-hailing platforms typically exhibit strong spatial heterogeneity, where demand patterns vary across areas due to differences in commuting structures, weather sensitivity, and local supply-demand dynamics. Therefore, the ability to generalize to unseen areas is crucial for practical deployment. We conduct zero-shot area forecasting experiments on the Ride-Hailing dataset. Specifically, the 200 areas are randomly divided into two disjoint subsets, with 150 areas used for training and validation and the 50 areas used for testing.

As shown in Table 5, ExoRAF improves the performance of different backbones. The results demonstrate that exogenous contexts across different areas exhibit certain shared characteristics, e.g., weather disturbance and holiday effect. ExoRAF can effectively leverage exogenous variables to capture transferable residual patterns across unseen areas, enabling more robust forecasting under spatial distribution shifts. However, for large-scale event impact scenarios, where cross-area commonality is limited, transferable residual patterns are more difficult to identify, resulting in relatively limited improvements.

Table 5: Zero-shot area forecasting results on the Ride-Hailing dataset under the regular forecasting setting. Average rank of exogenous-aware models across scenarios and metrics (lower is better).

Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact			Rank
Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	Avg.
CrossLinear	0.1054	4.124	11.71	0.1270	5.437	17.24	0.1338	5.654	18.05	0.1127	5.087	16.90	3
XLinear	0.1083	4.232	12.14	0.1339	5.737	18.49	0.1429	6.042	18.87	0.1134	5.111	16.96	7
TimeXer	0.1089	4.275	12.01	0.1387	5.948	18.65	0.1388	5.869	18.49	0.1144	5.161	17.05	8
DAG	0.1080	4.220	12.32	0.1391	5.966	19.18	0.1462	6.175	19.53	0.1138	5.133	17.26	9
TiDE	0.1095	4.273	12.42	0.1413	6.064	19.31	0.1511	6.382	19.86	0.1137	5.123	17.15	10
PMDformer	0.1072	4.186	12.40	0.1386	5.944	19.21	0.1498	6.335	20.00	0.1103	4.968	17.14	–
+ExoRAF	**0.1032**	**4.035**	**11.83**	**0.1251**	**5.356**	**17.41**	**0.1373**	**5.807**	**18.76**	**0.1098**	**4.951**	**16.86**	1
PETformer	0.1070	4.177	12.39	0.1398	6.001	19.27	0.1527	6.460	20.23	0.1107	4.982	17.16	–
+ExoRAF	**0.1024**	**4.005**	**11.78**	**0.1264**	**5.412**	**17.64**	**0.1372**	**5.801**	**18.87**	**0.1077**	**4.850**	**16.50**	2
PatchTST	0.1072	4.186	12.42	0.1394	5.983	19.29	0.1513	6.399	20.15	0.1108	4.993	17.16	–
+ExoRAF	**0.1033**	**4.041**	**11.89**	**0.1256**	**5.380**	**17.53**	**0.1380**	**5.841**	**19.07**	**0.1103**	**4.973**	**16.93**	4
iTransformer	0.1113	4.348	12.51	0.1413	6.064	19.29	0.1544	6.528	20.13	0.1139	5.134	17.24	–
+ExoRAF	**0.1062**	**4.154**	**11.87**	**0.1281**	**5.483**	**17.48**	**0.1386**	**5.856**	**18.65**	**0.1120**	**5.054**	**16.91**	5
SparseTSF	0.1123	4.391	12.65	0.1419	6.095	19.45	0.1565	6.624	20.40	0.1166	5.264	17.22	–
+ExoRAF	**0.1077**	**4.215**	**12.00**	**0.1292**	**5.531**	**17.52**	**0.1418**	**5.996**	**19.09**	**0.1118**	**5.039**	**16.72**	6
Avg. Improve	-4.07%	-3.93%	-4.81%	-9.50%	-9.72%	-9.25%	-9.38%	-9.41%	-6.41%	-1.88%	-1.84%	-2.33%	–

Transfer Evaluation. In real-world forecasting systems, different applications often rely on heterogeneous forecasting architectures. Training a separate exogenous adaptation module for every deployed backbone introduces additional costs. We therefore evaluate whether an ExoRAF module trained with one backbone can be reused by other backbones without fine-tuning. Specifically, we train ExoRAF using PatchTST (Nie et al., 2023) and iTransformer (Liu et al., 2024) as source backbones, respectively, and transfer each trained module to other target backbones. As shown in Figure 5, the transferred ExoRAF module improves performance under weather disturbance and holiday effect scenarios, including when applied to time series foundation models such as Timer-

S1 (Liu et al., 2026), TimesFM 2.5 (Das et al., 2023b), and Moirai 2.0 (Liu et al., 2025). It suggests that residual effects associated with recurring and broadly shared exogenous factors are transferable across forecasting backbones. Full results are provided in Table 13 of Appendix A.6.

In contrast, the improvement under large-scale event impact are limited. This suggests that transferability across backbones depends on the stability of the mapping between exogenous information and residual patterns. Weather and holiday effects are typically more repetitive, allowing different backbones to share similar residual structures under comparable exogenous information. In contrast, large-scale events are more sporadic and localized, and their impacts are difficult to fully characterize from available exogenous information. As a result, the learned residual corrections may entangle event-specific variations with backbone-specific errors, reducing their transferability across architectures.

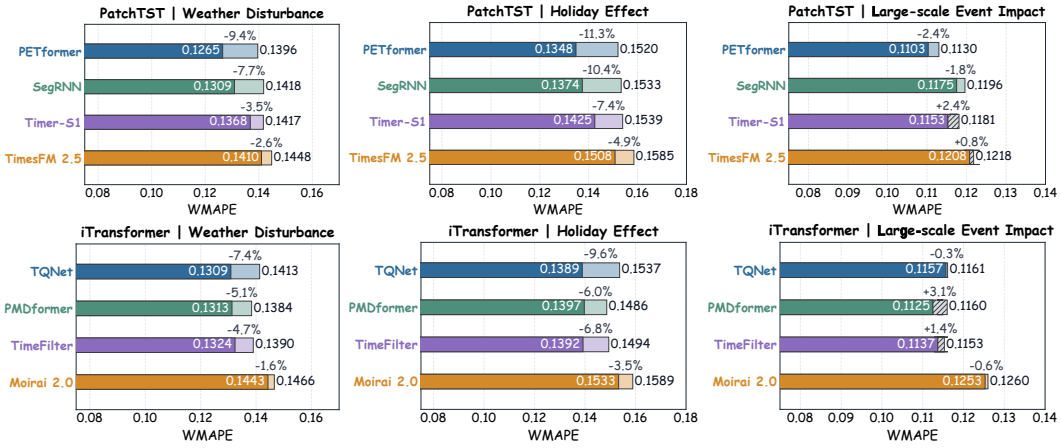

Figure 5: Cross-backbone transfer performance for regular forecasting on the Ride-Hailing dataset.

Furthermore, to further assess the practical effectiveness of ExoRAF, we conduct additional experiments, including evaluations under missing future exogenous variables and parameter sensitivity analyses. Detailed results are provided in the Appendix A.4.

5 Conclusion and Discussion

In this paper, we proposed **ExoRAF**, an exogenous residual retrieval framework for time series forecasting. Unlike existing approaches that directly incorporate exogenous variables as additional inputs, ExoRAF leverages exogenous contexts to retrieve historical residual patterns and generate adaptive corrections, while leaving endogenous temporal modeling to the forecasting backbone. By decoupling endogenous dynamics modeling from exogenous residual explanation, ExoRAF provides a lightweight and plug-and-play solution compatible with diverse forecasting architectures. Extensive experiments across various datasets, backbones, and exogenous scenarios demonstrate that ExoRAF consistently improves forecasting performance.

Limitations. Although ExoRAF demonstrates strong performance across diverse scenarios, it has several limitations. ***i) Memory Scalability.*** As the number of endogenous variables increases, ExoRAF requires a larger residual memory bank to store historical residual patterns, leading to increased memory overhead and retrieval costs. ***ii) Challenges in Extreme Scenarios.*** ExoRAF relies on historical residual patterns to capture recurring temporal behaviors, making it effective in scenarios with relatively stable and repetitive dynamics. However, under extreme events (e.g., severe weather factors), historical patterns may become insufficient to characterize abrupt distribution shifts, which can limit forecasting performance.

Future Work. We observe that these effects often exhibit substantial temporal-scale heterogeneity. For example, weather factors typically influence forecasts over short horizons, whereas holidays may introduce effects lasting for several weeks. Future research could explore more sophisticated multi-scale modeling strategies to better capture the heterogeneous temporal dynamics of exogenous-induced residual effects.

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

We promise that all experimental results can be reproduced. We have released our model code in an anonymous repository: https://anonymous.4open.science/r/ExoRAF-main

REFERENCES

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *Advances in Neural Information Processing Systems*, 38:57529–57580, 2026.

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.

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.

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

Yunfeng Ge, Ming Jin, Yiji Zhao, Hongyan Li, Bo Du, Chang Xu, and Shirui Pan. Eventtsf: Event-aware non-stationary time series forecasting. *arXiv preprint arXiv:2508.13434*, 2025.

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.

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

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

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.

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.

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.

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.

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*, 2025.

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.

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

Sisuo Lyu, Siru Zhong, Tiegang Chen, Weilin Ruan, Qingxiang Liu, Taiqiang Lv, Qingsong Wen, Raymond Chi-Wing Wong, and Yuxuan Liang. Ts-memory: Plug-and-play memory for time series foundation models, 2026. URL https://arxiv.org/abs/2602.11550.

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.

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, 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*, 2026.

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.

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.

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.

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*.

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, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

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.

Huanyu Zhang, Chang Xu, Yi-Fan Zhang, Zhang Zhang, Liang Wang, Jiang Bian, and Tieniu Tan. Timeraf: Retrieval-augmented foundation model for zero-shot time series forecasting, 2024. URL https://arxiv.org/abs/2412.20810.

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.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1234–1241, 2020.

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.

Shiqiao Zhou, Holger Schöner, Zipeng Wu, Edouard Fouché, IAG Wilson, and Shuo Wang. Stationarity-aware retrieval-augmented time series forecasting, 2026. URL https://arxiv.org/abs/2606.04135.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

A Appendix

In this section, we provide additional discussions, experimental details, and analyses to further validate the effectiveness of ExoRAF. The remainder of this section is organized as follows:

- Appendix A.1 presents a theoretical analysis of ExoRAF.
- Appendix A.2 provides additional details on the datasets, setup, and evaluation metrics.
- Appendix A.3 shows additional results of ExoRAF on Ride-Hailing dataset.
- Appendix A.4 provides additional experiments of ExoRAF on real-world applications.
- Appendix A.5 provides visualization to further illustrate the effectiveness of ExoRAF.
- Appendix A.6 reports the full forecasting results on the Ride-Hailing datasets.

A.1 Theoretical Analysis

Let X denote the historical endogenous series, $Z = [X^{\text{exo}}; Y^{\text{exo}}]$ denote the available exogenous information, and Y denote the future endogenous target. Conditioning on a given backbone parameter θ, let $\hat{Y} = \mathcal{F}_\theta(X)$ be the endogenous-only forecast and define the residual as $\tilde{r} = Y - \hat{Y}$. ExoRAF encodes the exogenous context as a query $U = \phi(Z)$ and retrieves a residual correction $\hat{r}(U)$ from memory. The corrected forecast is $\tilde{Y} = \hat{Y} + \hat{r}(U)$. The resulting forecast error satisfies

$$Y - \tilde{Y} = Y - (\hat{Y} + \hat{r}) = \tilde{r} - \hat{r}, \qquad \mathcal{L}_f = \rho(Y - \tilde{Y}) = \rho(\tilde{r} - \hat{r}). \tag{11}$$

Hence, optimizing the corrected forecast is equivalent to fitting the backbone residual, without introducing an additional training objective. Under squared loss, the optimal correction based only on the exogenous query U is the conditional mean residual

$$r^*(U) = \mathbb{E}[\tilde{r} \mid U]. \tag{12}$$

This quantity is the component of the backbone residual that can be explained by the exogenous information.

Proposition 1 (Risk decomposition)**.** *For any residual correction function $\hat{r}(U)$,*

$$\mathcal{L}(\hat{r}) = \mathbb{E}\left[\|\tilde{r} - r^*(U)\|^2\right] + \mathbb{E}\left[\|\hat{r}(U) - r^*(U)\|^2\right]. \tag{13}$$

Consequently, the improvement over the endogenous-only backbone $\hat{r}(U) = 0$ is

$$\mathcal{L}(0) - \mathcal{L}(\hat{r}) = \mathbb{E}\left[\|r^*(U)\|^2\right] - \mathbb{E}\left[\|\hat{r}(U) - r^*(U)\|^2\right]. \tag{14}$$

Proof. Since $r^*(U) = \mathbb{E}[\tilde{r} \mid U]$, we have

$$\mathbb{E}\left[\tilde{r} - r^*(U) \mid U\right] = 0. \tag{15}$$

Then

$$\begin{aligned} \mathcal{L}(\hat{r}) &= \mathbb{E}\left[\|\tilde{r} - r^*(U) + r^*(U) - \hat{r}(U)\|^2\right] \\ &= \mathbb{E}\left[\|\tilde{r} - r^*(U)\|^2\right] + \mathbb{E}\left[\|\hat{r}(U) - r^*(U)\|^2\right] \\ &\quad + 2\mathbb{E}\left[\langle \tilde{r} - r^*(U), r^*(U) - \hat{r}(U)\rangle\right]. \end{aligned} \tag{16}$$

The cross term is zero by the tower property of conditional expectation. Setting $\hat{r}(U) = 0$ and subtracting gives the improvement expression. The term $\mathbb{E}[\|r^*(U)\|^2]$ measures how much residual signal is contained in the exogenous information. If this term is zero, no method that only uses U can improve the backbone in expectation. If it is nonzero, ExoRAF improves the backbone whenever its retrieval error $\mathbb{E}[\|\hat{r}(U) - r^*(U)\|^2]$ is smaller than the explainable residual energy. □

Table 6: Dataset statistics. *Exo.* and *Endo.* denote exogenous and endogenous variables.

Dataset	Ride-Hailing	New York Taxi	Synthetic	Houston's Temperature	New York's Temperature
Domain	Traffic	Traffic	Synthetic waveform	Weather	Weather
Time-stamps	**14,025,600**	**4,704**	**26,280**	**45,216**	**45,216**
Frequency	30min	hourly	24 points per wave	hourly	hourly
Endo. Variable	9	1	1	1	1
Exo. Variable	9	3	2	6	6

A.2 Experimental Details

Datasets. We conduct experiments on five datasets with exogenous variables, including the Ride-Hailing dataset and four datasets from EventTSF (Ge et al., 2025)[3], i.e., New York Taxi, Synthetic, Houston's Temperature, New York's Temperature, as summarized in Table 6. Compared with existing datasets (Zeng et al., 2023), Ride-Hailing provides diverse endogenous and exogenous variables while better reflecting real-world scenarios. The exogenous variables consist of temporal indicators (i.e., half hour of day and day of week), area id, weather factor, and holiday-related factors, including public holidays, traditional festivals, western festivals, and two types of large-scale events.

Setup. Models are optimized using AdamW (Loshchilov & Hutter, 2019) under a unified hyperparameter search protocol. We split the four small-scale datasets into training, validation, and test sets with a ratio of 8:1:1, while the Ride-Hailing dataset is split with a ratio of 7:1.5:1.5. Following RideBench, we search the learning rate from $\{3\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}, 1\times10^{-3}, 3\times10^{-3}\}$ and report the performance under the best validation configuration. Models are trained for a maximum of 30 epochs, with early stopping if validation performance does not improve for 3 consecutive epochs.

To isolate the effect of our non-parametric design, we follow the training configurations of EventTSF and RideBench and keep the major architectural and optimization hyperparameters consistent across experiments. Specifically, the batch size is set to 64 for the four small-scale datasets and 128 for Ride-Hailing. If an out-of-memory (OOM) error occurs, we uniformly reduce the batch size to 32. We set the dropout rate to 0.5 and the hidden dimension to 256 across all experiments.

For pre-trained time-series foundation models, including Timer-S1, TimesFM 2.5, and Moirai 2.0, we use their officially released checkpoints and evaluate them without fine-tuning. These models are directly evaluated under a zero-shot inference setting to ensure a consistent evaluation protocol.

Evaluation Metrics. 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,

$$\text{WMAPE} = \frac{\sum_{i=1}^{n}|y_i-\tilde{y}_i|}{\sum_{i=1}^{n}|y_i|}, \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i-\tilde{y}_i|, \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}, \tag{17}$$

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

For the long-horizon forecasting setting on the Ride-Hailing dataset, pointwise errors alone are insufficient to assess whether a model captures the correct long-term trajectory. Therefore, we additionally evaluate day-level trend quality using Directional Accuracy of Day (Acc.) and Correlation of Day (Corr.). We first aggregate half-hourly predictions into daily values by averaging every 48 time steps. Let Y_i and $\tilde{Y}_i$ denote the ground-truth and predicted daily values on day i, respectively. Acc. is defined as

$$\text{Acc.} = \frac{1}{|\Omega|}\sum_{i\in\Omega}\mathbb{I}\left[\text{sign}(Y_i-Y_{i-1}) = \text{sign}(\tilde{Y}_i-\tilde{Y}_{i-1})\right], \tag{18}$$

where $\Omega = \{i : |Y_i - Y_{i-1}| > 10^{-6}\}$ denotes the set of days with non-negligible ground-truth changes. This metric measures whether the model predicts the correct direction of day-to-day de-

[3]https://github.com/WinfredGe/EventTSF

mand changes. Corr. is defined as

$$\text{Corr.} = \frac{\sum_{i=1}^{m}(Y_i - \mu)(\tilde{Y}_i - \tilde{\mu})}{\sqrt{\sum_{i=1}^{m}(Y_i - \mu)^2}\sqrt{\sum_{i=1}^{m}(\tilde{Y}_i - \tilde{\mu})^2}}, \tag{19}$$

where m is the number of evaluated days, and μ and $\tilde{\mu}$ are the means of the ground-truth and predicted daily sequences, respectively. Corr. measures whether the predicted daily trajectory follows the overall shape of the ground truth. Together, Acc. and Corr. provide a macro-level view of long-horizon forecasting quality beyond point-wise errors.

A.3 More Results and Analyses

Table 7: Long-horizon eight-week-ahead forecasting results of ExoRAF on the Ride-Hailing dataset. Average rank of exogenous-aware models across settings and metrics (lower is better).

Setting	Overall			Rank	Trend		Rank	First Week			Rank
Metrics	WMAPE	MAE	RMSE	Avg.	Acc.(↑)	Corr.(↑)	Avg.	WMAPE	MAE	RMSE	Avg.
CrossLinear	0.1293	4.941	14.66	7	0.6077	0.2671	9	0.1194	4.590	14.49	9
TimeXer	0.1319	5.048	14.63	12	0.6053	0.2673	10	0.1212	4.667	14.48	10
XLinear	0.1303	4.975	14.73	9	0.6072	0.2492	12	0.1211	4.658	14.53	11
TiDE	0.1330	5.070	14.56	11	0.6042	0.2970	11	0.1222	4.690	14.38	12
DAG	0.1335	5.099	14.63	13	0.5975	0.2428	13	0.1222	4.695	14.45	13
Crossformer	0.1348	5.187	15.21	–	0.6094	0.2476	–	0.1210	4.663	14.75	–
+ExoRAF	**0.1290**	**4.987**	**14.44**	5	**0.6474**	**0.4518**	1	**0.1137**	**4.385**	**13.87**	1
PatchTST	0.1299	4.957	14.58	–	0.6001	0.2457	–	0.1188	4.562	14.42	–
+ExoRAF	**0.1287**	**4.930**	**14.47**	4	**0.6330**	**0.3840**	4	**0.1158**	**4.453**	**14.05**	2
SegRNN	0.1273	4.862	14.66	–	0.6059	0.2470	–	0.1173	4.508	14.48	–
+ExoRAF	**0.1273**	**4.883**	**14.59**	2	**0.6268**	**0.3584**	7	**0.1140**	**4.388**	**14.19**	3
iTransformer	0.1302	4.972	14.70	–	0.6094	0.2527	–	0.1207	4.637	14.56	–
+ExoRAF	**0.1274**	**4.874**	**14.37**	1	**0.6346**	**0.3845**	3	**0.1171**	**4.499**	**14.19**	8
TQNet	0.1328	5.075	14.74	–	0.6019	0.2367	–	0.1213	4.659	14.53	–
+ExoRAF	**0.1287**	**4.942**	**14.39**	3	**0.6355**	**0.3677**	5	**0.1165**	**4.484**	**14.09**	5
PMDformer	0.1316	5.022	14.63	–	0.5959	0.2402	–	0.1202	4.613	14.40	–
+ExoRAF	**0.1298**	**4.977**	**14.39**	6	**0.6291**	**0.3797**	6	**0.1168**	**4.485**	**14.00**	4
DUET	0.1338	5.122	14.89	–	0.6106	0.2476	–	0.1235	4.756	14.69	–
+ExoRAF	**0.1297**	**4.997**	**14.46**	8	**0.6481**	**0.4447**	2	**0.1169**	**4.511**	**14.00**	7
SparseTSF	0.1338	5.106	14.92	–	0.5956	0.2121	–	0.1191	4.572	14.54	–
+ExoRAF	**0.1302**	**4.977**	**14.76**	10	**0.6226**	**0.3326**	8	**0.1163**	**4.465**	**14.34**	6
Avg. Improve	-2.20%	-1.80%	-2.06%	–	+5.14%	+60.73%	–	-3.60%	-3.50%	-3.12%	–

Long-Horizon Forecasting. Unlike week-ahead forecasting, the long-horizon setting evaluates models from three aspects: pointwise accuracy, day-level trend quality (Acc. and Corr.), and first-week forecasting performance, reflecting fine-grained numerical precision, long-term trajectory modeling, and near-term accuracy preservation, respectively. This setting extends forecasting from short-term fluctuation prediction to long-term temporal evolution modeling, evaluating whether models can capture macro-level demand trends across multiple weeks, which is essential for strategic planning and large-scale resource allocation.

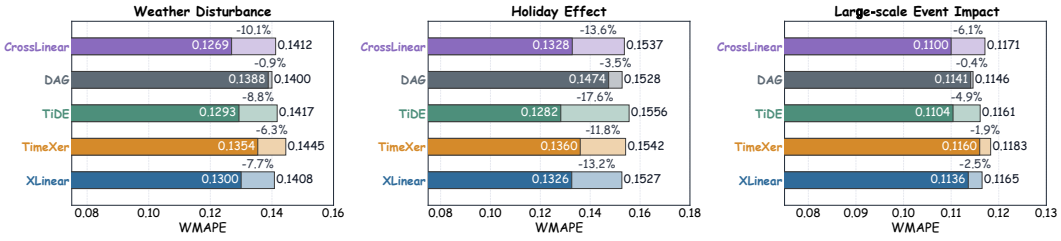

Figure 6: Effect of exogenous variables on exogenous-aware models on the Ride-Hailing dataset.

As shown in Table 7, ExoRAF consistently improves all three aspects. Notably, even a simple architecture such as SegRNN (Lin et al., 2025b) augmented with ExoRAF can outperform all exogenous-aware forecasting methods. No method achieves the best performance across all three objectives.

However, by integrating ExoRAF with PMDFormer (Hu et al., 2026), we achieve a favorable balance between trend accuracy and first-week forecasting performance.

Exogenous-aware forecasting models with ExoRAF. As shown in Figure 6, existing exogenous-aware forecasting methods, including CrossLinear (Zhou et al., 2025), DAG (Qiu et al., 2026), TiDE (Das et al., 2023a), TimeXer (Wang et al., 2024b), and XLinear (Chen et al., 2026), exhibit varying abilities in utilizing exogenous information. To answer the question of whether *ExoRAF can complement existing exogenous-aware forecasting methods?*

As shown in Table 8, ExoRAF improves forecasting performance across scenarios. The gains are more significant for models with limited exogenous modeling capabilities (i.e., TimeXer and DAG), while smaller improvements are observed for models with stronger exogenous modeling abilities (i.e., TiDE and CrossLinear). We think that this is due to the potential interference introduced by two-stage exogenous utilization: when the backbone already captures stable exogenous effects, ExoRAF may provide redundant signals that hinder residual refinement (Li et al., 2026).

Table 8: Results of exogenous-aware forecasting models with ExoRAF on the Ride-Hailing dataset.

Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact		
Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE
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
+ExoRAF	0.1050	4.024	12.50	**0.1266**	**5.280**	**18.86**	0.1330	5.502	17.91	0.1105	4.871	**15.73**
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**
+ExoRAF	**0.1047**	**4.013**	**12.65**	**0.1258**	**5.250**	18.82	**0.1318**	**5.460**	**18.23**	0.1104	4.869	15.95
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
+ExoRAF	**0.1068**	**4.104**	**12.65**	**0.1261**	**5.261**	**18.55**	0.1339	5.542	**18.36**	**0.1133**	**5.003**	**15.95**
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
+ExoRAF	**0.1064**	**4.085**	**12.65**	**0.1284**	**5.355**	**19.03**	**0.1346**	**5.580**	**18.33**	**0.1110**	**4.896**	**15.87**
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
+ExoRAF	**0.1054**	**4.040**	**12.80**	**0.1272**	**5.309**	**19.18**	**0.1358**	**5.624**	**18.76**	**0.1107**	**4.882**	**15.94**
Avg. Improve	-1.26%	-1.27%	-1.40%	-4.36%	-4.48%	-3.43%	-1.17%	-1.22%	-0.93%	-1.76%	-1.81%	-1.81%

ExoRAF vs. CrossLinear on different backbones. Similar to ExoRAF, CrossLinear (Zhou et al., 2025) is also designed as a plug-and-play method. We therefore integrate CrossLinear with existing forecasting models for further comparison. Different from the parallel integration strategy of ExoRAF, CrossLinear adopts a sequential architecture that incorporates exogenous information before the forecasting backbone. We evaluate ExoRAF and CrossLinear under two settings on the weather disturbance scenario. In the regular forecasting setting (see Figure 7(a)), each adaptation method is trained with its corresponding forecasting backbone. In the transfer setting (see Figure 7(b)), the ExoRAF module trained with PatchTST (Nie et al., 2023) and the independently trained CrossLinear module are transferred to target backbones without further fine-tuning; all target backbones use their independently trained and frozen parameters.

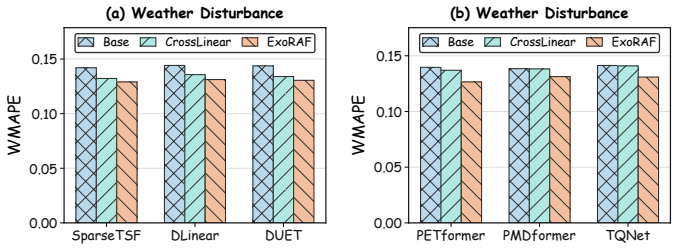

Figure 7: Comparison of the efficiency of ExoRAF and CrossLinear with different backbones.

Across both settings, ExoRAF consistently outperforms CrossLinear. We attribute this improvement to two factors. First, ExoRAF adopts a parallel design that preserves the original backbone representations, avoiding potential interference with the input features introduced by sequential integration. Second, CrossLinear primarily relies on directly fusing endogenous and exogenous variables, without explicitly modeling the dynamic effects induced by exogenous factors. In contrast, ExoRAF captures exogenous-driven residual variations and adaptively retrieves transferable residual patterns, enabling consistent improvements across diverse forecasting models regardless of the underlying architecture. Consequently, ExoRAF maintains stable performance gains even under the transfer setting, demonstrating its robustness and generalizability to unseen forecasting backbones.

A.4 More Application Analysis

Memory Slots. Figure 8(a) illustrates the impact of memory capacity with different numbers of memory slots (i.e., 10, 20, 30, and 50). Increasing memory capacity does not always improve performance. A larger memory may introduce redundant or less relevant residual prototypes, which weakens retrieval attention and consequently degrades the quality of retrieved corrections. Moreover, an excessively large memory increases storage and retrieval overhead, making it less suitable for deployment.

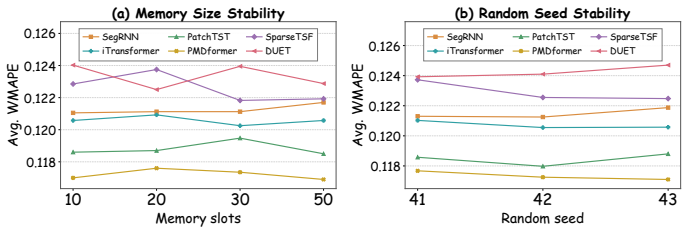

Figure 8: Sensitivity analysis of ExoRAF: (a) the impact of different parameter configurations and (b) the impact of different random seeds.

Random Seed. We further evaluate the robustness of ExoRAF under different random seeds. Figure 8(b) reports the forecasting performance with different random seeds (41, 42, and 43) under the regular forecasting setting. ExoRAF maintains stable performance with low variance across different seeds, demonstrating its robustness to random initialization.

Table 9: Regular forecasting results when future exogenous variables are unavailable (i.e., only historical *exo.* is available). Here, *exo.* denotes exogenous variables. ExoRAF consistently improves the forecasting performance of the underlying backbones under this setting.

Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact		
Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE
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
only *historical exo.*	0.1093	4.192	12.92	0.1306	5.453	19.18	0.1400	5.797	18.97	0.1133	4.996	15.93
historical + future exo.	**0.1084**	**4.158**	**12.87**	**0.1290**	**5.384**	**18.98**	**0.1361**	**5.637**	**18.80**	**0.1131**	**4.993**	**15.97**
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
only *historical exo.*	0.1051	4.032	12.81	0.1276	5.328	19.27	0.1368	5.673	19.06	0.1103	4.867	15.89
historical + future exo.	**0.1038**	**3.981**	**12.64**	**0.1247**	**5.204**	**18.72**	**0.1346**	**5.584**	**18.75**	**0.1094**	**4.830**	**15.78**
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
only *historical exo.*	0.1070	4.101	12.94	0.1294	5.403	19.36	0.1374	5.695	19.05	0.1123	4.954	16.11
historical + future exo.	**0.1060**	**4.064**	**12.78**	**0.1274**	**5.320**	**19.04**	**0.1344**	**5.569**	**18.61**	**0.1111**	**4.901**	**15.94**
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
only *historical exo.*	0.1112	4.275	13.11	0.1330	5.554	19.37	0.1434	5.955	19.26	0.1158	5.116	16.30
historical + future exo.	**0.1095**	**4.207**	**12.82**	**0.1305**	**5.445**	**18.84**	**0.1395**	**5.787**	**18.67**	**0.1149**	**5.072**	**16.11**

Table 10: Long-horizon forecasting results when future exogenous variables are unavailable (i.e., only historical exo. is available). Here, *exo.* denotes exogenous variables.

Setting	Overall			Trend		First Week		
Metrics	WMAPE	MAE	RMSE	Acc.(↑)	Corr.(↑)	WMAPE	MAE	RMSE
SparseTSF	0.1338	5.106	14.92	0.5956	0.2121	0.1191	4.572	14.54
only *historical exo.*	0.1308	4.998	14.94	0.6120	0.2693	0.1174	4.505	14.52
historical + future exo.	**0.1302**	**4.977**	**14.76**	**0.6226**	**0.3326**	**0.1163**	**4.465**	**14.34**
PatchTST	0.1299	4.957	14.58	0.6001	0.2457	0.1188	4.562	14.42
only *historical exo.*	**0.1282**	**4.910**	**14.40**	0.6314	**0.3860**	**0.1149**	**4.419**	**13.95**
historical + future exo.	0.1287	4.930	14.47	**0.6330**	0.3840	0.1158	4.453	14.05
iTransformer	0.1302	4.972	14.70	0.6094	0.2527	0.1207	4.637	14.56
only *historical exo.*	**0.1268**	**4.851**	**14.30**	**0.6379**	**0.3986**	**0.1167**	**4.488**	**14.10**
historical + future exo.	0.1274	4.874	14.37	0.6346	0.3845	0.1171	4.499	14.19
DUET	0.1338	5.122	14.89	0.6106	0.2476	0.1235	4.756	14.69
only *historical exo.*	**0.1287**	**4.957**	**14.38**	0.6441	0.4153	**0.1165**	**4.493**	**13.95**
historical + future exo.	0.1297	4.997	14.46	**0.6481**	**0.4447**	0.1169	4.511	14.00

Forecasting without Future Exogenous Variables. In real-world applications, future exogenous variables are often unavailable or difficult to obtain in a timely manner. To evaluate the robustness of ExoRAF under this practical setting, we conduct experiments on the Ride-Hailing dataset for both regular and long-horizon forecasting using only historical endogenous and exogenous variables. When future exogenous variables are available, Table 9 shows that incorporating them consistently

improves forecasting accuracy in the standard setting, demonstrating ExoRAF's ability to effectively leverage future exogenous information.

However, Table 10 shows that future exogenous variables provide limited benefits in the long-horizon forecasting setting. As the forecasting horizon increases, long-term temporal structures, such as trends and periodic patterns, become increasingly dominant, while short-term dependencies between future exogenous variables and endogenous dynamics become weaker. Moreover, future exogenous variables over extended horizons are often associated with higher uncertainty, which may introduce noise and reduce the effectiveness of residual correction. Therefore, historical exogenous contexts are generally sufficient for ExoRAF to provide effective residual retrieval, particularly in long-horizon forecasting scenarios.

A.5 More Visualizations

Figures 9 and 10 present additional comparisons between the ground-truth residuals and the residuals generated by ExoRAF under similar exogenous contexts. As shown in Figure 11, residual features from different scenarios form clearly separated clusters, indicating that forecasting residuals contain structured patterns associated with exogenous contexts rather than random noise. It further validates the effectiveness of leveraging exogenous contexts to retrieve historical residual patterns for adaptive forecasting correction.

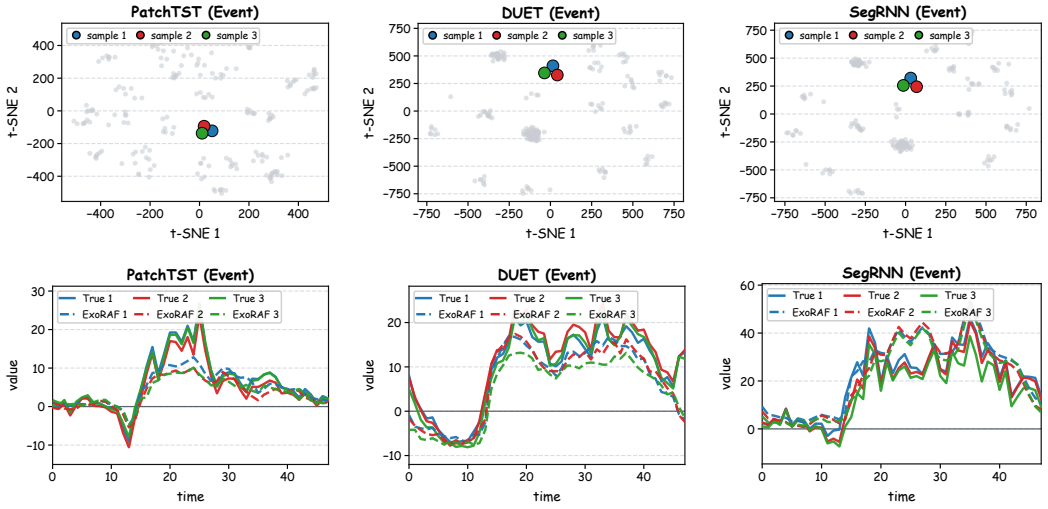

Figure 9: Visualization of residuals under large-scale event impact scenarios.

A.6 Full Results

Table 11 summarizes the performance of other baselines, including PETformer (Lin et al., 2024c), TQNet (Lin et al., 2025a), Crossformer (Zhang & Yan, 2023), SegRNN (Lin et al., 2025b), and DLinear (Zeng et al., 2023), on the Ride-Hailing dataset under the regular forecasting setting. Table 12 summarizes the results of other baselines, including Crossformer (Zhang & Yan, 2023), DUET (Qiu et al., 2025), and DLinear (Zeng et al., 2023), under the zero-shot area forecasting setting. Table 13 shows the full results of transfer evaluation.

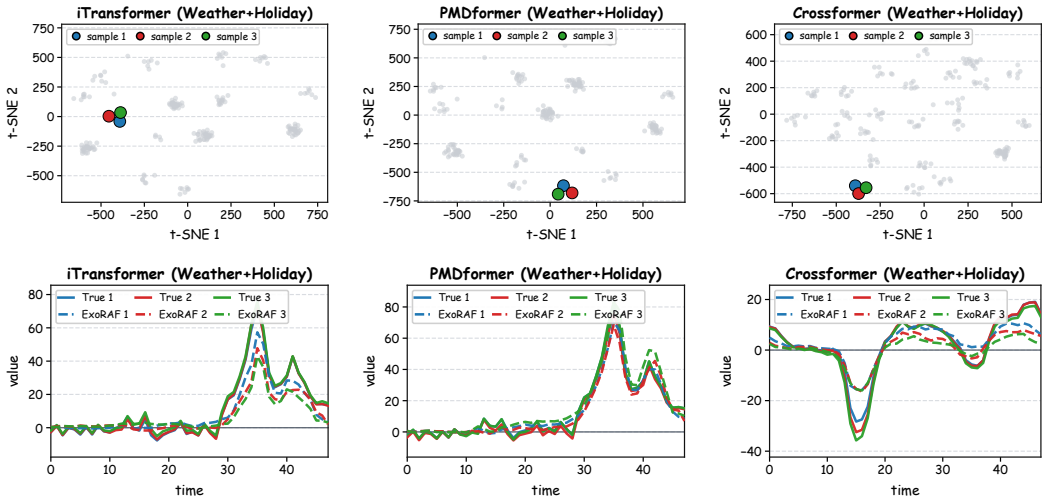

Figure 10: Visualization of residuals under weather-holiday interaction scenarios.

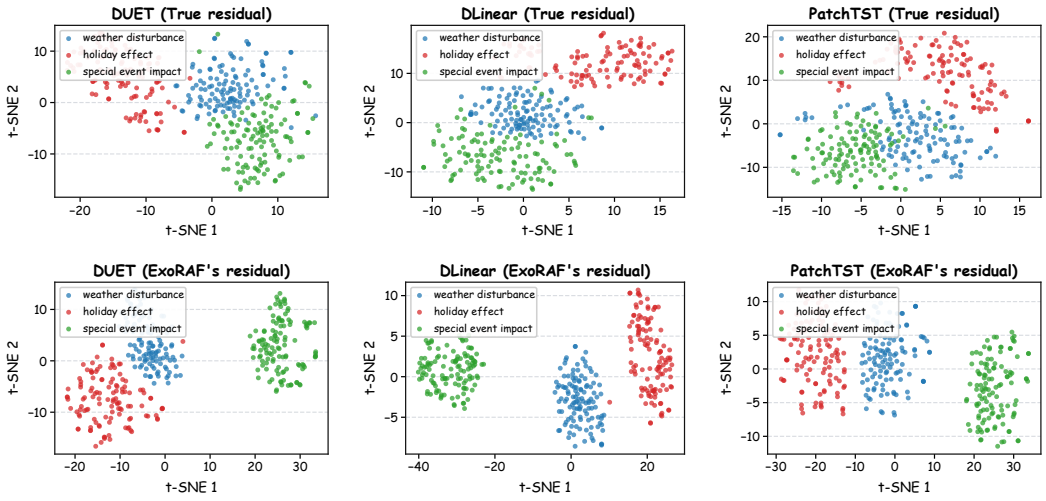

Figure 11: t-SNE visualization of residuals under three scenarios.

Table 11: Full results of ExoRAF under the regular setting on the large-scale Ride-Hailing dataset.

Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact			Rank
Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	Avg.
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	3
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	4
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	9
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	14
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	15
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	–
+ExoRAF	**0.1027**	**3.939**	**12.52**	**0.1238**	**5.168**	**18.62**	**0.1327**	**5.505**	**18.40**	**0.1085**	**4.787**	**15.60**	1
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	–
+ExoRAF	**0.1038**	**3.981**	**12.64**	**0.1247**	**5.204**	**18.72**	**0.1346**	**5.584**	**18.75**	**0.1094**	**4.830**	**15.78**	2
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	–
+ExoRAF	**0.1031**	**3.953**	**12.73**	**0.1262**	**5.268**	**19.13**	**0.1364**	**5.660**	**19.02**	**0.1087**	**4.791**	**15.81**	5
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	–
+ExoRAF	**0.1059**	**4.060**	**12.74**	**0.1272**	**5.309**	**18.96**	**0.1335**	**5.530**	**18.50**	**0.1110**	**4.896**	**15.91**	6
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	–
+ExoRAF	**0.1060**	**4.064**	**12.78**	**0.1274**	**5.320**	**19.04**	**0.1344**	**5.569**	**18.61**	**0.1111**	**4.901**	**15.94**	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	–
+ExoRAF	**0.1049**	**4.027**	**12.80**	**0.1264**	**5.277**	**19.11**	**0.1368**	**5.673**	**18.94**	**0.1112**	**4.914**	**16.03**	8
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	–
+ExoRAF	**0.1084**	**4.158**	**12.87**	**0.1290**	**5.384**	**18.98**	**0.1361**	**5.637**	**18.80**	**0.1131**	**4.993**	**15.97**	10
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	–
+ExoRAF	**0.1064**	**4.082**	**12.92**	**0.1279**	**5.343**	**19.05**	**0.1342**	**5.565**	**18.87**	**0.1146**	**5.059**	**16.38**	11
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	–
+ExoRAF	**0.1120**	**4.291**	**12.88**	**0.1312**	**5.476**	**18.96**	**0.1398**	**5.788**	**18.64**	**0.1136**	**5.013**	**15.67**	12
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	–
+ExoRAF	**0.1095**	**4.207**	**12.82**	**0.1305**	**5.445**	**18.84**	**0.1395**	**5.787**	**18.67**	**0.1149**	**5.072**	**16.11**	13
Avg. Improve	-4.53%	-4.37%	-5.12%	-9.63%	-9.83%	-9.13%	-11.07%	-11.07%	-7.64%	-3.84%	-3.83%	-4.33%	–

Table 12: Full zero-shot area forecasting results on the Ride-Hailing dataset under the regular setting.

Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact			Rank
Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	Avg.
CrossLinear	0.1054	4.124	11.71	0.1270	5.437	17.24	0.1338	5.654	18.05	0.1127	5.087	16.90	3
DAG	0.1080	4.220	12.32	0.1391	5.966	19.18	0.1462	6.175	19.53	0.1138	5.133	17.26	12
XLinear	0.1083	4.232	12.14	0.1339	5.737	18.49	0.1429	6.042	18.87	0.1134	5.111	16.96	10
TimeXer	0.1089	4.275	12.01	0.1387	5.948	18.65	0.1388	5.869	18.49	0.1144	5.161	17.05	11
TiDE	0.1095	4.273	12.42	0.1413	6.064	19.31	0.1511	6.382	19.86	0.1137	5.123	17.15	13
PMDformer	0.1072	4.186	12.40	0.1386	5.944	19.21	0.1498	6.335	20.00	0.1103	4.968	17.14	–
+ExoRAF	**0.1032**	**4.035**	**11.83**	**0.1251**	**5.356**	**17.41**	**0.1373**	**5.807**	**18.76**	**0.1098**	**4.951**	**16.86**	1
PETformer	0.1070	4.177	12.39	0.1398	6.001	19.27	0.1527	6.460	20.23	0.1107	4.982	17.16	–
+ExoRAF	**0.1024**	**4.005**	**11.78**	**0.1264**	**5.412**	**17.64**	**0.1372**	**5.801**	**18.87**	**0.1077**	**4.850**	**16.50**	2
PatchTST	0.1072	4.186	12.42	0.1394	5.983	19.29	0.1513	6.399	20.15	0.1108	4.993	17.16	–
+ExoRAF	**0.1033**	**4.041**	**11.89**	**0.1256**	**5.380**	**17.53**	**0.1380**	**5.841**	**19.07**	**0.1103**	**4.973**	**16.93**	4
iTransformer	0.1113	4.348	12.51	0.1413	6.064	19.29	0.1544	6.528	20.13	0.1139	5.134	17.24	–
+ExoRAF	**0.1062**	**4.154**	**11.87**	**0.1281**	**5.483**	**17.48**	**0.1386**	**5.856**	**18.65**	**0.1120**	**5.054**	**16.91**	5
Crossformer	0.1094	4.278	12.56	0.1407	6.041	19.33	0.1538	6.502	20.21	0.1139	5.141	17.36	–
+ExoRAF	**0.1054**	**4.127**	**12.03**	**0.1271**	**5.439**	**17.67**	**0.1408**	**5.955**	**19.15**	**0.1108**	**4.994**	**16.93**	6
DUET	0.1116	4.364	12.56	0.1422	6.104	19.42	0.1540	6.515	20.17	0.1144	5.162	17.35	–
+ExoRAF	**0.1083**	**4.243**	**11.87**	**0.1296**	**5.544**	**17.46**	**0.1379**	**5.826**	**18.56**	**0.1141**	**5.144**	**16.87**	7
SparseTSF	0.1123	4.391	12.65	0.1419	6.095	19.45	0.1565	6.624	20.40	0.1166	5.264	17.22	–
+ExoRAF	**0.1077**	**4.215**	**12.00**	**0.1292**	**5.531**	**17.52**	**0.1418**	**5.996**	**19.09**	**0.1118**	**5.039**	**16.72**	8
DLinear	0.1156	4.512	12.52	0.1432	6.142	19.09	0.1549	6.547	19.63	0.1172	5.286	16.70	–
+ExoRAF	**0.1112**	**4.349**	**12.05**	**0.1321**	**5.655**	**17.76**	**0.1415**	**5.976**	**18.72**	**0.1144**	**5.162**	**16.41**	9
Avg. Improve	-3.85%	-3.70%	-4.69%	-9.22%	-9.46%	-8.99%	-9.31%	-9.34%	-6.24%	-1.85%	-1.85%	-2.33%	–

Table 13: Cross-backbone transfer performance for regular forecasting on the Ride-Hailing dataset.

	Scenarios	Overall			Weather Disturbance			Holiday Effect			Large-scale Event Impact		
	Metrics	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE	WMAPE	MAE	RMSE
PatchTST	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
	+ExoRAF	**0.1052**	**4.041**	**12.53**	**0.1265**	**5.278**	**18.62**	**0.1348**	**5.591**	**18.56**	**0.1103**	**4.861**	**15.58**
	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
	+ExoRAF	**0.1112**	**4.268**	**12.86**	**0.1309**	**5.467**	**18.83**	**0.1374**	**5.695**	**18.66**	**0.1175**	**5.189**	**16.26**
	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
	+ExoRAF	0.1176	4.528	**12.74**	**0.1368**	**5.714**	**18.56**	**0.1425**	**5.910**	**18.38**	0.1181	5.218	**15.65**
	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
	+ExoRAF	0.1227	4.738	**13.01**	**0.1410**	**5.889**	**18.79**	**0.1508**	**6.257**	**18.75**	0.1218	5.385	**15.93**
iTransformer	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
	+ExoRAF	**0.1097**	**4.219**	**12.96**	**0.1309**	**5.468**	**19.28**	**0.1389**	**5.762**	**18.87**	**0.1157**	**5.115**	**16.20**
	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
	+ExoRAF	0.1106	4.277	**12.87**	**0.1313**	**5.484**	**19.12**	**0.1397**	**5.806**	**18.86**	0.1160	5.132	**15.92**
	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
	+ExoRAF	0.1127	4.330	**12.65**	**0.1324**	**5.531**	**18.65**	**0.1392**	**5.766**	**18.27**	0.1153	5.091	**15.59**
	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
	+ExoRAF	0.1267	4.882	**13.27**	**0.1443**	**6.022**	**19.09**	**0.1533**	**6.354**	**18.83**	**0.1253**	**5.536**	**16.19**
Avg. Improve		+1.92%	+2.41%	-4.68%	-5.25%	-5.41%	-9.54%	-7.49%	-7.48%	-8.18%	+0.34%	+0.43%	-4.60%