ON THE ROLE OF CONTEXT IN ZERO-SHOT MULTIVARIATE TIME-SERIES FORECASTING

Rohith Saai Pemmasani Prabakaran
University of Amsterdam
Amsterdam, The Netherlands
{rohith,saai,pemmasani,prabakaran}@uva.nl

Stijn Verdenius
WAIR
Amsterdam, The Netherlands
{stijn}@wairforretail.com

Partha Das
WAIR
Amsterdam, The Netherlands
{partha,das}@wairforretail.com

Christian A. Naesseth
University of Amsterdam
Amsterdam, The Netherlands
{c,a,naesseth}@uva.nl

ABSTRACT

Accurate multivariate time-series forecasting is essential for decision making in domains such as retail, yet most neural forecasting models remain task-specific and perform poorly in data-scarce or zero-shot settings. Prior-Fitted Networks (PFNs) provide a principled Bayesian meta-learning framework, but existing PFN-based approaches to time-series forecasting make limited use of context, restricting their ability to condition on structurally related examples. We propose a *context-centric framework* for zero-shot multivariate forecasting and introduce In-Context TimePFN, a Transformer-based PFN trained exclusively on synthetic tasks with explicit *context–query* structure. By treating forecasting as in-context Bayesian inference over structured temporal contexts, our approach performs probabilistic prediction without task-specific fine-tuning. Experiments on four real-world benchmarks show that In-Context TimePFN achieves competitive or superior zero-shot performance compared to standard baselines and TimePFN. Controlled ablations further demonstrate that performance is highly sensitive to the quality and alignment of context, identifying structured context as a key enabler of effective zero-shot time-series forecasting with PFNs.

Track: Research

1 INTRODUCTION

Multivariate time-series (MTS) forecasting is a critical component of decision-making in fast-paced industries such as retail (Wanchoo, 2019; Fildes et al., 2022), yet most neural network-based forecasting models remain task-specific and degrade under distribution shift or data scarcity. This limitation is exacerbated in large architectures such as Transformers, which offer limited support for zero-shot generalization across datasets.

Prior-Fitted Networks (PFNs) (Müller et al., 2021) provide a meta-learning alternative by framing prediction as task-level inference, where forecasts are produced by conditioning on a set of related

context examples drawn from the same task distribution. For example, if a retailer introduces a new blue sweater in a style previously sold in red, a PFN can treat the historical sales, pricing, and weather-driven demand of the red sweater as context, and use this to predict future sales of the blue variant (the query) without retraining. However, existing PFN-based forecasting models depart from this formulation in a crucial way: models such as ForecastPFN (Dooley et al., 2024) and TimePFN (Taga et al., 2025) treat the query history itself as context, effectively reducing inference to sequence extrapolation from a single trajectory rather than conditioning on related context instances. Figure 1 illustrates this distinction: existing PFN-based forecasting models treat the query history as the sole context, whereas our approach explicitly separates context examples from the query series, enabling task-level in-context inference.

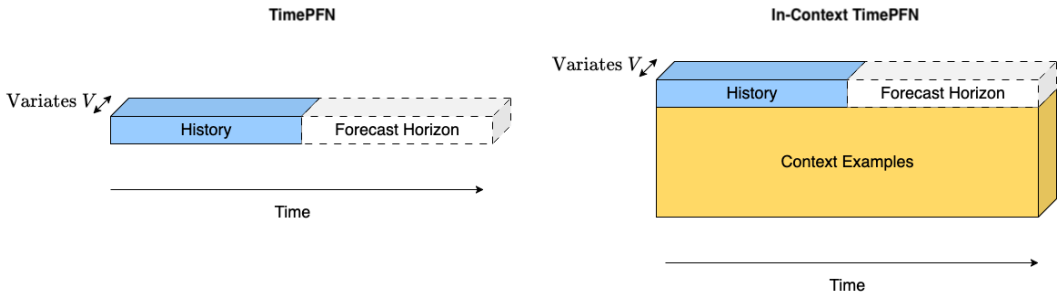


Figure 1: Comparison of context usage in PFN-based time-series forecasting. Left: TimePFN treats the query history as the sole context, reducing inference to history-based sequence extrapolation. Right: Our approach separates the query history from an explicit set of context examples drawn from the same task distribution, enabling task-level in-context inference.

In this paper, we argue that explicit task-level context is not interchangeable with query history, and that this distinction fundamentally shapes zero-shot forecasting behavior. To study this effect, we introduce an in-context PFN framework that operates on sets of related time-series instances, explicitly separating context examples from the query series. The model is trained exclusively on synthetic multivariate forecasting tasks designed to enforce shared structure across context–query pairs, enabling genuine in-context inference rather than history-only extrapolation.

Using this framework, we analyze zero-shot forecasting performance across three real-world benchmarks (Zhou et al., 2021) and one retail-related benchmark that permits explicit context definition. We find that structurally aligned context substantially improves performance on datasets such as ECL and Traffic, while random or misaligned context can actively harm inference, often underperforming a with-context baseline. Conversely, when curating structurally meaningful context requires domain knowledge or specialized retrieval techniques, as in ETT and Walmart M5, history-based models such as TimePFN can outperform context-based inference by avoiding reliance on poorly aligned context. These results clarify the functional difference between sequence-level extrapolation and task-level in-context inference in time-series forecasting. Rather than proposing a new forecasting architecture, this work isolates how explicit context changes the inference regime of PFN-based models and identifies the difficulty of constructing structurally relevant context sets in practice as a key bottleneck for effective zero-shot generalization.

2 Related work

Recent work on zero-shot time-series forecasting has increasingly adopted in-context learning paradigms trained on large-scale synthetic data. ForecastPFN (Dooley et al., 2024) introduced PFNs for univariate forecasting, demonstrating zero-shot generalization from synthetic training alone. However, the observed history of the query series is treated as the sole context, limiting conditioning on structurally related examples. TimePFN (Taga et al., 2025) extends this approach to multivariate settings using variate-wise tokenization and kernel-based synthetic priors, but similarly conflates query history with context, reducing inference to history-based sequence extrapolation.

This departs from the original PFN formulation, which casts prediction as task-level inference conditioned on an explicit context set. In the PFN framework, a forecasting task is defined as

$(\mathbb{D}_{\text{context}}, \boldsymbol{x}_{\text{query}})$, where $\mathbb{D}_{\text{context}}$ is a set of related time-series instances and $\boldsymbol{x}_{\text{query}}$ is the observed history of a target series. The model aims to approximate the posterior predictive distribution over the future trajectory $\boldsymbol{y}_{\text{query}}$,

$$P(\boldsymbol{y}_{\text{query}} \mid \boldsymbol{x}_{\text{query}}, \mathbb{D}_{\text{context}}), \tag{1}$$

which collapses when the query history itself is treated as context.

LatPFN (Verdenius et al., 2024) moves closer to this inference paradigm by conditioning on multiple related time series sampled from a shared distribution, highlighting the importance of explicit context examples. However, its synthetic prior remains univariate and does not capture multivariate dependency structure.

More broadly, existing synthetic priors based on kernel composition (Ansari et al., 2024; Taga et al., 2025) or trend–seasonality decomposition (Dooley et al., 2024; Verdenius et al., 2024) capture marginal temporal behaviour but offer limited control over cross-variable dependencies, making it difficult to study the role of context alignment in multivariate zero-shot forecasting.

3 METHODOLOGY

We study zero-shot multivariate time-series forecasting under an explicit context–query task formulation. Our framework comprises two components: 1. a synthetic task generator that enforces shared structure between context and query instances, and 2. an in-context TimePFN trained on these tasks to perform zero-shot forecasting via task-level inference.

3.1 SYNTHETIC TASK GENERATION

To isolate the effect of context alignment, we train exclusively on synthetic multivariate forecasting tasks in which context and query instances are structurally related by construction. Each task is generated from a shared latent causal mechanism that governs both the context set and the query series, ensuring that context examples provide meaningful task-specific information.

Concretely, for each task we sample a stationary causal graph over observed variables and latent factors, which defines the temporal dynamics and cross-variable dependencies shared across all task instances. Context and query series are generated from the same underlying mechanism, while independently sampled parameters and noise introduce controlled variability across instances. This yields task families that are internally coherent but non-identical.

By enforcing shared structure between context and query, this synthetic setup enables genuine in-context inference and allows us to study the role of context alignment independently of real-world data availability or context retrieval heuristics. More details are provided in Appendix A.1

3.2 IN-CONTEXT TIMEPFN

We train a Transformer-based Prior-Fitted Network (Figure 2) to perform zero-shot multivariate forecasting via in-context inference on the synthetic tasks described above. The model takes as input a set of context time-series instances and a query series with an observed history, and outputs a predictive distribution over the future trajectory of the query.

The architecture is designed for task-level inference rather than single-sequence extrapolation. A context encoder aggregates information across the context set to infer a latent representation of shared task structure, capturing common temporal patterns and cross-variable dependencies. A query decoder then conditions on this representation while processing the query history, allowing predictions to adapt to task-specific dynamics without parameter updates at inference time.

As evidenced by Liu et al. (2023), the model operates on variate-level representations to preserve multivariate structure and produces probabilistic forecasts, enabling uncertainty estimation consistent with posterior predictive inference conditioned on the context set.

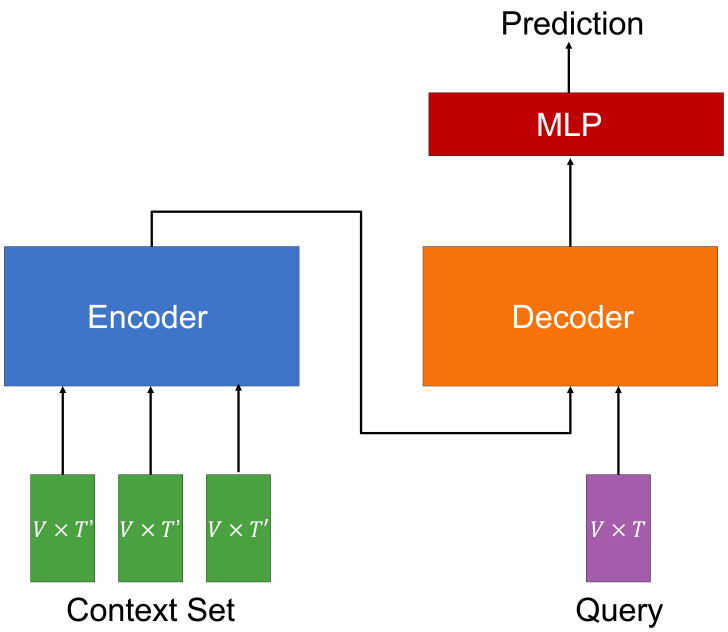


Figure 2: Model architecture for zero-shot forecasting using the In-Context TimePFN framework. The context set consists of C multivariate time-series examples of dimension $V \times T'$, each embedded and passed through the Transformer encoder. The query instance, a single $V \times T$ history window, is embedded and processed by the Transformer decoder using cross-attention over the encoded context. The decoder outputs a representation of shape $V \times d_{embed}$, which is passed to the forecasting MLP head, producing predictions for each query.

4 EXPERIMENTS

We evaluate our approach in a strict zero-shot setting by training exclusively on synthetic tasks and testing directly on real-world multivariate datasets without any task-specific fine-tuning. The In-Context TimePFN is pretrained on one million synthetic tasks, each consisting of a query series and a context set of seven structurally related examples.

Experiments are conducted on three standard multivariate benchmarks (ETT, ECL, and Traffic) (Zhou et al., 2021) and the Walmart M5 dataset (Howard et al., 2020) to reflect a retail forecasting setting. All benchmark datasets are normalized and evaluated using a rolling-window setup with a history length and forecast horizon of 96 time steps. For each query instance, context examples are constructed from temporally or structurally related series (e.g., parallel entities or matching calendar periods), with no temporal overlap with the target. For ETT, ECL and Traffic, for every query, context examples are constructed from the same dataset one calender year ahead. For Walmart M5, for every query i.e product sold at a store, context examples are constructed from the time-series of the same product sold at other stores during the same time period. Performance is compared against standard baselines and TimePFN using MSE and MAE. Other probabilistic metrics are not considered as the other baselines provide point-estimates, making comparison difficult.

4.1 ZERO-SHOT FORECASTING ON REAL-WORLD DATASETS

Table 1 reports zero-shot forecasting performance across all datasets. Qualitative results are provided in Appendix A.2. In-Context TimePFN achieves the strongest results on ECL and Traffic, where our context definition is aligned with the query series.. On ECL, it attains the lowest error among all methods, outperforming all baselines. A similar pattern is observed on Traffic, with substantial gains over all baselines.

Table 1: Mean (standard deviation) for MSE and MAE across the models and benchmark datasets. Best results per dataset and metric are in red color and bold.

Dataset	**ETT**		**ECL**		**Traffic**		**Walmart**	
Model	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
Mean	0.12 (0.11)	0.24 (0.09)	0.81 (0.58)	0.72 (0.22)	1.32 (0.36)	0.99 (0.11)	**0.96** (0.20)	**0.69** (0.19)
Naive	0.16 (0.12)	0.28 (0.10)	1.37 (1.35)	0.90 (0.42)	2.57 (1.82)	1.29 (0.35)	1.89 (3.52)	0.81 (0.68)
S.Naive	0.17 (0.12)	0.28 (0.09)	1.40 (1.14)	0.92 (0.37)	2.60 (1.15)	1.31 (0.25)	2.12 (1.64)	0.89 (0.42)
TimePFN	**0.06** (0.06)	**0.19** (0.09)	0.94 (0.65)	0.77 (0.25)	1.42 (0.38)	1.02 (0.12)	2.41 (21.60)	0.87 (0.58)
Ours	0.35 (0.27)	0.48 (0.20)	**0.79** (0.44)	**0.70** (0.20)	**1.23** (0.41)	**0.91** (0.13)	2.52 (21.84)	0.89 (0.59)

In contrast, In-Context TimePFN underperforms TimePFN on ETT and Walmart. On these datasets, our definition of context does not accurately reflect the structure of the query series, and conditioning on such context does not yield consistent improvements. In these cases, history-based models such as TimePFN benefit from avoiding reliance on potentially misaligned context.

These results suggest that zero-shot performance differences across datasets are largely driven by context quality rather than model capacity, a hypothesis we examine directly through controlled context ablations.

4.2 Ablation on the importance of the context set

We analyse the role of context quality by comparing two variants: *With Context*, which uses query-aligned examples, and *Random Context*, which provides unaligned examples. Results are summarized in Table 2.

Aligned context consistently improves performance on ECL and Traffic, where the *With Context* variant achieves the lowest error across both metrics. Random context often degrades performance relative to with context, indicating that the model actively conditions on provided examples even when they are uninformative. On ETT and Walmart, performance differences across variants are small, confirming that context curation is a limiting factor in these settings.

Table 2: Evaluating the influence of context quality on PFN performance. Mean values are reported with standard deviations in parentheses. Lowest values per dataset and metric are highlighted in blue.

Setting	**With Context**		**Random Context**	
Dataset	**MSE**	**MAE**	**MSE**	**MAE**
ETT	0.35 (0.27)	0.48 (0.20)	**0.34** (0.31)	**0.45** (0.22)
ECL	**0.79** (0.44)	**0.70** (0.20)	1.06 (0.61)	0.83 (0.23)
Traffic	**1.23** (0.41)	**0.91** (0.13)	1.78 (0.73)	1.09 (0.21)
Walmart	2.52 (21.84)	**0.89** (0.59)	**2.52** (21.59)	**0.89** (0.59)

5 Conclusion

We showed that zero-shot multivariate forecasting with Prior-Fitted Networks is highly sensitive to context quality. While existing PFN-based time-series models collapse context into the query history, our analysis demonstrates that explicit context can substantially improve performance when it is structurally aligned with the target task. Conversely, misaligned or uninformative context can be actively harmful, often underperforming a with context baseline.

These findings explain why context-based PFN models succeed on some datasets but not others, and highlight context alignment, rather than model capacity, as a primary bottleneck for effective in-context forecasting. Future work should therefore focus on principled, structure-aware context selection mechanisms that enable reliable task-level inference in real-world settings.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Samuel Dooley, Gurnoor Singh Khurana, Chirag Mohapatra, Siddartha V Naidu, and Colin White. Forecastpfn: Synthetically-trained zero-shot forecasting. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Addison Howard, inversion, Spyros Makridakis, and vangelis. M5 forecasting - accuracy. `https://kaggle.com/competitions/m5-forecasting-accuracy`, 2020. Kaggle.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Timepfn: Effective multivariate time series forecasting with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20761–20769, 2025.

Stijn Verdenius, Andrea Zerio, and Roy LM Wang. Lat-pfn: A joint embedding predictive architecture for in-context time-series forecasting. *arXiv preprint arXiv:2405.10093*, 2024.

Karan Wanchoo. Retail demand forecasting: a comparison between deep neural network and gradient boosting method for univariate time series. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp. 1–5. IEEE, 2019.

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.

A APPENDIX

A.1 SYNTHETIC TASK GENERATION

We generate synthetic multivariate forecasting tasks using a structural causal model (SCM)-based prior designed to produce coherent context-query pairs that share an underlying generative mechanism.

Each task is defined by a stationary directed acyclic graph (DAG) over V observed variables $\{X_1, \ldots, X_V\}$ and V latent confounders $\{U_1, \ldots, U_V\}$. For every observed variable X_i^t, temporal structure is introduced via (i) an autoregressive self-edge from X_i^{t-1}, and (ii) lagged edges from latent parents $U_j^{t-\ell}$, where lags ℓ are sampled uniformly from a fixed range.

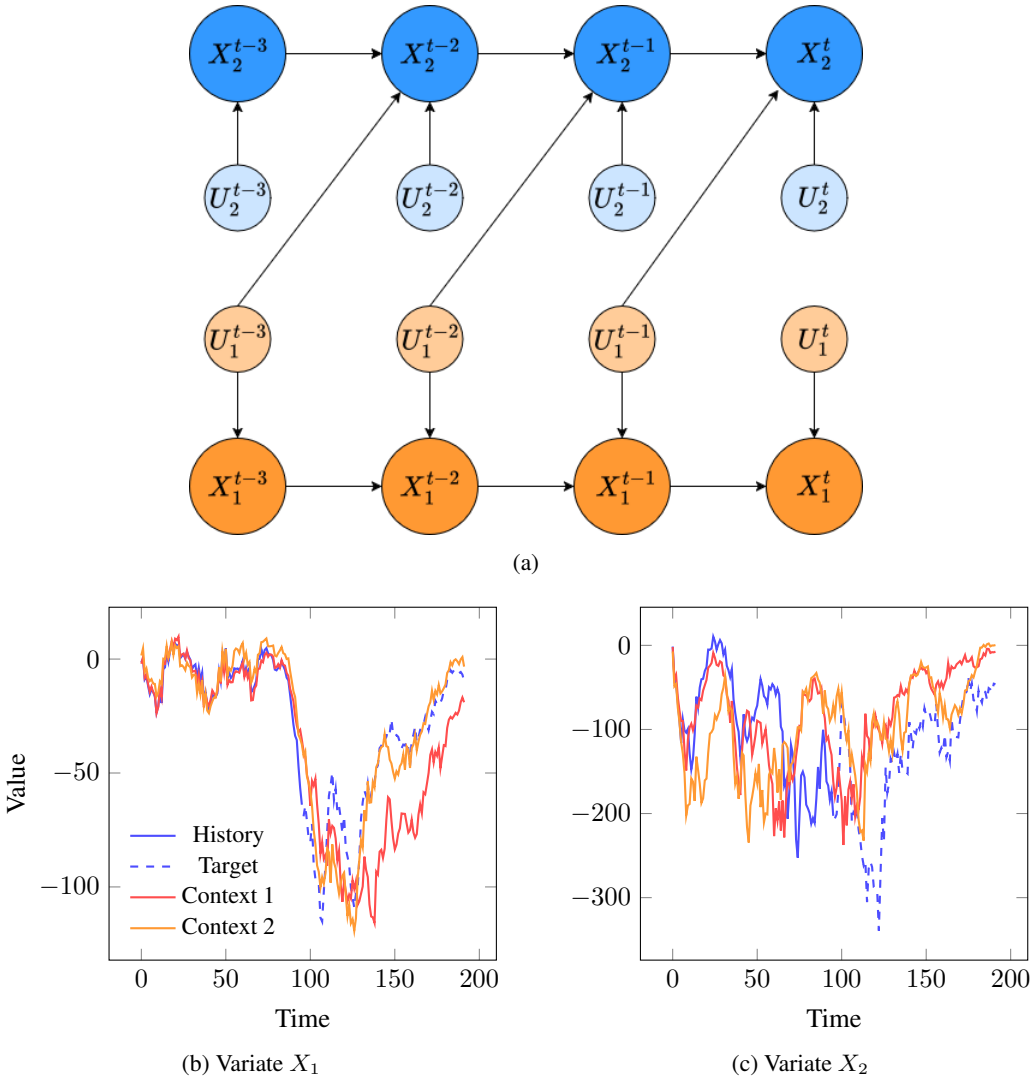
Given a sampled graph, each observed variable evolves according to:

$$X_i^t = w_i X_i^{t-1} + \sum_{(j,\ell)\in \mathrm{Pa}(X_i)} w_{ij}^{(\ell)} U_j^{t-\ell} + \varepsilon_i^t,$$

where w_i are autoregressive coefficients, $w_{ij}^{(\ell)}$ are latent influence weights, and ε_i^t is additive noise (Gaussian, Student-t, or Laplace). Structural weights are sampled independently for each context and query instance within a task, while the causal graph and latent processes are shared. This ensures structural coherence across the task with instance-level variability.

Each task consists of a context set of $C = 7$ multivariate time-series and one query series, all generated from the same SCM but with independently sampled structural weights and noise. This construction ensures that context examples are conditionally informative for the query while preserving diversity across instances.

An illustrative example of a sampled causal graph and the corresponding context–query realizations generated by TimeSCM is shown in Figure 3, demonstrating shared structural mechanisms with instance-level variability across context and target series.



(a)

(b) Variate X_1

(c) Variate X_2

Figure 3: Schematic illustration of a synthetic forecasting task generated using the synthetic task generation framework. (a) Sampled structural causal graph defining the dependencies among observed and latent variables across time. (b, c) Time-series realizations of variates X_1 and X_2 for the query series (history and target) and two context examples. All trajectories are generated from the same underlying SCM but differ due to independently sampled structural weights, demonstrating structural consistency with contextual variation. The figure highlights the key property of the framework, where examples drawn from the same causal graph backbone can be used to produce related instances—enabling conditioning on the query instance during training of the In-Context TimePFN.

A.2 Qualitative performance of In-Context TimePFN on benchmark datasets

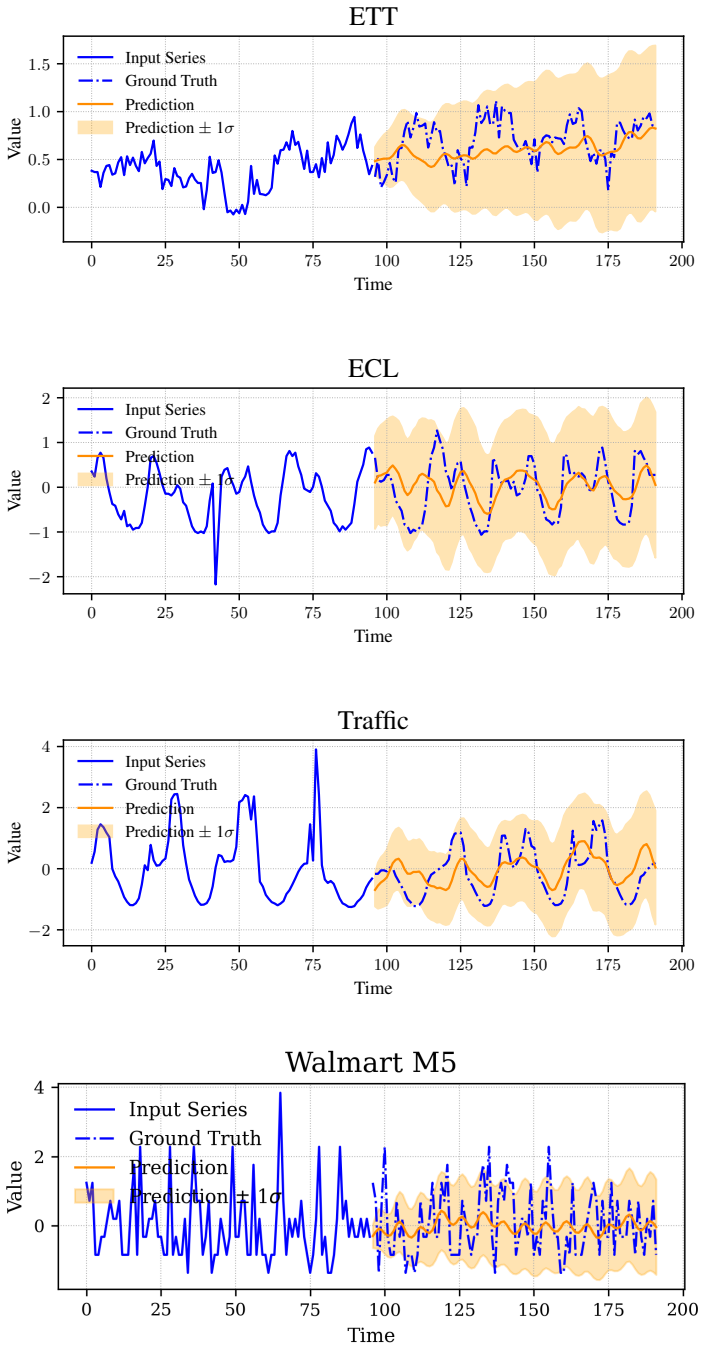


Figure 4: Zero-shot forecasting results from the In-Context TimePFN model across four real-world datasets: ETT, ECL, Traffic, and Walmart M5. Each plot shows the input series (first 96 steps), the ground truth target series (dashed blue), the PFN prediction (orange line), and the associated one-standard-deviation uncertainty band (shaded region).