Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.AS32414

A Review of Transformer Models and Their Variants for Predictive Tasks

Kejun Shen

Department of International Economics and Trade, Central University of Finance and Economics, Beijing, China
lorelai_shen@foxmail.com

Abstract. With an increase in task complexity (e.g., long horizon, high dimensions, multivariate and multimodal),However, traditional statistical models and recurrent neural networks are limited to the issues of scalability and modelling of longrange dependencies [The Transformer].which is based on the self-attention mechanism has been proved to be an effective alternative because of its parallel computing feature and capacity to capture global temporal dependency.The purpose of this section is to provide an overview about Transformers (and predictors based on it). We first summarize the basic building blocks of the vanilla Transformer and their modifications towards time series prediction. Next,efficient attention, multi-scale temporal modelling,cross-variable dependency learning, and multimodal fusion.We summarize typical uses in the context of univariate, multivariate, and multimodal forecast settings along with comparisons between their predictive performance and computation speed. Finally,major issues including the computational complexity, robustness to distributional shift and explainability are pointed out,and possible further research directions are given.The purpose of this survey is to offer an organized point of reference for the researcher or practitioner developing Transformersbased forecasting models.

Keywords: Transformer, time series prediction, efficient attentions, multivariate modelling, multimodal integration.

1. Introduction

In the last few years, forecasting problems are increasingly moving away from previous works focusing on univariate and small length time series towards harder scenarios with large horizons, multidimensionality, multivariate correlation and heterogenous multimodal data. This shift imposes much more stringent requirements for the prediction model, especially on the aspect of nonlinear representation ability, long-range dependency modeling and real-time inference performance. Traditional statistical approaches, recurrent neural networks (RNN), and long short-term memory (LSTM) based approaches, fail to provide an appropriate trade-off between the three aspects: predictive performance, scalability and computation cost. In reality, the conflict of precise prediction and lightweight implementation becomes more serious, underscoring both the scientific interest, as

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AS32414

well as the urgent need to study novel methods in deep learning that are better suited for prediction problems [1].

In this regard, the Transformer network based on self-attentions has been increasingly applied to areas outside of NLP, to diverse forecasting problems. Motivated with addressing the issue of tackling domain specific problem for time series predictions, many different Transformers have been suggested, improved upon, etc... All these models seek for a more efficient long-sequence modelling, better representation of both local and global temporal pattern, and provide improved robustness against the presence of noise and non-stationarity which is often encountered with real world data [2]. Reported results suggest that Transformers achieve better predictive performance than conventional recurrent neural network (RNN)/long short-term memory (LSTM) architectures for many practical time series prediction problems. However, such improvement comes with the cost of huge computation resource usage (e.g., large memory footprint or heavy GPU utilization), which makes them impractical for larger scale problems or applications that require low-latency predictions. In practice, a lot of work has gone into improving the efficiency of Transformers as well as how they are used at inference time [3].

In such a context, this survey aims at providing an organized view on both the evolution of prediction-focused Transformer flavors as well as their improving behaviors for predictions. More precisely, it summarizes the major model structure and transfer learning techniques for tasks, compares state-of-the-art methods in single-multivariate, and multimodal forecasting environments, and critically reviews the individual advantages and drawbacks of each approach. Besides pointing out some of the longstanding technical difficulties and avenues of further research, including efficient attention mechanisms, improved generalization in the small-data regime and lightweight model deployment. In doing so, we hope to provide a more systematic understanding and give a unified reference framework of researcher or practitioner who are involved in researching state-of-art forecasting methods.

2. Introduction to the Transformer

2.1. Core mechanisms of the original Transformer

For the prediction tasks, most of the basic units in the vanilla Transformer are as follows: attention layer, positional encoding, and the encoder–decoder framework.All together these parts allow for an efficient and flexible representation of sequences. Unlike recurrent structures, self-attention calculates pairwise dependency among any pair of time steps by the linear projection and scaled dot product of Query, Key, and Value vectors. This formulation enables long-range dependencies as well as the local temporal patterns to be learned in one forward pass, thus avoiding the notorious problems of gradient vanishing/exploding, and poor parallelism with respect to sequence length that plague recurrent networks (RNNs) on long sequences .

The multi-head attention further extends such a capacity, as it allows for learning simultaneously various kinds of dependency patterns within distinct subspaces. By decomposing the attention process into multiple heads, the Transformer can jointly capture diverse time series patterns such as trends and seasonality and sudden variations. In practical forecasting tasks, this multi-head structure provides a flexible mechanism for disentangling diverse temporal behaviors embedded in complex time series data [2].

DOI: 10.54254/2755-2721/2026.AS32414

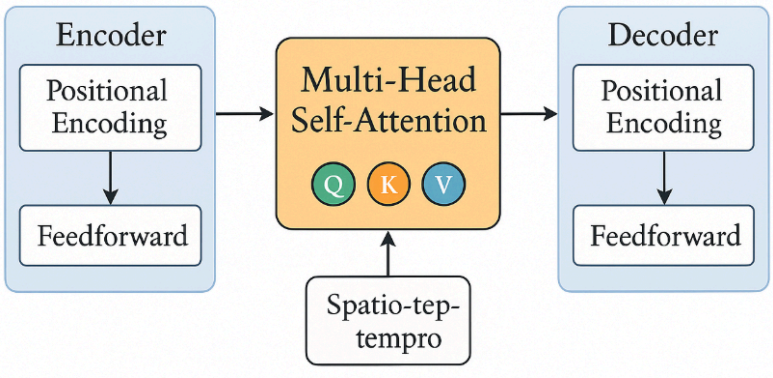


Figure 1. Core mechanisms of the original Transformer architecture

Given a feature vector for any particular time step of an input sequence (e.g., a power load value vector), three separate linear transformations produce the Query(Q), Key(K) and Value(V) representations where Q is the information demand in current time, K is the information from others, and V represent the contents that should be attended to, respectively. The attention output can be calculated by the scaled dot-product attention equation as follows :

$$\mathrm{Attention}\left(\mathrm{Q},\mathrm{K},\mathrm{V}\right)=\mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}$$

The scale constant is used to avoid huge dot-product numbers at a higher dimension which may cause the gradients of softmax to saturate. The multi-head attention projects Q, K, and V into multiple lower-dimensional subspaces, within which self-attention is computed independently and then concatenated. This design allows different attention heads to focus on distinct temporal characteristics at different scales (e.g., Head A captures fluctuations while Head B focuses on patterns).

Since the self-attention operation itself does not encode positional order, position encoding provides an important mechanism for introducing temporal structure within the Transformer. By utilizing either fixed sinusoidal encodings or learned positional embedding, the model can differentiate time steps and implicitly encode the relative time distance and periodical structure. This ability plays an especially crucial role in time series with strong daily, weekly or yearly patterns.

Structurally, the encoder extracts features from historical sequences hierarchically through stacking several layers consisting of multi-head attentions and position-wise feedforward networks while the decoder, on the other hand, generates future sequence in an auto-regressive way with masked self-attention restrictions that enables an end-to-end map from past observation to future prediction. Depending on the forecasting task, either the full encoder–decoder architecture can be used to make sequence-to-sequence predictions, or even just the encoder is stacked on top of a separate regression head, which outputs numeric forecasts (or distribution parameters) for future time points .

2.2. Adaptation strategies for prediction tasks

In order to fully benefit from the modelling advantages offered by Transformers or their variants in prediction tasks, some adaptations tend to be made at the input encoding level, output formulation,

DOI: 10.54254/2755-2721/2026.AS32414

and optimization objectives. As for the input side, a common practice consists in partitioning sequential data streams into overlapping "past-future" chunks via moving windows, thus transforming time series to a supervised machine learning problem amenable for an attention architecture. Moreover, structured inputs are often built on top of raw observations and enriched with features about the timestamps such as hour-of-day, day-of-week, and holiday indicators, along with domain knowledge priors. For the multivariate setting, normalization/standardization are commonly used in order to mitigate training instabilities due to different scales of variables, whereas data are also cleansed by means of the imputation of missing values and the removal of outliers. In long-sequence forecasting problems, the size of the historical window can be adaptively tuned depending on tasks, trying to achieve an operational compromise between the modeling of long-range time dependencies and maintaining acceptable computational efficiency .

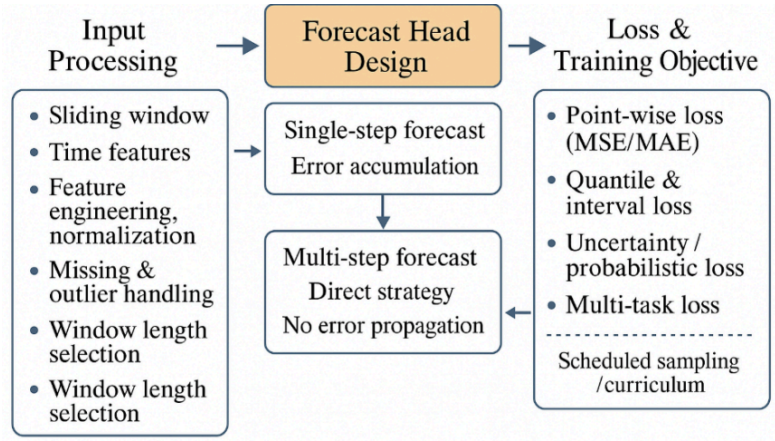


Figure 2. Designing Transformer-based forecasting: input processing, output strategy, and loss function adaptation

As illustrated in Fig. 2, the design of Transformer-based forecasting frameworks further requires careful consideration of output strategies and loss function selection. The form of the output depends on application needs. For instance, if a model must respond quickly (low-latency) to user queries or operate in near-real time, one-step ahead prediction is commonly used for predicting the value in current time step, or just after one more step. On the other hand, when dealing with medium-and long-term prediction/scheduling problems, one tends to favor multi-step ahead forecast because of its ability to provide the whole future sequence at once. Comparing the two multi-step methods we can see that recursive method reuses past predicted values which is in line with temporal causality, however it accumulates errors across a larger horizon. The straight forward solution, however, predicts several future time-steps in one go, which mitigates error accumulation but imposes more stringent global modelling assumptions on the network.

In terms of the objective function, recent works have progressively moved away from standard point estimation loss, such as mean squared error (MSE) and mean absolute error (MAE). More recently there have been efforts in developing quantile based losses, uncertainty aware formulation, and multi-task joint objectives which try to jointly optimize the trade-off between prediction accuracy, interval coverage and robustness in one objective function. In addition, methods like scheduled sampling were proposed to address the mismatch in distribution of inputs at train time versus inference, thus increasing the robustness of models in production settings.

DOI: 10.54254/2755-2721/2026.AS32414

3. Introduction to Transformer variants

While it is evident that the original Transformer has certain strengths in terms of modeling long-range dependency, as well as enabling parallel computation, the quadratic time/space complexity of its vanilla attention operation turns out to be the bottleneck for long-sequence prediction or resource-efficient deployment.With a main goal to keep good performance on forecast while not increasing too much computation, various forecasting-specific Transformers are developed based on this structure in forecasting task. All these variants generally improve the Transformer's applicability to longer temporal horizons, higher dimensional observations, and even more complicated settings through controlled perturbations of the attention process, input representation, and interactions between variables [4,5].

From the perspective of efficient attention design, mainstream approaches aim to reduce the overhead of fully connected self-attention through sparsification, approximation, or implementation-level optimization. Representative techniques include ProbSparse attention for long-sequence forecasting, hashing-based compression, and hardware-aware exact attention implementations such as FlashAttention, which improve memory efficiency and practical throughput while preserving the core attention computation [6]. In practice, these approaches can retain critical temporal dependencies while substantially lowering the training and inference burden, thereby making Transformer-based forecasting more feasible for long historical windows and online rolling prediction.

Many Transformer variants also explicitly add decomposition for the long sequence to global trend and local fluctuation, typically combined with multi-scale representations. This choice improves the capacity of the model to represent long-term regularities, whilst still being sensitive to shorter term variation.For example, with the addition of trend–seasonality decomposition, frequency domain transformations or segmented and patch-based representations, which enable models to independently capture trends and periodicities as well as short-term noise with various time scales; this helps address several limitations of long-term prediction with the advantages of avoiding oversmoothing as well as phase shift while increasing robustness to non-stationary time series.

As for the multivariate forecasting problem, what is important to tackle for the Transformer variant is how we could catch not only the temporal dependence inside each time series but also the correlation among various variables simultaneously.

As stated above, several papers showed that by adding extra modules like a variable selection network, feature-wise attentions, and cross-attentions, these modules help the model dynamically select important driving features and capture relations between these features.

Meanwhile, some studies change the dimensionality structure of the attention operation in order to enhance the expression ability of high-dimensional correlation and decrease useless calculation, that is of particular importance for large scale multivariate problems .

As forecasting applications extend beyond single numerical sequences to richer contextual settings that may involve textual, spatial, visual, or graph-structured signals, integrating heterogeneous information has become an important direction in forecasting model design [7]. In this setting, a key objective is no longer limited to modeling temporal dependence within one sequence, but also includes coordinating complementary information from multiple sources so that contextual signals can assist prediction. Accordingly, recent forecasting studies increasingly emphasize contextual fusion, structured exogenous-variable modeling, and flexible attention-based aggregation mechanisms.

In general, the Transformer variations for forecasting keep developing in multiple correlated directions such as efficient attention mechanisms, multi-scale temporal modelling, cross-variable

DOI: 10.54254/2755-2721/2026.AS32414

dependency modelling and multimodal fusion. These developments provide the methodological underpinning to the application-oriented analyses that we present next.

To enable a useful comparison and discussion of the predictive performance between different Transformers, we first briefly explain the evaluation criteria which are widely used for prediction problems.In practice, for Transformer-based forecasting, researchers usually evaluate their model in terms of some pre-defined evaluation criteria measuring forecast accuracy as well as robustness. Mean squared error (MSE) and root mean squared error (RMSE) are the most commonly employed ones as they penalize larger errors more severely than others, having a stronger sensitivity for extreme fluctuations and peaks.In order to be able to more effectively measure the scale invariant behaviour (i.e., when dealing with variables of different orders of magnitude and/or which may not have a stationary distribution), absolute error measures like the mean absolute error (MAE), but also relative ones, e.g., the mean relative error (MRE) or symmetric relative error (SRE) are used.

Relative measures focus on relative errors and complement absolute error measurements in their ability to assess a models' performance with respect to distributional shift.

While we do not perform a common set of experiments and compare results, the comparisons made between various types of Transformers in later sections rely on an improvement in metrics that is commonly seen across papers.In summary, the available results seem to indicate that effective attention mechanism and multiscale modeling strategy could constantly decrease MSE and RMSE for long-term forecasts with a rather steady relative error, and thus provide an operational tradeoff between the model's forecasting power versus its complexity.

4. Typical applications of Transformers and their variants in forecasting tasks

4.1. Univariate time series forecasting

Univariate time series forecasting predicts future values solely from the historical observations of a single target variable and is widely applied in scenarios such as power load, electricity price, weather, stock price, and exchange rate forecasting. Compared with traditional statistical models and recurrent neural networks, Transformer-based models show stronger capability in capturing long-range dependencies, modeling periodic patterns, and handling sudden fluctuations through parallel self-attention mechanisms. Their practical advantage lies in the ability to flexibly match different domain characteristics: long-sequence models are well suited to strongly periodic tasks such as power load forecasting, while lightweight or frequency-aware variants are more appropriate for highly volatile financial and exchange-rate series, where rapid changes and deployment efficiency must both be considered [8]. Existing studies indicate that Transformer variants can achieve better long-horizon forecasting accuracy and computational efficiency, while frequency-domain decomposition and lightweight architectural design further improve their ability to capture local fluctuations, trend reversals, and overall temporal trajectories in real-world forecasting tasks [9].

DOI: 10.54254/2755-2721/2026.AS32414

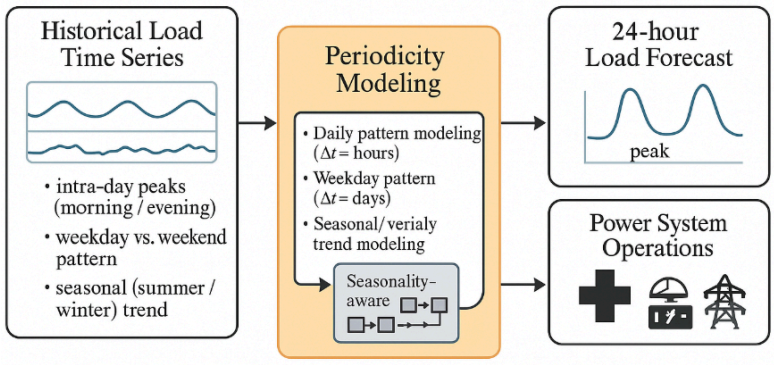


Figure 3. Framework of periodicity modeling for univariate power load forecasting

4.2. Multivariate time series forecasting

The key problem for multivariate TS prediction is how to simultaneously model both the pattern of temporal evolution as well as the coupling relationship between variables along the variable dimension. In practice, multiple observed variables may cause the system states to change in a non-linear and multi-scale way. For instance, in weather prediction, temperature, humidity, wind speed and air pressure collectively affect rain generation, while in industrial equipment monitoring, the combined variation of vibration, temperature, and pressure signals potential fault risks. Traditional approaches, such as the ARIMA model and LSTM based methods, often rely on handcrafted cross-features or pre-defined fusion rules. Hence, they are unable to effectively model long-range dependencies in the case of large-scale multivariate input datasets . Transformer-based models overcome these issues as they are capable of jointly learning both time series dynamics within each variable, i.e., long-term memory effects influencing the present value of some variable, and cross-correlations between variables that represent cooperative/inhibitory interaction between various variables. By using the multi-head attention mechanism, these models could automatically select important variables to learn the interaction pattern among them, which does not rely on any strong priors. It is shown through experiments that such a feature brings obvious advantages for multivariate prediction tasks, especially in complicated, dynamic and tightly-coupled situations.

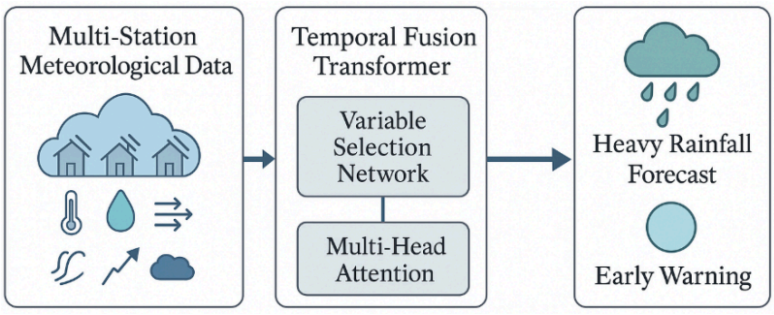


Figure 4. Multivariate time-series forecasting: a multi-station Transformer framework for heavy rainfall prediction

Multi-station meteorological forecasting provides a representative illustration of this capability. In such settings, forecasting models must jointly process observations from multiple stations and multiple variables, including temperature, humidity, wind speed, and atmospheric pressure. Beyond learning nonlinear cross-variable relationships, the model must also capture cross-station temporal

DOI: 10.54254/2755-2721/2026.AS32414

interactions and horizon-dependent feature relevance. The Temporal Fusion Transformer (TFT), for example, introduces variable selection mechanisms that can adaptively emphasize informative covariates at different prediction horizons, while attention layers help model long-range temporal dependencies within the forecasting window . From a methodological perspective, this makes TFT and related architectures well suited to multivariate forecasting problems in which feature importance changes over time and strong prior fusion rules are unavailable.

4.3. Multimodal fusion forecasting

Multimodal fusion forecasting aims to jointly exploit time series data together with heterogeneous information sources such as text, images, graphs, or other contextual signals, so that complementary cues dispersed across different modalities can be incorporated into a unified predictive framework [10]. Compared with unimodal approaches, this line of research places greater emphasis on cross-source coordination, contextual representation, and adaptive fusion. The main difficulties arise from semantic alignment across heterogeneous inputs and from the fact that the relative contribution of each information source may vary over time and across operating conditions. Owing to their flexible attention-based aggregation mechanisms, Transformer architectures are often regarded as a promising foundation for building modality-specific encoders and for performing dynamic information integration in richer forecasting pipelines .

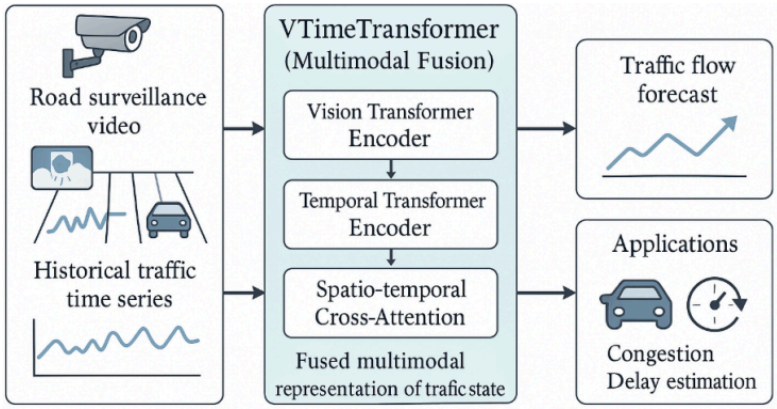


Figure 5. Multimodal fusion for traffic flow forecasting with VTimeTransformer

Taking traffic flow forecasting as a representative example, intelligent transportation systems often require not only historical traffic flow sequences, but also richer contextual information such as periodic patterns, network-level interactions, and exogenous operating conditions. In this context, the Traffic Transformer proposed by Cai et al. focuses on capturing continuity and periodicity in traffic time series, showing how Transformer-style architectures can model recurring temporal structure in urban traffic systems [11]. Although this work is primarily centered on traffic time-series forecasting rather than full visual–temporal fusion, it usefully illustrates the broader trend of extending forecasting models beyond single-sequence inputs toward more context-aware predictive frameworks.

More broadly, traffic forecasting studies suggest that integrating temporal attention with structured contextual information can improve robustness under rush-hour congestion, holiday effects, and other nonstationary operating conditions, thereby providing useful support for congestion warning, dispatch coordination, and intelligent transportation management. These

DOI: 10.54254/2755-2721/2026.AS32414

observations reinforce the practical value of richer-input forecasting systems even when the available information sources are heterogeneous and dynamically changing.

5. Existing challenges and future directions

Despite the substantial advances of forecasting-oriented Transformer variants, their large-scale and reliable deployment in real-world forecasting tasks remains constrained by several persistent challenges. Although mechanisms such as sparse attention and efficient architectural design have reduced the computational burden of standard self-attention, achieving an effective balance among forecasting accuracy, latency, and memory consumption is still a key issue, especially in resource-constrained scenarios [12]. At the same time, practical time series forecasting often involves noisy, incomplete, heterogeneous, and dynamically shifting data, which limits model robustness and generalization under long-term structural changes, extreme events, and complex multimodal settings. Future research should therefore focus on developing more lightweight and hardware-efficient Transformer architectures through strategies such as sparse or low-rank approximation, pruning, quantization, and distillation, so as to improve deployment efficiency while maintaining predictive performance[13]. In addition, uncertainty modeling and ensemble-based forecasting deserve greater emphasis, as quantile prediction, ensemble aggregation, and related post-processing techniques can enhance robustness, better characterize prediction risk, and improve operational reliability under regime shifts and extreme conditions [14].

6. Summary

This paper provides a structured review of transformer-based forecasting models by summarizing their modeling foundations, representative variants, application scenarios, and practical performance. It first introduces the core components of Transformers, including self-attention, positional encoding, and the encoder–decoder architecture, and highlights their advantages over traditional ARIMA and RNN/LSTM models in capturing long-range dependencies and enabling parallel computation. On this basis, the review further outlines major forecasting-oriented architectural enhancements, such as sparse attention, seasonal–trend decomposition, frequency-domain modeling, and multivariate or multimodal fusion mechanisms. Representative applications are then discussed across univariate, multivariate, and richer-context forecasting tasks, covering power load and electricity price forecasting, high-frequency financial analysis, exchange rate prediction, industrial process monitoring, and traffic flow modeling, where different Transformer variants demonstrate strong adaptability to scenario-specific characteristics and complex dependency structures. A comparative synthesis of existing experimental results suggests that the Transformer family has generally outperformed many traditional approaches in long-horizon forecasting and flexible representation learning for complex tasks. However, important challenges remain in computational complexity, resource-efficient deployment, data distribution drift, heterogeneous contextual alignment, and robustness under extreme conditions. Overall, this review aims to provide a coherent reference and methodological insight for future research and practical deployment of Transformer-based forecasting models in power systems, finance, industrial processes, and urban computing.

References

[1] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting, " Int. J. Forecast., vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

DOI: 10.54254/2755-2721/2026.AS32414

[2] A. Vaswani et al., "Attention is all you need, " in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, pp. 5998–6008, 2017.
[3] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting, " in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 12, pp. 11106–11115, 2021.
[4] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, " in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, pp. 22419–22430, 2021.
[5] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer, " in Proc. Int. Conf. Learn. Represent. (ICLR), 2020.
[6] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness, " arXiv preprint arXiv: 2205.14135, 2022.
[7] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey, " arXiv preprint arXiv: 2202.07125, 2022.
[8] H. Luo, "Research on Time Series Forecasting Algorithm Based on Transformer, " M.S. thesis, Beijing Jiaotong Univ., Beijing, China, Sep. 2024.
[9] A. Nguyen, S. Ha, and N. Phien, "LiteFormer: An encoder-only multi-head attention transformer for financial time series forecasting, " SSRN Electron. J., 2024, doi: 10.2139/ssrn.4729648.
[10] J. Kim, H. Kim, H. Kim, D. Lee, and S. Yoon, "A comprehensive survey of time series forecasting: Architectural diversity and open challenges, " arXiv preprint arXiv: 2411.05793, Oct. 2024.
[11] L. Cai, K. Janowicz, G. Mai, B. Yan, and R. Zhu, "Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting, " Trans. GIS, vol. 24, no. 3, pp. 736–755, Jun. 2020, doi: 10.1111/tgis.12644.
[12] M. Wang, J. Yang, B. Yang, H. Li, T. Gong, B. Yang, and J. Cui, "Towards lightweight time series forecasting: A patch-wise transformer with weak data enriching, " arXiv preprint arXiv: 2501.10448, Jan. 2025.
[13] Y. Wang, Y. Qiu, P. Chen, Y. Shu, Z. Rao, L. Pan, B. Yang, and C. Guo, "LightGTS: A lightweight general time series forecasting model, " arXiv preprint arXiv: 2506.06005, Jun. 2025.
[14] D. D. Modi and R. Pan, "Enhancing transformer-based foundation models for time series forecasting via bagging, boosting and statistical ensembles, " arXiv preprint arXiv: 2508.16641, Aug. 2025.