Satellite-Text-Prompted Large Language Model for Photovoltaic Power Forecasting

Pengfei Jia, Jianghong Ma, Baoquan Zhang, Kenghong Lin, Xinyu Zhang, Chuyao Luo, Xutao Li, Yunming Ye*

Shenzhen Key Laboratory of Internet Information Collaboration, Harbin Institute of Technology, Shenzhen
jiapengfei@stu.hit.edu.cn, yeyunming@hit.edu.cn

Abstract

Photovoltaic (PV) power forecasting is critical for the operation of solar power plants and the coordination of energy within power grids. This work aims to predict future PV power time series by leveraging multimodal data. While recent studies have incorporated numerical modalities such as satellite image sequences and numerical weather prediction (NWP) time series, they often overlook textual modalities—such as the spatio-temporal context of PV plants—and the potential of pretrained large language models (LLMs). In this paper, we build upon existing numerical inputs and further explore the use of spatio-temporal text prompts, generated based on plant coordinates and forecast start time, to enhance the forecasting process. We propose PV-LLM, a satellite-text-prompted framework that integrates a pretrained LLM to improve PV power forecasting. The framework consists of three key components: Text Prompt Construction, Modality-Specific Encoding, and Adaptive Prompt Tuning. First, the Text Prompt Construction module generates spatio-temporal prompts that offer high-level semantic guidance. Next, the Modality-Specific Encoding module encodes each modality according to its unique characteristics, capturing modality-specific patterns while managing varying context lengths. Finally, the Adaptive Prompt Tuning module fine-tunes the LLM to integrate multimodal embeddings, while an adaptive gating mechanism retains its pretrained knowledge. We validate the effectiveness of the proposed framework on a real-world dataset containing multiple PV plants. Experimental results demonstrate that our approach outperforms existing state-of-the-art methods.

Introduction

Photovoltaic (PV) power forecasting is a complex and critical challenge, especially as the global installed capacity of PV systems continues to grow at an unprecedented pace. According to recent data from the International Energy Agency, solar PV capacity additions reached approximately 451.9 GW in 2024, marking a 32% increase compared to 2023 (International Renewable Energy Agency (IRENA) 2025), with solar PV accounting for three-quarters of all newly installed power capacity. Accurate and timely PV power forecasting is essential for effective grid management and the optimized operation of solar power plants (Babaei, Zhao, and Fan 2019; Makarov et al. 2009). Unlike conventional power generation, PV systems exhibit a pronounced diurnal production cycle, closely tied to sunlight availability. Additionally, their efficiency is significantly influenced by geographic and meteorological conditions, leading to considerable fluctuations in power output (Nguyen et al. 2016). These characteristics make PV power forecasting fundamentally different from general-purpose time series forecasting tasks.

Traditional PV forecasting methods (Lorenz et al. 2011; Saint-Drenan et al. 2015) rely on physical modeling and numerical weather prediction (NWP) data to estimate future PV output. However, the predictive accuracy of such methods is constrained by the nonlinear nature of physical systems. Recent data-driven methods (Li, Su, and Shu 2014; Ma et al. 2024; Kim and Suh 2024) have leveraged deep neural architectures to improve pattern extraction and temporal representation. To overcome the coarse spatio-temporal resolution of NWP data, satellite imagery has been incorporated as an auxiliary modality. Yet, the vast amount of visual data can introduce redundant information and noise, making effective multimodal alignment a persistent challenge.

PV power forecasting exhibits strong diurnal periodicity and spatial correlations across geographically distributed PV plants (Mayer and Gróf 2021). In practice, human forecasters often rely on contextual knowledge—such as the geographic location of PV plants—which can be naturally conveyed through language. Despite this, existing methods primarily focus on numerical data and largely neglect the textual modality. Unlike raw sensory inputs, textual descriptions provide high-level semantic cues, such as plant coordinates, forecast start time, or expected peak values—attributes that are especially important for accurately predicting power peaks. Designing effective text prompts that align with the spatio-temporal nature of PV forecasting remains largely underexplored.

Meanwhile, large language models (LLMs) have demonstrated exceptional modeling and generative capabilities in natural language processing (Radford et al. 2019; Ouyang et al. 2022). Recent studies (Hu et al. 2025; Jin et al. 2024) have begun to investigate the use of pretrained LLMs for time series forecasting, but results have been mixed across domains. The applicability of LLMs to PV power forecasting has not yet been thoroughly studied, in part due to the

*Corresponding author

significant modality gap between multimodal PV data and the textual pretraining corpus of LLMs. Effectively bridging this gap and enabling LLMs to reason over PV-related multimodal data remains a significant open problem.

To address these issues, we propose PV-LLM, a satellite-text-prompted framework for PV power forecasting that leverages pretrained LLMs. PV-LLM integrates numerical data (including satellite image sequences and NWP time series) with spatio-temporal text prompts, and consists of three main components: Modality-Specific Encoding, Text Prompt Construction, and Adaptive Prompt Tuning. The modality-specific encoder processes numerical data while reducing the modality gap between sequential inputs and language representations. A Temporal-Semantic Encoder is designed for PV and NWP time series, while a satellite image encoder compresses remote sensing inputs into compact satellite prompts. The Text Prompt Construction module generates rich spatio-temporal descriptions using metadata such as plant coordinates and forecast start time, along with dataset descriptions, task descriptions, and statistical summaries to provide semantic guidance for LLMs. Finally, the Adaptive Prompt Tuning module fine-tunes the LLM to integrate multimodal embeddings, utilizing an adaptive gating mechanism to preserve pretrained knowledge. Satellite prompts are injected into intermediate layers of the LLM, and the gating is initialized to mitigate early training noise.

Our main contributions are as follows:

- We introduce the textual modality into PV power forecasting by generating spatio-temporal prompts from plant metadata, enabling synergistic integration with time series and satellite modalities.
- We propose PV-LLM, a novel framework that leverages pretrained LLMs and modality-specific encoders to balance pattern differences and context lengths, while adaptively tuning the LLM for multimodal fusion.
- We validate our approach on a real-world dataset containing multiple PV plants, and experimental results demonstrate that PV-LLM significantly outperforms existing state-of-the-art methods.

Related Work

PV Power Forecasting

Based on modeling principles, PV power forecasting methods can be broadly categorized into non-machine-learning methods (Lorenz et al. 2011; Mayer and Gróf 2021) and machine learning methods (Li, Su, and Shu 2014; Alfadda et al. 2017; Elizabeth Michael et al. 2022).

Non-machine-learning methods construct mathematical models of PV systems using physical principles and statistical relationships. For example, one study (Mayer and Gróf 2021) developed 32,400 sub-models by combining irradiance transposition techniques with specific PV module parameters, and validated them using high-resolution data from 16 PV plants. Persistence method (Sarmas et al. 2022) uses the previous day's power output as the prediction for the current day. Mean method (Ma et al. 2024) computes the average PV output from all historical sequences in the training set. Clear Sky (Ineichen 2016) estimates theoretical global horizontal irradiance based on the prediction time and geographic location, assuming no cloud cover. However, due to limitations in modeling precision and the scale of available data, non-machine-learning approaches often underperform in complex or rapidly changing operational scenarios.

Machine learning methods apply data-driven methods to automatically extract features for prediction. Early work utilized traditional techniques such as ARIMA (Li, Su, and Shu 2014) and support vector regression (Alfadda et al. 2017). More recently, deep learning-based methods (Elizabeth Michael et al. 2022; Ma et al. 2024) have become mainstream. For instance, one method (Elizabeth Michael et al. 2022) introduced a hybrid CNN-LSTM model to enhance short-term PV forecasting. Multimodal fusion strategies have also been proposed to improve prediction accuracy (Kim and Suh 2024; Ma et al. 2024). FusionSF (Ma et al. 2024), for example, employed a cross-attention mechanism to fuse satellite image sequences with PV data. However, satellite imagery introduces a substantially larger input volume than time series data, which may dilute task-relevant patterns and hinder effective learning. More importantly, existing methods often overlook the textual modality, which can convey high-level semantic information such as plant coordinates and forecast start time.

Deep Learning for Time Series Forecasting

Transformer-based (Zhou et al. 2021; Wu et al. 2021; Zhang and Yan 2023; Nie et al. 2023; Wang et al. 2024b) and MLP-based methods (Zhang et al. 2022; Zeng et al. 2023; Wu et al. 2023; Zhang et al. 2024; Murad, Aktukmak, and Yilmaz 2025) currently dominate deep learning for time series forecasting. Autoformer (Wu et al. 2021) incorporates a decomposition architecture that models trend and seasonal components using autocorrelation-based attention. DLinear (Zeng et al. 2023) critiques the permutation-invariant nature of self-attention and instead employs a decomposition-driven linear structure to preserve temporal order. TimeXer (Wang et al. 2024b) combines exogenous information via variate-wise cross-attention and patch-wise self-attention, improving multivariate time series modeling. Although these methods effectively model time-series patterns, they remain unimodal and fail to leverage complementary information from other modalities such as images or text, which can enhance contextual understanding.

Recent advances have shown the potential of integrating text into time series forecasting (Wang et al. 2024a, 2025), though its application in PV forecasting remains underexplored. In parallel, several studies (Hu et al. 2025; Zhou et al. 2023; Jin et al. 2024) have adapted LLMs for time series forecasting with promising results. However, other works (Tan et al. 2024) have reported limited or inconsistent improvements in certain forecasting tasks, suggesting that the effectiveness of LLMs strongly depends on inherent domain characteristics and data modalities. The applicability of LLMs to PV power forecasting, which involves complex multimodal and spatio-temporal dependencies, remains largely unexplored and represents an important open research challenge.

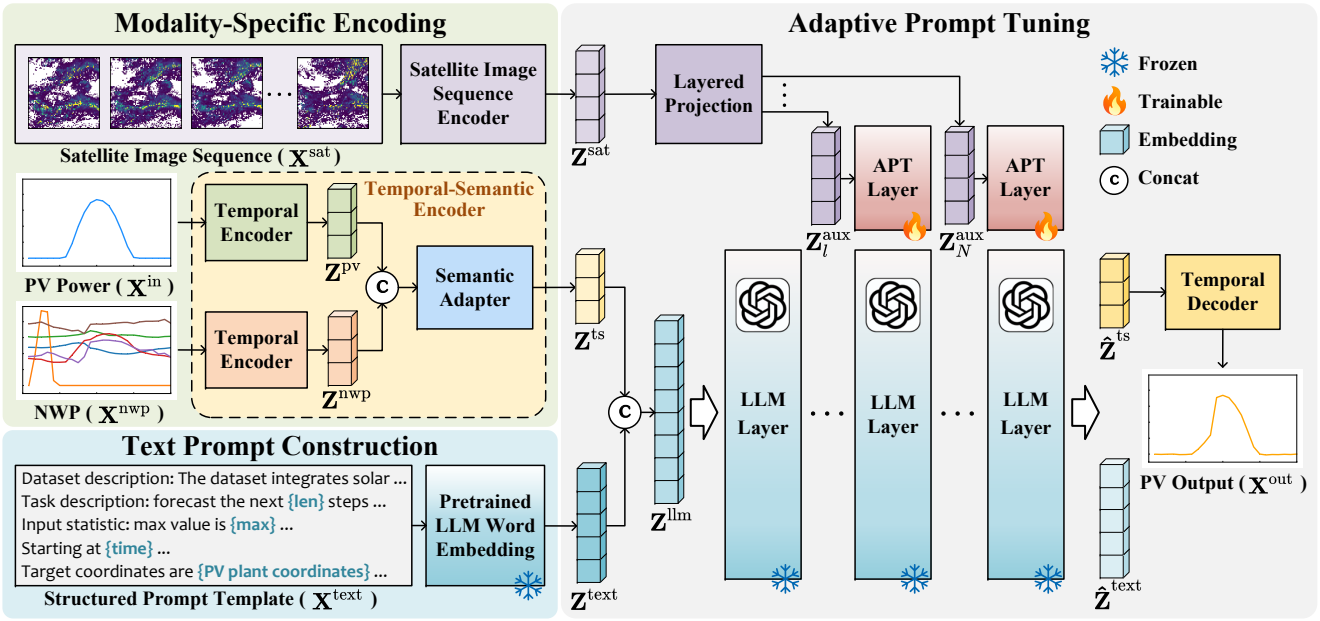

Figure 1: The overall architecture of the proposed PV-LLM method. The model accepts multimodal inputs, including PV power sequences, NWP time-series data, satellite image sequences, and a structured prompt template. The Modality-Specific Encoding (MSE) module processes time-series and satellite inputs, producing embeddings $\mathbf{Z}^{\text{ts}}$ and $\mathbf{Z}^{\text{sat}}$. The Text Prompt Construction (TPC) module generates spatio-temporal text prompts composed of dataset descriptions, task descriptions, input statistics, and PV plant coordinates. These prompts are encoded into embeddings $\mathbf{Z}^{\text{text}}$ via a pretrained LLM. All embeddings are then integrated by the Adaptive Prompt Tuning (APT) module to generate the final PV power forecast.

Methodology

Problem Formulation

The solar PV power forecasting task aims to predict a future sequence of PV power values, denoted as $\mathbf{X}^{\text{out}} = \{x_{t+1}, \ldots, x_{t+m}\}$, over a forecasting horizon of m time steps. The input consists of a historical PV power time series $\mathbf{X}^{\text{in}} = \{x_{t-n+1}, \ldots, x_t\}$, where n is the input length and $x_t \in \mathbb{R}$ represents the PV power at time t.

This study focuses on PV power forecasting based on multimodal inputs. Specifically, we incorporate both conventional NWP time-series data and the recently emerging satellite image sequence data. Furthermore, we introduce a novel modality—spatio-temporal text prompts. These prompts are constructed from a structured prompt template. Let $\mathbf{X}^{\text{nwp}} \in \mathbb{R}^{m \times V}$ denote the NWP data and $\mathbf{X}^{\text{sat}} \in \mathbb{R}^{n \times C \times H \times W}$ denote the satellite image sequence. Let $\mathbf{X}^{\text{text}}$ represent the structured prompt template. Here, V is the number of NWP variables, C is the number of image channels, and $H \times W$ is the spatial resolution.

To effectively integrate these heterogeneous modalities, we propose a satellite-text-prompted method, denoted as $f_{\boldsymbol{\theta}}(\cdot)$, where $\boldsymbol{\theta}$ represents the model parameters. In addition, we include auxiliary spatial information: the coordinates of the PV plant, $\mathbf{C}^{\text{in}} \in \mathbb{R}^2$, and the coordinates associated with the satellite image grid, $\mathbf{C}^{\text{sat}} \in \mathbb{R}^{2 \times H \times W}$. The PV power forecasting problem in this study can be formulated as:

$$\hat{\mathbf{X}}^{\text{out}} = f_{\boldsymbol{\theta}}(\mathbf{X}^{\text{in}}, \mathbf{X}^{\text{nwp}}, \mathbf{X}^{\text{sat}}, \mathbf{X}^{\text{text}}, \mathbf{C}^{\text{in}}, \mathbf{C}^{\text{sat}}). \quad (1)$$

Overview

Building on the above multimodal setting, we propose a satellite-text-prompted PV power forecasting model with a pretrained LLM backbone, termed **PV-LLM**. As shown in Fig. 1, PV-LLM contains three modules: MSE, TPC, and APT. MSE encodes the PV power time series, NWP data, and satellite image sequences to produce embeddings $\mathbf{Z}^{\text{ts}}$ and $\mathbf{Z}^{\text{sat}}$. TPC constructs spatio-temporal text prompts using dataset and task descriptions, input statistics, forecast start time, and PV plant coordinates. These prompts are converted into text embeddings $\mathbf{Z}^{\text{text}}$ by the pretrained LLM's embedding layer. APT integrates all embeddings and performs prompt tuning to generate the final PV forecast while preserving pretrained knowledge.

Modality-Specific Encoding

To capture modality-specific patterns, the Modality-Specific Encoding module encodes each modality according to its unique characteristics. It consists of two primary components: the Temporal-Semantic Encoder (TE) and the Satellite Image Sequence Encoder (SE). The TE module encodes PV power and NWP sequences to produce time-series embeddings $\mathbf{Z}^{\text{ts}}$, while the SE module processes satellite image sequences and learns spatial correlations based on satellite coordinates $\mathbf{C}^{\text{sat}}$, yielding satellite prompt embeddings $\mathbf{Z}^{\text{sat}}$. All intermediate embeddings are represented as per-sample tensors $\mathbf{Z} \in \mathbb{R}^{L \times D}$, where L is the sequence length, and D is the embedding dimension determined by the corresponding encoder.

Formally, let $f_{\boldsymbol{\theta}_{te}}(\cdot)$ and $f_{\boldsymbol{\theta}_{se}}(\cdot)$ denote the TE and SE encoders, respectively. The MSE module is defined as:

$$\mathbf{Z}^{\text{sat}} = f_{\boldsymbol{\theta}_{se}}(\mathbf{X}^{\text{sat}}, \mathbf{C}^{\text{sat}}), \tag{2}$$

$$\mathbf{Z}^{\text{ts}} = f_{\boldsymbol{\theta}_{te}}(\mathbf{X}^{\text{in}}, \mathbf{X}^{\text{nwp}}). \tag{3}$$

Temporal-Semantic Encoder The TE encodes PV power and NWP time-series data while bridging the modality gap between time-series data and natural language. It consists of three main components: a PV power temporal encoder, an NWP temporal encoder, and a Semantic Adapter (SA).

The PV power temporal encoder processes the PV data to produce temporal embeddings $\mathbf{Z}^{\text{pv}}$, while the NWP temporal encoder encodes the NWP data into embeddings $\mathbf{Z}^{\text{nwp}}$. These embeddings are then fused via the Semantic Adapter to generate the final time-series embeddings $\mathbf{Z}^{\text{ts}}$. For $\mathbf{Z}^{\text{pv}}$ and $\mathbf{Z}^{\text{nwp}}$, the embedding dimension D follows the configuration in FusionSF (Ma et al. 2024). Let $f_{\boldsymbol{\theta}_{pe}}(\cdot)$, $f_{\boldsymbol{\theta}_{ne}}(\cdot)$, and $f_{\boldsymbol{\theta}_{sa}}(\cdot)$ denote the PV power encoder, the NWP encoder, and the Semantic Adapter, respectively. The overall formulation of the TE module is as follows:

$$\mathbf{Z}^{\text{pv}} = f_{\boldsymbol{\theta}_{pe}}(\mathbf{X}^{\text{in}}), \tag{4}$$

$$\mathbf{Z}^{\text{nwp}} = f_{\boldsymbol{\theta}_{ne}}(\mathbf{X}^{\text{nwp}}), \tag{5}$$

$$\mathbf{Z}^{\text{ts}} = f_{\boldsymbol{\theta}_{sa}}(\text{Concat}(\mathbf{Z}^{\text{pv}}, \mathbf{Z}^{\text{nwp}})). \tag{6}$$

Temporal Encoder. The Temporal Encoder comprises a patch embedding layer followed by a lightweight Transformer backbone (Vaswani et al. 2017). The patch embedding layer segments the input time series into fixed-length patches and projects them linearly. The lightweight Transformer backbone encodes these patches into temporal embeddings. Its core component is Multi-Head Attention (MHA), which captures dependencies across time steps and attends to diverse representation subspaces. Given queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, MHA is formulated as:

$$\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^{\mathbf{O}}, \tag{7}$$

$$\text{head}_i = \text{Attention}(\mathbf{Q}W_i^{\mathbf{Q}}, \mathbf{K}W_i^{\mathbf{K}}, \mathbf{V}W_i^{\mathbf{V}}), \tag{8}$$

where $W_i^{\mathbf{Q}}$, $W_i^{\mathbf{K}}$, $W_i^{\mathbf{V}}$, and $W^{\mathbf{O}}$ are learnable projection matrices. When $\mathbf{Q} = \mathbf{K} = \mathbf{V}$, the MHA operation becomes Multi-Head Self-Attention (MHSA).

Semantic Adapter. Pretrained LLMs are primarily optimized for textual input and cannot directly process time-series modalities. To address this limitation, we introduce a Semantic Adapter that maps time-series embeddings into the LLM word embedding space. The adapter first constructs a set of word prototypes $\mathbf{Z}^{\text{wp}}$ by linearly projecting the pretrained word embeddings $\boldsymbol{\theta}_{we}$. It then applies a single-layer cross-attention module, where $\mathbf{Z}^{\text{wp}}$ serves as both the key and the value, and the concatenated temporal embeddings $\text{Concat}(\mathbf{Z}^{\text{pv}}, \mathbf{Z}^{\text{nwp}})$ serve as the query:

$$\mathbf{Z}^{\text{wp}} = \text{Linear}(\boldsymbol{\theta}_{we}), \tag{9}$$

$$\mathbf{Z}^{\text{ts}} = \text{MHA}(\text{Concat}(\mathbf{Z}^{\text{pv}}, \mathbf{Z}^{\text{nwp}}), \mathbf{Z}^{\text{wp}}, \mathbf{Z}^{\text{wp}}). \tag{10}$$

The embedding dimension D of $\mathbf{Z}^{\text{sat}}$, $\mathbf{Z}^{\text{ts}}$, and $\mathbf{Z}^{\text{wp}}$ equals the hidden size C^{llm} of the pretrained LLM.

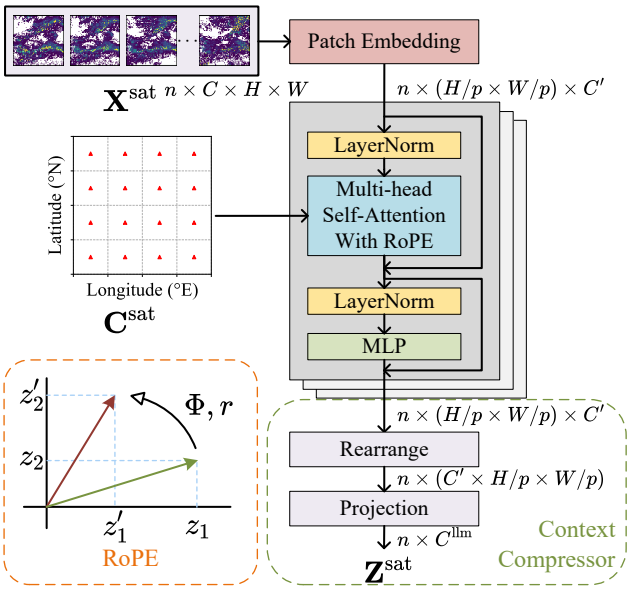

Figure 2: Satellite Image Sequence Encoder. The SE module generates satellite prompt embeddings while effectively reducing the context length.

Satellite Image Sequence Encoder The Satellite Image Sequence Encoder (SE) encodes satellite image sequences into satellite prompt embeddings $\mathbf{Z}^{\text{sat}}$. Since geographical location plays a critical role in PV power forecasting, the SE incorporates spatial coordinates $\mathbf{C}^{\text{sat}}$ of image patches as auxiliary inputs. Specifically, the coordinates of each patch center are used as its positional encoding. To model spatial relationships, we apply Rotational Position Encoding (RoPE) (Su et al. 2024), a relative position encoding technique that integrates position information into the self-attention mechanism via geometric transformations. RoPE represents embeddings as complex vectors and applies rotation matrices based on relative positions. The RoPE-enhanced Multi-Head Self-Attention (MHSAR) is computed as:

$$\text{MHSAR}(\mathbf{Z}) = \text{MHA}(\mathbf{R}_{\boldsymbol{\Phi},r}\mathbf{Z}, \mathbf{R}_{\boldsymbol{\Phi},r}\mathbf{Z}, \mathbf{Z}), \tag{11}$$

where $\mathbf{R}_{\boldsymbol{\Phi},r}$ is a rotation matrix determined by the relative position r and angle set $\boldsymbol{\Phi}$, which is a constant determined by the embedding dimension.

Context Compressor. As illustrated in Fig. 2, the SE consists of a patch embedding layer, a lightweight Transformer backbone, and a Context Compressor (CC). The patch embedding layer divides satellite images of shape $n \times C \times H \times W$ into patches of size $p \times p$. The resulting patch embeddings have the shape $n \times (H/p \times W/p) \times C'$, where C' is the Transformer embedding dimension. The CC module contains a rearrangement step and a projection operation, which significantly reduces the context length and balances the token-length scale across temporal and spatial modalities. The projection operation, implemented as an MLP, changes the embedding shape to $n \times C^{\text{llm}}$.

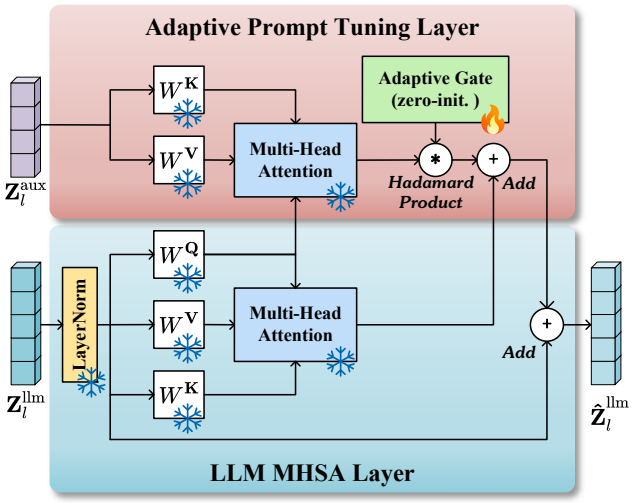

Figure 3: Adaptive Prompt Tuning Layer. The APT layer applies MHA to fuse $\hat{\mathbf{Z}}_l^{\text{llm}}$ and $\hat{\mathbf{Z}}_l^{\text{aux}}$. The fusion result is added to the output of the MHSA layer in the LLM through an Adaptive Gate.

Text Prompt Construction

Most existing PV power forecasting approaches overlook the potential of textual modalities. In practice, experts often refer to additional information such as the geographical context of PV plants, which can be effectively represented through natural language. Recent work (Yang, Li, and Zhang 2024) in satellite image classification has demonstrated the effectiveness of integrating text prompts for improved accuracy and generalization. Motivated by this, we design spatio-temporal text prompts based on structured metadata, including dataset descriptions, task descriptions, input statistics, forecast start time, and PV plant coordinates. These prompts guide the model to focus on the prediction objective and its contextual meaning. The dataset and task descriptions activate domain-specific knowledge in the LLM, while statistical and spatial features enhance grounding.

Each text prompt is constructed from three inputs: a structured template $\mathbf{X}^{\text{text}}$, the PV input sequence $\mathbf{X}^{\text{in}}$, and the plant coordinates $\mathbf{C}^{\text{in}}$. The template contains placeholders for forecast time, statistical summaries, and coordinates, which are dynamically filled for each sample. To prevent noise caused by high-precision values, all decimals are truncated to six digits. The LLM's pretrained word embedding layer $f_{\boldsymbol{\theta}_{we}}(\cdot)$ converts the prompt into embeddings:

$$\mathbf{Z}^{\text{text}} = f_{\boldsymbol{\theta}_{we}}(\text{TPC}(\mathbf{X}^{\text{text}}, \mathbf{X}^{\text{in}}, \mathbf{C}^{\text{in}})). \tag{12}$$

Adaptive Prompt Tuning

The Adaptive Prompt Tuning (APT) module fuses multimodal embeddings to generate the final PV power forecast. The APT module consists of a pretrained LLM, a layered projection module $f_{\boldsymbol{\theta}_{lp}}(\cdot)$, multiple APT layers, and a temporal decoder. In this work, we adopt GPT-2 (Radford et al. 2019) due to its efficiency and strong language prior. The input embeddings to APT include $\mathbf{Z}^{\text{ts}}$ and $\mathbf{Z}^{\text{sat}}$ from the MSE module, and $\mathbf{Z}^{\text{text}}$ from the TPC module. $\mathbf{Z}^{\text{text}}$ can be directly interpreted by the LLM, while $\mathbf{Z}^{\text{ts}}$, generated by the Temporal-Semantic Encoder, carries information aligned with the word prototype space. These two embeddings are concatenated along the sequence-length dimension and used as the input $\mathbf{Z}^{\text{llm}}$ to the pretrained LLM. Since the LLM cannot directly understand the spatio-temporal satellite embeddings $\mathbf{Z}^{\text{sat}}$, PV-LLM employs APT layers to integrate $\mathbf{Z}^{\text{sat}}$ into the model.

Adaptive Prompt Tuning Layer. Each APT layer is aligned with an MHSA layer in the LLM. The layered projection module takes satellite prompt embeddings $\mathbf{Z}^{\text{sat}}$ and produces the auxiliary information embeddings $\mathbf{Z}_l^{\text{aux}}$ for each APT layer, where l corresponds to the l-th MHSA layer of the LLM. The structure of the APT layer is illustrated in Fig. 3. Each APT layer contains an Adaptive Gate. During training, the parameters of the pretrained MHSA layer in the LLM are frozen, while the parameters of the Adaptive Gate are learnable. The MHSA layer receives $\mathbf{Z}_l^{\text{llm}}$ from the previous LLM layer as input, and the APT layer integrates $\mathbf{Z}_l^{\text{llm}}$ with the auxiliary information embeddings $\mathbf{Z}_l^{\text{aux}}$. The fusion is performed using MHA, where the query is $\mathbf{Z}_l^{\text{llm}}$, and the key and value are $\mathbf{Z}_l^{\text{aux}}$. The result of MHA is element-wise multiplied (Hadamard product) with the Adaptive Gate, and then added to the MHSA output of the LLM to produce the fused output $\hat{\mathbf{Z}}_l^{\text{llm}}$. The Adaptive Gate is initialized to zero, which allows PV-LLM to preserve more of the original LLM knowledge during early training, and gradually incorporate auxiliary information as training progresses. In summary, the APT module can be formalized as follows:

$$\mathbf{Z}^{\text{llm}} = \text{Concat}(\mathbf{Z}^{\text{text}}, \mathbf{Z}^{\text{ts}}), \tag{13}$$

$$f_{\boldsymbol{\theta}_{lp}}(\mathbf{Z}^{\text{sat}}) = \mathbf{Z}_l^{\text{aux}}, \mathbf{Z}_{l+1}^{\text{aux}}, ..., \mathbf{Z}_N^{\text{aux}}, \tag{14}$$

$$\begin{aligned}\hat{\mathbf{Z}}_l^{\text{llm}} &= \text{MHSA}(\text{LN}(\mathbf{Z}_l^{\text{llm}})) + \mathbf{Z}_l^{\text{llm}} \\ &\quad + \text{MHA}(\text{LN}(\mathbf{Z}_l^{\text{llm}}), \mathbf{Z}_l^{\text{aux}}, \mathbf{Z}_l^{\text{aux}}) \odot \mathbf{g},\end{aligned} \tag{15}$$

where $\odot$ denotes the Hadamard product operator.

Temporal Decoder. The temporal decoder $f_{\boldsymbol{\theta}_{td}}(\cdot)$ generates the final PV power prediction from the LLM output representations. The LLM produces the last hidden states $\hat{\mathbf{Z}}_N^{\text{llm}}$ (with shape $L \times D$), where L and D denote the sequence length and hidden dimension, respectively. These hidden states contain both time-series and textual information, concatenated along the sequence dimension. To recover modality-specific representations, we apply a split operation along L to obtain the fused time-series embeddings $\hat{\mathbf{Z}}^{\text{text}}$ (with shape $L^{\text{text}} \times D$) and the text embeddings $\hat{\mathbf{Z}}^{\text{ts}}$ (with shape $L^{\text{ts}} \times D$), with $L^{\text{text}} + L^{\text{ts}} = L$. The time-series part $\hat{\mathbf{Z}}^{\text{ts}}$ serves as the final multimodal fusion result of $\mathbf{Z}^{\text{ts}}$, $\mathbf{Z}^{\text{text}}$, and $\mathbf{Z}^{\text{sat}}$. The temporal decoder follows the structure used in the output head of PatchTST (Nie et al. 2023). It applies a flatten operation followed by an MLP to reconstruct the PV power time series from the patch-level embeddings, forming the final prediction. The decoding process is expressed as:

$$\hat{\mathbf{Z}}^{\text{text}}, \hat{\mathbf{Z}}^{\text{ts}} = \text{Split}(\hat{\mathbf{Z}}_N^{\text{llm}}), \tag{16}$$

$$\hat{\mathbf{X}}^{\text{out}} = f_{\boldsymbol{\theta}_{td}}(\hat{\mathbf{Z}}^{\text{ts}}). \tag{17}$$

Category	Methods	All (25210)		Easy (18014)		Hard (7196)	
		MAE(↓)	RMSE(↓)	MAE(↓)	RMSE(↓)	MAE(↓)	RMSE(↓)
Non-machine-learning	Persistence	0.06500	0.13909	0.04763	0.10279	0.10838	0.20319
	Mean	0.07632	0.12849	0.07674	0.12614	0.07528	0.13417
	Clear Sky	0.07347	0.15682	0.05589	0.12196	0.11748	0.22119
Unimodal (Transformer-based)	Informer (2021)	0.07973	0.13086	0.07952	0.12867	0.08025	0.13613
	Autoformer (2021)	0.07830	0.11702	0.07015	0.09876	0.10285	0.15505
	Crossformer (2023)	0.06599	0.11259	0.06201	0.10173	0.08440	0.14645
	PatchTST (2023)	0.06575	0.11755	0.06056	0.10192	0.08320	0.14783
	TimeXer (2024)	0.06620	0.11497	0.05806	0.09429	0.08651	0.15416
Unimodal (MLP-based)	LightTS (2022)	0.06474	0.11048	0.05724	0.09347	0.08324	0.14413
	Dlinear (2023)	0.07609	0.12310	0.06364	0.09762	0.10682	0.17035
	TimesNet (2023)	0.07551	0.13528	0.07120	0.12559	0.08653	0.15665
	WPMixer (2025)	0.06493	0.11670	0.05600	0.09636	0.08729	0.15642
Multimodal	CrossViViT (2023)	0.05789	0.11818	0.04891	0.09924	0.08007	0.15535
	Time-LLM (2024)	0.06150	0.11557	0.04986	0.08811	0.09064	0.16539
	MSP-RS (2024)	0.04320	<u>0.08889</u>	0.03859	<u>0.07868</u>	0.05475	0.11038
	FusionSF (2024)	<u>0.04038</u>	0.08893	<u>0.03658</u>	0.07939	0.04989	<u>0.10922</u>
	SolarFusionNet (2025)	0.04100	0.09051	0.03776	0.08174	<u>0.04913</u>	0.10944
	PV-LLM (Ours)	**0.03962**	**0.08449**	**0.03615**	**0.07620**	**0.04829**	**0.10234**

Table 1: Quantitative results of model performance on MMSP(S) dataset across "All", "Easy", and "Hard" scenarios. The best results are highlighted in **bold** and the second best results are <u>underlined</u>. The numbers in parentheses in the first row denote the sample counts for each scenario.

Experimental Results

Experimental Setup

Dataset. In this work, we conduct experiments on the MMSP(S) dataset (Ma et al. 2024), which includes PV power data, corresponding satellite image sequences, and NWP data from multiple real-world photovoltaic plants. The dataset spans from January 2021 to June 2022 with a temporal resolution of one hour. All data are normalized using min-max scaling. Details of the modalities are as follows:

- **PV Power Data:** Collected from ten photovoltaic power stations in Shandong Province, China, providing ground-truth measurements of solar power generation.
- **Satellite Image Sequence Data:** Acquired from the Advanced Himawari Imagers onboard Japan's Himawari-8/9 satellites, providing high-resolution cloud dynamics crucial for spatio-temporal understanding.
- **NWP Data:** Obtained from the European Centre for Medium-Range Weather Forecasts, offering single-point meteorological predictions at the grid point closest to each PV plant.

Baselines. We compare our proposed method against four categories of baselines, which correspond to the group indices in Table 1: (1) **Non-machine-learning methods** including Persistence, Mean, and Clear Sky methods (Ineichen 2016); (2) **Transformer-based unimodal methods** such as Informer (Zhou et al. 2021), Autoformer (Wu et al. 2021), Crossformer (Zhang and Yan 2023), PatchTST (Nie et al. 2023), and TimeXer (Wang et al. 2024b); (3) **MLP-based unimodal methods** including LightTS (Zhang et al. 2022), DLinear (Zeng et al. 2023), TimesNet (Wu et al. 2023), and WPMixer (Murad, Aktukmak, and Yilmaz 2025); (4) **Multimodal methods** that integrate heterogeneous inputs, including CrossViViT (Boussif et al. 2023), Time-LLM (Jin et al. 2024), MSP-RS (Kim and Suh 2024), FusionSF (Ma et al. 2024), and SolarFusionNet (Jing et al. 2025). The modalities used by each baseline follow their original designs.

Training Details. PV-LLM is trained using the AdamW optimizer (Loshchilov and Hutter 2019) with a weight decay of 0.05, a peak learning rate of 3×10^{-5}, and cosine learning rate scheduling. All experiments are conducted on a single NVIDIA RTX A6000 GPU with a batch size of 64. Training for 100 epochs takes approximately one day.

Evaluation Metrics. Following common practice in PV forecasting (Ma et al. 2024; Jing et al. 2025), we evaluate model performance using two standard metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Forecasting Scenarios. To better understand model performance under varying difficulty, we categorize the test set of MMSP(S) into two forecasting scenarios based on the difficulty level. Specifically, we compute the ratio r between the area under the curve (AUC) of the target PV power sequence and that of the input historical sequence:

$$r = \left| \ln \frac{\mathrm{AUC}_{\mathbf{X}^{\text{out}}}}{\mathrm{AUC}_{\mathbf{X}^{\text{in}}}} \right| \tag{18}$$

Samples satisfying $r < \left|\ln \frac{2}{3}\right|$ are classified as "easy" scenarios, while the remaining ones are considered "hard" scenarios. A larger value of r indicates greater deviation from historical patterns, complicating the PV forecasting task.

Methods	MAE(All)	RMSE(All)
w/o TPC	0.04255	0.08954
w/o APT	0.04405	0.09021
w/o Semantic Adapter	0.04750	0.09330
w/o Context Compressor	0.04376	0.08893
w/o LLM	0.05500	0.10781
w/o LLM Pretraining	0.04610	0.09446
w/o LLM Freezing	0.04209	0.09109
w/o Gate Zero Initialization	0.04202	0.08772
PV-LLM (Full Model)	**0.03962**	**0.08449**

Table 2: Ablation Study on Key Modules

Quantitative Results

Table 1 presents a comprehensive comparison between PV-LLM and four categories of baseline methods. The best and second-best results are highlighted in bold and underlined, respectively. From these results, we derive the following key observations:

- **Effectiveness of Multimodal Methods:** Integrating multimodal data significantly improves PV power forecasting. While unimodal methods perform similarly to non-machine-learning baselines, multimodal methods achieve superior performance across both metrics.
- **Performance Discrepancy Across Scenarios:** All methods exhibit degraded performance under "hard" scenarios compared to "easy" ones, with the largest gap observed in the Persistence baseline. This validates the rationality of our scenario classification. Interestingly, under "easy" scenarios, many methods underperform compared to Persistence—likely due to their focus on generalization at the expense of simpler patterns.
- **Superiority of PV-LLM:** PV-LLM consistently outperforms all baselines across both scenarios. The advantage is particularly evident in RMSE, indicating that PV-LLM is more robust to extreme variations by leveraging complementary information from multiple modalities. We additionally performed the Wilcoxon signed-rank test using five independent runs, and the results show that the median MAE and RMSE of PV-LLM are significantly lower than those of FusionSF ($p < 0.05$).

Ablation Study

We conduct extensive ablation studies to assess the contribution of individual components within PV-LLM, including the LLM backbone, text prompt components, and positional encoding strategies in the Satellite Image Sequence Encoder. The main findings are summarized below:

Methods	MAE(All)	RMSE(All)
PV-LLM (Sinusoidal)	0.04070	0.08625
PV-LLM (Learned)	0.04152	0.08468
PV-LLM (RoPE)	**0.03962**	**0.08449**

Table 3: Ablation Study on Positional Encoding

Methods	MAE(All)	RMSE(All)
w/o Dataset Descriptions	0.04027	0.08601
w/o Task Descriptions	0.04027	0.08530
w/o Input Statistics	0.04032	0.08611
w/o Forecast Start Time	0.04314	0.08671
w/o Plant Coordinates	0.04103	0.08603
PV-LLM (Full Model)	**0.03962**	**0.08449**

Table 4: Ablation Study on Text Prompt

- **Impact of TPC, APT, SA, and CC:** As shown in Table 2, excluding any of the Text Prompt Construction (TPC), Adaptive Prompt Tuning (APT), Semantic Adapter (SA), or Context Compressor (CC) modules leads to performance degradation. Among them, removing SA results in the largest drop, highlighting the critical role of modality alignment. For the APT ablation, we directly concatenate $\mathbf{Z}^{\text{sat}}$ and $\mathbf{Z}^{\text{llm}}$ and input them into the pretrained LLM.
- **Role of the Pretrained LLM:** Table 2 further explores variants involving the LLM: "w/o LLM" completely removes it; "w/o LLM Freezing" enables parameter updates during training; "w/o LLM Pretraining" initializes from scratch without freezing; and "w/o Gate Zero Initialization" denotes randomly initializing the adaptive gate in APT with values from a uniform distribution over $(0, 1)$. All these variants degrade performance, underscoring the value of pretrained knowledge and the necessity of proper parameter constraints.
- **Effectiveness of RoPE:** Table 3 shows that RoPE outperforms sinusoidal and learnable encodings, indicating superior encoding of spatio-temporal patterns in satellite image sequences for PV forecasting.
- **Effectiveness of Text Prompt Components:** As reported in Table 4, all components of the text prompt contribute positively to performance. Notably, the inclusion of forecast start time and plant coordinates yields the most significant gains, outperforming other elements such as task descriptions, dataset annotations, and statistical summaries. This highlights the value of explicitly incorporating spatio-temporal cues in prompt design.

Conclusion

This work presents PV-LLM, a novel satellite-text-prompted framework for photovoltaic power forecasting that integrates multimodal inputs through a pretrained large language model. Unlike prior approaches that focus primarily on numerical features, PV-LLM leverages spatio-temporal text prompts to incorporate high-level semantic information relevant to the forecasting task, including plant coordinates, forecast start time, and input statistics. Extensive experiments demonstrate the effectiveness of introducing the text modality and highlight the potential of LLMs for modeling multimodal dependencies in PV power forecasting.

Acknowledgments

This work was supported in part by National Science and Technology Major Project (No. 2022ZD0119503), National Natural Science Foundation of China under Grants 62272130, 62376072, 62202122, and 62502120, Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011949), and Shenzhen Science and Technology Program under Grants KCXFZ20230731094905010, KCXFZ20240903093006009, SYSPG20241211173609009, GXWD20231130110308001, and JCYJ20250604145617023.

References

Alfadda, A.; Adhikari, R.; Kuzlu, M.; and Rahman, S. 2017. Hour-ahead solar PV power forecasting using SVR based approach. In *2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference*, 1–5.

Babaei, S.; Zhao, C.; and Fan, L. 2019. A data-driven model of virtual power plants in day-ahead unit commitment. *IEEE Transactions on Power Systems*, 34(6): 5125–5135.

Boussif, O.; Boukachab, G.; Assouline, D.; Massaroli, S.; Yuan, T.; Benabbou, L.; and Bengio, Y. 2023. Improving day-ahead solar irradiance time series forecasting by leveraging spatio-temporal context. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2342–2367.

Elizabeth Michael, N.; Mishra, M.; Hasan, S.; and Al-Durra, A. 2022. Short-term solar power predicting model based on multi-step CNN stacked LSTM technique. *Energies*, 15(6): 2150.

Hu, Y.; Li, Q.; Zhang, D.; Yan, J.; and Chen, Y. 2025. Context-Alignment: Activating and Enhancing LLMs Capabilities in Time Series. In *The Thirteenth International Conference on Learning Representations*.

Ineichen, P. 2016. Validation of models that estimate the clear sky global and beam solar irradiance. *Solar Energy*, 132: 332–344.

International Renewable Energy Agency (IRENA). 2025. Renewable Capacity Statistics 2025. Technical report, International Renewable Energy Agency, Abu Dhabi.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-f.; Pan, S.; et al. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Jing, T.; Chen, S.; Navarro-Alarcon, D.; Chu, Y.; and Li, M. 2025. SolarFusionNet: Enhanced Solar Irradiance Forecasting via Automated Multi-Modal Feature Selection and Cross-Modal Fusion. *IEEE Transactions on Sustainable Energy*, 16(2): 761–773.

Kim, B.; and Suh, D. 2024. Solar PV Generation Prediction Based on Multisource Data Using ROI and Surrounding Area. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–11.

Li, Y.; Su, Y.; and Shu, L. 2014. An ARMAX model for forecasting the power output of a grid connected photovoltaic system. *Renewable Energy*, 66: 78–89.

Lorenz, E.; Scheidsteger, T.; Hurka, J.; Heinemann, D.; and Kurz, C. 2011. Regional PV power prediction for improved grid integration. *Progress in Photovoltaics: Research and Applications*, 19(7): 757–771.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *The seventh International Conference on Learning Representations*.

Ma, Z.; Wang, W.; Zhou, T.; Chen, C.; Peng, B.; Sun, L.; and Jin, R. 2024. FusionSF: Fuse Heterogeneous Modalities in a Vector Quantized Framework for Robust Solar Power Forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5532–5543.

Makarov, Y. V.; Loutan, C.; Ma, J.; and De Mello, P. 2009. Operational impacts of wind generation on California power systems. *IEEE Transactions on Power Systems*, 24(2): 1039–1050.

Mayer, M. J.; and Gróf, G. 2021. Extensive comparison of physical models for photovoltaic power forecasting. *Applied Energy*, 283: 116239.

Murad, M. M. N.; Aktukmak, M.; and Yilmaz, Y. 2025. Wpmixer: Efficient multi-resolution mixing for long-term time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19581–19588.

Nguyen, A.; Velay, M.; Schoene, J.; Zheglov, V.; Kurtz, B.; Murray, K.; Torre, B.; and Kleissl, J. 2016. High PV penetration impacts on five local distribution networks using high resolution solar resource assessment with sky imager and quasi-steady state distribution system simulations. *Solar Energy*, 132: 221–235.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 27730–27744.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Saint-Drenan, Y.-M.; Bofinger, S.; Fritz, R.; Vogt, S.; Good, G.; and Dobschinski, J. 2015. An empirical approach to parameterizing photovoltaic plants for power forecasting and simulation. *Solar Energy*, 120: 479–493.

Sarmas, E.; Dimitropoulos, N.; Marinakis, V.; Mylona, Z.; and Doukas, H. 2022. Transfer learning strategies for solar power forecasting under data scarcity. *Scientific Reports*, 12(1): 14643.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tan, M.; Merrill, M.; Gupta, V.; Althoff, T.; and Hartvigsen, T. 2024. Are language models actually useful for time series

forecasting? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 60162–60191.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Wang, C.; Qi, Q.; Wang, J.; Sun, H.; Zhuang, Z.; Wu, J.; Zhang, L.; and Liao, J. 2025. ChatTime: A Unified Multimodal Time Series Foundation Model Bridging Numerical and Textual Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, X.; Feng, M.; Qiu, J.; Gu, J.; and Zhao, J. 2024a. From news to forecast: integrating event analysis in LLM-based time series forecasting with reflection. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 58118–58153.

Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; and Long, M. 2024b. TimeXer: empowering transformers for time series forecasting with exogenous variables. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 469–498.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 22419–22430.

Yang, C.; Li, Z.; and Zhang, L. 2024. Bootstrapping interactive image–text alignment for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11121–11128.

Zhang, T.; Zhang, Y.; Cao, W.; Bian, J.; Yi, X.; Zheng, S.; and Li, J. 2022. Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures. *CoRR*, abs/2207.01186.

Zhang, X.; Feng, S.; Ma, J.; Lin, H.; Li, X.; Ye, Y.; Li, F.; and Ong, Y. S. 2024. Frnet: Frequency-based rotation network for long-term time series forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3586–3597.

Zhang, Y.; and Yan, J. 2023. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *International Conference on Learning Representations*.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11106–11115.

Zhou, T.; Niu, P.; Wang, X.; Sun, L.; and Jin, R. 2023. One fits all: power general time series analysis by pretrained LM. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 43322–43355.