Research on Engineering Structures & Materials
www.jresm.org
MIM Research Group

Research Article

Distributed photovoltaic operation characteristics based on static grid topology and multi-source heterogeneous data fusion: A potential edge deployment scheme

Ge Xiao [*,a], Hanmei Ma [b], Xinyue Guo [c], Jimeng Song [d]

State Grid Information & Telecommunication Center (Big Data Center), Beijing, China

Article Info

Article History:

Received 09 Sep 2026

Accepted 12 Oct 2026

Keywords:

Distributed photovoltaic power prediction;
Multi-source heterogeneous data;
Graph attention networks;
Spatial-temporal modeling;
Knowledge distillation;
Edge deployment;
Multi-time interval forecasting

Abstract

In order to address the challenges of spatiotemporal coupling and edge deployment difficulties in distributed photovoltaic power generation prediction, a multi granularity graph gated long short-term memory network coupled with an output knowledge distillation architecture is proposed. Firstly, utilizing dynamic collaboration mechanisms to unify multi-source heterogeneous observations and grid structure priors. Next, the three-branch graph encoder composed of the original nodes and the multi hop graph attention network extracts spatial context through the Softmax gating algorithm. Then, these spatial features are injected into the time step through conditional bias for joint multi time scale prediction. Finally, a lightweight student network was constructed using knowledge distillation for potential edge deployment. The results indicate that the model achieved an average absolute error of 25.31 MW and an average absolute percentage error of 1.48% within a prediction interval of 1 hour. During the merged time interval, the global average absolute percentage error decreased to 2.64%, which is 42.39% less than the traditional gated cyclic unit model. In addition, the lightweight student model successfully compressed the number of parameters from 319.7K to 44.9K, achieving a compression rate of 85.96%. Although knowledge distillation has brought about a small overall improvement in comprehensive error, its statistical significance is manifested as marginal benefits, and it exposes a certain trade-off of performance degradation in the 24-hour long-term prediction field. This framework achieves deep collaboration between data fusion, grid aware modeling, and lightweight compression, providing engineering reference for potential edge deployment in resource constrained scenarios.

© 2026 MIM Research Group. All rights reserved.

1. Introduction

In the context of a high proportion of renewable energy grid connection, the installed capacity of distributed photovoltaic continues to grow, and the volatility, randomness and spatio-temporal coupling characteristics of new energy output become increasingly prominent [1-3]. Compared with traditional centralized power sources, distributed photovoltaic power is not only driven by meteorological factors such as irradiation, temperature, humidity, wind speed and rainfall, but also jointly affected by geographical differences of grid-connected nodes, resource endowments and grid connection relationships [4-5]. Specifically, high penetration rates of distributed photovoltaics can cause reverse power flow and severe feeder voltage fluctuations. In addition, the intermittency of photovoltaics has increased the uncertainty of predictions, leading to increased procurement costs for rotating backup and operational pressure on voltage regulating equipment. Improving prediction accuracy directly alleviates these risks by optimizing backup allocation and reducing

*Corresponding author: xiaoge202608@163.com
[a]orcid.org/0009-0000-3420-267X; [b]orcid.org/0009-0002-1605-7274; [c]orcid.org/0009-0004-5227-2216;
[d]orcid.org/0009-0009-4779-0509
DOI: http://dx.doi.org/10.17515/resm2026-2322vk0909rs

dynamic climbing burden. When the prediction object expands from a single site to a regional-level grid-connected unit aggregated from multiple nodes, it is often difficult to fully characterize key features such as cross-regional weather propagation, grid coupling, and power linkage by relying only on local historical sequences for modeling. Therefore, constructing prediction models that can uniformly integrate multi-source heterogeneous observations and grid structure priors for distributed photovoltaic scenarios becomes an important research issue in new energy dispatching, consumption analysis and operation perception [6].

From the perspective of application requirements, distributed photovoltaic power prediction is not a single time interval problem. The prediction for the next one hour mainly serves short-term scheduling and real-time control, the prediction for the next six hours supports intraday rolling optimization and standby adjustment, and the prediction for the next 24 hours is closely related to day-ahead planning, energy storage coordination and risk assessment. Different prediction intervals correspond to different sources of uncertainty and decision-making objectives, but all rely on the same set of historical power, meteorological information, temporal priors, and spatial coupling relationships. Therefore, conducting multi-time distributed photovoltaic power prediction under a unified framework not only helps to improve the consistency of model expression, but also helps to enhance the availability of prediction results on the dispatch side and the operation side [7].

Among the existing photovoltaic power prediction methods, hybrid machine learning and deep learning models provide an effective way to overcome the sharp fluctuations and nonlinear characteristics of the data. Souhe et al. proposed a prediction model based on parallel hybrid Support Vector Machine-Gated Recurrent Unit (SVM-GRU) and ant colony optimization algorithm to address the issue of data variability in smart grid photovoltaic prediction. By combining the strengths of both algorithms, the method achieved a standard correlation coefficient of 0.9986[8]. Wu S et al. used an adaptive noise fully integrated empirical mode decomposition algorithm to break down the data into several subsequences to mitigate the interference of large fluctuations in photovoltaic power. The Convolutional Neural Network-Gated Recurrent Unit (CNN-GRU) model was constructed. This approach effectively overcame the limitations of a single network in capturing historical features [9]. To further enhance the supply-demand balance capacity of energy management systems, Boucetta L N et al. proposed a hybrid deep learning model that integrates GRU and attention mechanisms. The results showed that the model had a decision coefficient of up to 0.992 in wind power advance prediction, significantly capturing the inherent volatility of the sequence [10]. In addition, Ma Y et al. proposed a two-stage prediction method based on the optimized Transformer architecture for the nonlinear and periodic characteristics of photovoltaic power generation, introducing the ProbSparse attention mechanism and weighted sequence decomposition module. Compared with the traditional Transformer, the Root Mean Square Error (RMSE) of this method was reduced by 31.23% [11]. Zarghami M et al. proposed a novel deep learning model for evaluating the flexibility of power systems - the spatio-temporal hybrid convolutional converter network. In power and load predictions for the next 24 hours and 168 hours, the network reduced the prediction error to 2.5% and the maximum Mean Absolute Error (MAE) to 10.1% [12]. Heng S et al. proposed a method based on multi-task learning and Gru-attention networks for the planning and scheduling problem of multi-energy load systems. The results showed that the method reduced the average Mean Absolute Percentage Error (MAPE) of predictions by 43% [13].

In addition, in recent years, spatiotemporal graph neural networks (ST-GNN) have made breakthrough progress in the field of power system prediction, with representative frontier models emerging such as adaptive graph convolutional recurrent networks (AGCRN), attention mechanism spatiotemporal graph neural networks (ASTGNN), spatiotemporal graph convolutional networks (STGCN), and time series-oriented Transformer variants (such as PatchTST). However, applying it to distributed photovoltaic scenarios faces computational and theoretical constraints. Given that the main physical connectivity between regional substations remains relatively static during most scheduling cycles, the study adopts the assumption of static graphs. The theoretical boundary lies in sacrificing absolute sensitivity to instantaneous topology fine-tuning in exchange for stability

and computational robustness of spatial coupling expression, effectively avoiding edge overfitting and computational overhead caused by frequent reconstruction of graph structures.

In summary, previous research has made significant progress in distributed photovoltaic power prediction and multi energy collaborative analysis based on data-driven and hybrid temporal networks. However, there are still many limitations in the face of regional level grid connected scenarios with a high proportion of new energy access. One is that multi-source heterogeneous data is usually aggregated at a single scale, lacking a unified heterogeneous dynamic observation and power grid topology fusion mechanism [14]. The second issue is that existing graph time series models often come with high parameter sizes and computational costs. When deployed on the edge or in real-time perception scenarios with limited resources, there is a lack of deep collaboration solutions that balance prediction accuracy and model light weighting [15]. To this end, research is conducted on distributed photovoltaic power prediction and grid perception modeling methods driven by multi-source heterogeneous data, and a Multi-granularity Graph Gated Long Short-Term Memory (MGG-LSTM) network is proposed. This model is based on a static graph composed of real grid connection relationships, using original node branches, 1-hop GATv2 branches, and 2-hop GATv2 branches to extract multi granularity graph context, and implementing graph side fusion using Softmax gating. Subsequently, the obtained graph context will be used simultaneously for the initial state construction and time step bias injection of the conditional LSTM, and then combined with the shared output head, time-division output head, and persistent skip branch to complete multi temporal photovoltaic power prediction for the next 1 hour, 6 hours, and 24 hours. This design can integrate multi-source heterogeneous dynamic observations and grid perception priors in a unified framework, thus balancing spatial structure expression, time-dependent modeling, and multi temporal output requirements. On this basis, further research is conducted to construct a lightweight student model, and the feasibility of compressing and deploying the model is analyzed using the output knowledge distillation strategy. The main contributions are as follows:

- A multi-source heterogeneous data modeling framework for distributed photovoltaic grid-connected node prediction is constructed, and historical power, supplementary dynamic variables, meteorological statistics, time features, differential features, static node attributes, and static edge attributes are uniformly organized as graph time series inputs.
- A grid-aware MGG-LSTM model is proposed, effectively injecting multi-granularity graph context into the time series prediction process through three-branch graph coding, Softmax gating fusion, and conditional LSTM design.
- Experimental analyses are conducted on future 1-hour, 6-hour, and 24-hour multi-time interval prediction tasks to verify the positive effects of multi-source heterogeneous feature fusion and grid-aware modeling on the performance of distributed photovoltaic power prediction.
- To meet the requirements of lightweight deployment, the study constructs the Student Baseline model and output distillation model, and analyzes the role of knowledge distillation in parameter compression and maintaining the accuracy of multi-time interval prediction.

2. Data Modeling and Multi-Source Heterogeneous Input Construction

2.1 Dataset and Scenario Description

The study uses existing data related to the power system in a certain region as a case to construct a multi-node grid-connected scenario for distributed photovoltaic power prediction. The original data covers the period from 2019 to early 2021, including hourly organized dynamic observations, node static attributes, and grid connection relationships. Given that the original data is presented in the form of regional-level nodes, the study interprets them as distributed photovoltaic aggregation units or regional-level grid-connected nodes to analyze the multi-node distributed photovoltaic power prediction problem without changing the data source. The dataset is sourced from the SCADA system operated by a provincial power grid, covering an area of approximately 5000 square kilometers in the central eastern part of the province. The 27 defined nodes physically represent 27 regional substations, with installed capacities ranging from 15MW to 120MW. 92 directed edges represent the actual physical transmission lines connecting these substations and

their flow directions. The physical basis for interpreting these regional level nodes as distributed photovoltaic aggregation units is that in actual power grid scheduling, scattered roofs and small distributed photovoltaics are often difficult to be directly observed by the dispatch center one by one, but are collected through the distribution network to the upper-level substation nodes. Therefore, treating the net photovoltaic output of the substation node as a distributed aggregation unit not only conforms to the real physical connectivity of the power grid topology, but also to the actual data collection and operation perception mode of the scheduling end. Regarding the modeling of directed edges, although the actual power flow is essentially bidirectional, the directed graph edges in the study represent the main transmission line directions and energy transmission paths derived from the historical SCADA power flow direction, effectively capturing the dominant spatial propagation patterns. In addition, due to the confidentiality agreement governing the proprietary provincial SCADA dataset, the research has not yet been validated on the open-source standard benchmark network. The prediction target is the photovoltaic power output of each node at different time intervals in the future, and the static graph structure is used to describe the connection relationship between the nodes and the grid perception prior.

Table 1 summarizes the input information used. Overall, distributed photovoltaic power prediction does not rely solely on a single time series, but is composed of data from multiple sources, different update frequencies, and different structural forms. Among them, historical photovoltaic power, supplementary dynamic variables, and meteorological statistics belong to hourly dynamic time series data. The time features belong to the rule prior constructed based on timestamps. Node attributes are used to characterize the spatial location and resource endowment of grid-connected units. Edge attributes are used to describe the connection capabilities and physical constraints between nodes. There are obvious differences in the physical meaning and representation of the above information, which form the basis of the multi-source heterogeneous input that the study focuses on.

2.2 Data Preprocessing

To ensure consistency and reproducibility in the modeling process, the study completes data cleaning, feature construction, and sample generation following the actual process in the preprocessing script. The core steps are as follows:

- Time division: In order to overcome the severe seasonal distribution shift caused by traditional chronological partitioning, this study redesigned and corrected the data partitioning strategy. We mainly adopted a seasonal stratified rolling cross validation/mixed sampling scheme to ensure that the training set, validation set, and test set can evenly cover the complete meteorological and photovoltaic output dynamic characteristics of all seasons, eliminating fundamental seasonal mismatches.
- Outlier correction: The study performs non-negative truncation on the dynamic power sequence to prevent negative observations from interfering with subsequent training.
- Meteorological statistics mapping: The study extracts statistics such as rainfall, air pressure, global irradiance, air temperature, dew point temperature, humidity, wind direction, gusts, and wind speed from the original field as node-level exogenous drive inputs.
- Variation Feature Construction: The study further constructs differential features such as irradiance, rainfall, gusts, humidity, air pressure, air temperature, wind speed, and dew point temperature to enhance the model's perception of short-term weather changes.
- Baseline-assisted features: Construct by node, based on the training set statistics, the study organizes the photovoltaic baseline items by month and hour, and calculates the deviation of the current power from the baseline to supplement the typical intraday pattern information of the node.
- Feature scaling: Robust scaling strategies are adopted for dynamic inputs, node attributes, and edge attributes to mitigate the influence of abnormal samples and dimensional differences on model training. The prediction target is processed with an independent scaler.
- Graph structure construction: With grid-connected nodes as graph nodes and grid connection relationships as graph edges, the study constructs a static graph containing 27 nodes and 92 directed edges, and saves node mappings and edge attribute tensors.

- Sample generation: The study constructs the input sequence using a historical window of length 24 and generate multi-time interval prediction targets for the next 1 hours, 6 hours, and 24 hours.

Table 1. Overview of multi-source heterogeneous input data

Category	Feature
Static image G	Node: Static attributes such as latitude and longitude, photovoltaic potential, etc.; Edge: source node, target node, transmission capacity, line efficiency, line length, energy load type Historical power: photovoltaic power PV and its historical target sequence
Dynamic input X	Meteorological statistics: mean and standard deviation of rainfall, atmospheric pressure, global radiation, temperature, dew point temperature, relative humidity, wind direction, gusts, and wind speed Time characteristics: year, month, day, hour, week, week sequence number, season, holidays, etc. Change characteristics: baseline difference of photovoltaics, as well as differential characteristics such as irradiation, rainfall, gusts, humidity, air pressure, temperature, wind speed, dew point temperature, etc.

After the above processing, the original multi-source heterogeneous observations are uniformly organized into two types of model inputs: One is a dynamic sequence tensor arranged in chronological order, and the other is a static graph structure composed of node and edge attributes. The former corresponds to observations and priors that change over time, while the latter corresponds to structural constraints that do not change frequently over time, providing the basis for subsequent grid perception modeling.

2.3 Multi-Source Heterogeneous Fusion Mechanism

The multi-source heterogeneous data-driven approach emphasized in the study is mainly based on the collaborative modeling approach that combines graph-side static priors with sequence-side dynamic observations according to its structural characteristics. Specifically, node attributes and edge attributes are first encoded on the static graph as grid-aware graph context to express the structural associations and physical constraints between nodes. Historical power, supplementary dynamic variables, meteorological statistics, time features, and differential features are retained as dynamic sequence inputs to describe the operating state and exogenous drivers of nodes over time.

In subsequent models, the static graph-side information is mainly extracted from the node-level graph context through multi-granularity graph encoding and further injected into the time series modeling module. Among them, the graph-side encoding is responsible for providing stable structural prior and spatial coupling information, and the sequence-side encoding is responsible for expressing short-term dynamic changes. The two work together under a unified framework complete multi-time interval distributed photovoltaic power prediction. Based on this idea, the specific structure and modeling process of the MGG-LSTM will be further presented in Section 3.

3. Methods

3.1 Problem Definition and Overall Framework

Based on the multi-source heterogeneous input constructed in Chapter 2, this study aims to establish a multi-time interval photovoltaic power prediction model MGG-LSTM for regional-level distributed photovoltaic scenarios. Set the static grid diagram $G = (V, E)$, where the node set V represents the region-level grid-connected nodes and the edge set E represents the inter-regional connection relationship. Based on the preprocessing results, the static graph used in the study contains 27 nodes and 92 directed edges. For any node$i$, $X_i = \{x_{i,t-23}, x_{i,t-22}, \dots, x_{i,t}\}$, $x_{i,\tau} \in \mathbb{R}^d$,a historical dynamic sequence of length 24 was given. Here, d represents the dimension of the dynamic feature at each time step. The aim of the study is to simultaneously predict the

photovoltaic power output of the node for the next 1 h, 6 h, and 24 h, namely $Y_i = \{y_{i,t+1}, y_{i,t+6}, y_{i,t+24}\}$.

The overall structure of the MGG-LSTM consists of three parts: static graph coding, conditional time series modeling, and multi-time interval prediction. The model first encodes node attributes and edge attributes across the entire graph to obtain the graph context representation of each node. Then, the model extracts the corresponding graph context based on the node number corresponding to the sample and injects it into the temporal branch to constrain the modeling process of the historical dynamic sequence. Finally, by combining the shared prediction head, time-interval linear head, and continuous skip connection branch, the model outputs the estimated photovoltaic power at multiple predicted time intervals in the future. Unlike traditional models that rely solely on single-node time series, MGG-LSTM utilizes both the spatial coupling relationship between regions and the dynamic evolution law of nodes themselves to form a unified spatio-temporal prediction framework.

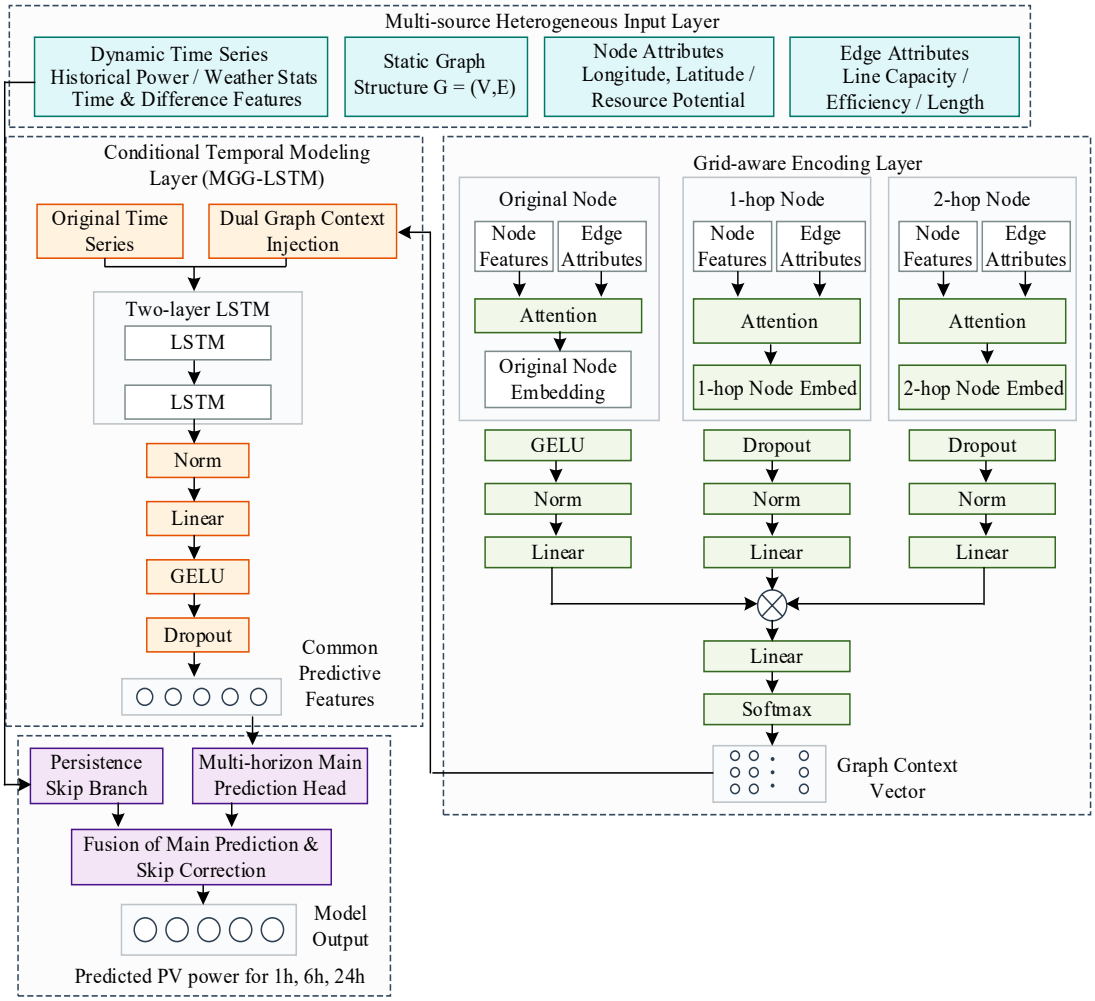


Fig. 1. Overall structure of the MGG-LSTM model

In terms of the implementation mechanism, MGG-LSTM has the following three characteristics. First, the model uses a full-graph GATv2 encoder instead of batch-to-batch subgraph extraction to learn a stable spatial context directly on a fixed static graph. Second, the model retains the original static node branches, 1-hop GAT branches, and 2-hop GAT branches simultaneously, and completes multi-granularity spatial information fusion through the gating mechanism. Third, the graph context is not only used to construct the initial latent state of the LSTM, but also to inject the dynamic sequence moment by moment through conditional bias, thereby enhancing the constraint ability of the spatial prior on time series prediction [16-18]. The overall topological architecture and data flow of the MGG-LSTM network are shown in Figure 1.

As shown in Figure 1, first, the multi-source heterogeneous input layer separates and characterizes the hourly dynamic sequence from the static topological prior. In the grid-aware coding layer on the right, the original nodes, 1-hop and 2-hop spatial branches capture multi-granularity spatial coupling relationships through the GATv2 attention mechanism, and are condensed into graph context vectors reflecting the physical constraints of the grid by Softmax gated weighted. This vector is then used as a conditional excitation to co-guide the initial state construction and stepwise conditional bias correction of the two-layer LSTM in the conditional time series modeling layer, generating highly cohesive common prediction features. Finally, at the bottom output layer, the model performs parallel fusion correction of deep spatiotemporal features extracted from the multi-time distance master prediction head and continuous skip connection branches directly extracted from historical observations to co-output photovoltaic power estimates for the next 1 h, 6 h, and 24 h.

3.2 Multi-Granularity Graph Coding and Spatial Context Modeling

Region-level distributed photovoltaic power shows obvious spatial correlation. On the one hand, adjacent regions are often jointly driven by similar weather processes. On the other hand, structural information such as grid topology, line capacity and transmission efficiency affect the coupling strength between regions [19-21]. Therefore, it is difficult to fully characterize the correlations between different regions starting from the single-node historical sequence alone, and graph structure priors must be introduced in the modeling. For this purpose, MGG-LSTM builds a three-branch graph encoder on a static graph G. Let the node static attribute matrix be $X^{(g)}$, then the first branch linearly encodes the original node features, resulting in Equation (1).

$$H^{(r)} = LN(\phi(X^{(g)}W_r + b_r)) \tag{1}$$

In Equation (1), $\phi(\cdot)$ represents the GELU activation function, $LN(\cdot)$ represents Layer Normalization (LN), and W_r and b_r are the learnable weight matrix and bias vector of the branch. This branch directly retains the static attribute information of the node itself and can be regarded as the original static node branch. The second branch uses GATv2 convolution with edge attributes to extract the first-order neighborhood space context. Let the edge index be E and the edge attribute be $A^{(e)}$, then the 1-hop branch is denoted as Equation (2).

$$H^{(1)} = LN(Drop(GATv2_1(X^{(g)}, E, A^{(e)}))) \tag{2}$$

In Equation (2), $GATv2_1(\cdot)$ represents the first layer GATv2 convolution operation, $Drop(\cdot)$ represents the random inactivation operation to prevent overfitting. Because edge_dim is explicitly set in the implementation, edge attributes such as line capacity, efficiency, length, and energy type are directly involved in the attention calculation, so that the spatial encoding can reflect not only the topological connectivity but also the physical characteristics of the grid connection. Based on this, the third branch continues to propagate a second $H^{(1)}$GATv2 on the same full graph, forming a 2-hop spatial representation under a larger receptive field, as shown in Equation (3).

$$H^{(2)} = LN(Drop(GATv2_2(H^{(1)}, E, A^{(e)}))) \tag{3}$$

In Equation (3), $GATv2_2(\cdot)$represents the second layer GATv2 convolution operation. It should be noted that the 2-hop branches here do not construct new independent subgraphs separately, but rather the second-order neighborhood representation obtained by continuing propagation based on the first-order graph state on the same static graph, thus expanding the spatial receptive field while maintaining implementation stability.

For any node i, let the representations corresponding to the three branches be, and $h_i^{(r)}$,$h_i^{(1)}$, and $h_i^{(2)}$, then concatenate the three first, as shown in Equation (4).

$$u_i = h_i^{(r)} \parallel h_i^{(1)} \parallel h_i^{(2)} \tag{4}$$

Then the fusion weights are generated through the linear layer and normalized using Softmax, as shown in Equation (5).

$$\alpha_i = \text{softmax}(W_f u_i + b_f) \tag{5}$$

In Equation (5), $\alpha_i = [\alpha_i^{(r)}, \alpha_i^{(1)}, \alpha_i^{(2)}]$. W_f and b_f are the learnable weight matrix of the gated fusion layer with the bias vector. In order to overcome the shortcomings of traditional Softmax gating, which only outputs global scalar multipliers and severely limits feature representation capabilities, research has upgraded this mechanism to a channel/dimension attention mechanism. By calculating weights for each feature dimension of the concatenated vector, the network can adaptively allocate the representation capacity of multi granularity graph context for different feature channels. The multi-granularity graph fusion results are shown in Equation (6).

$$\tilde{h}_i = \alpha_i^{(r)} h_i^{(r)} \parallel \alpha_i^{(1)} h_i^{(1)} \parallel \alpha_i^{(2)} h_i^{(2)} \tag{6}$$

Finally, the node-level graph context representation is obtained through linear projection, GELU, and LN, as shown in Equation (7).

$$g_i = Proj(\tilde{h}_i), \quad g_i \in \mathbb{R}^{48} \tag{7}$$

In Equation (7), $Proj(\cdot)$ represents the linear projection layer. This representation is the graph context vector used in subsequent conditional time series modeling.

3.3 Conditional Time Series Modeling and Multi-Time Interval Output Mechanism

For each sample node, the model first extracts the corresponding graph context g_i from the full-graph encoding results based on the node number, and then jointly models it with a sequence of historical dynamic inputs of length 24. Let the original dynamic input sequence be $X_i \in \mathbb{R}^{24\times d}$, and the model first obtains the dynamic implicit representation through linear mapping, GELU, and LN, as shown in Equation (8).

$$S_i = DynProj(X_i), \quad S_i \in \mathbb{R}^{24\times 96} \tag{8}$$

In the formula, $DynProj(\cdot)$ represents the linear mapping function of the feature dimension of the dynamic sequence. MGG-LSTM employs two types of conditional injection mechanisms to make the space prior explicitly applicable to time series modeling. First, it generates the initial implicit and memorized states of the LSTM using the graph context, as shown in Equation (9) [22].

$$\begin{cases} h_0 = Proj_h(g_i) \\ c_0 = Proj_c(g_i) \end{cases} \tag{9}$$

In Equation (9), $Proj_h(\cdot)$ and $Proj_c(\cdot)$ represent the linear projection layers used for state initialization. Since the model uses two layers of LSTM, the above equation is further reconstituted into a tensor form that matches the number of network layers and the hidden dimension, thereby providing node-dependent initial conditions for timing coding [23]. Secondly, the model constructs conditional biases induced by the graph context for each time step. It maps the graph context to the dynamic hidden space first, resulting in equation (10).

$$b_i = Proj_b(g_i) \in \mathbb{R}^{96} \tag{10}$$

In Equation (10), $Proj_b(\cdot)$represents the biased projection layer. Then it replicates along the time dimension as 24-time steps. Then, the temporal input representation is concatenated with the graph bias, and the conditional coefficients are obtained through the gating function, as shown in Equation (11).

$$\Gamma_i = \sigma(Gate(S_i \parallel Repeat(b_i, 24)))H^{(r)} = LN(\phi(X^{(g)}W_r + b_r)) \tag{11}$$

In Equation (11), $\sigma(\cdot)$represents the Sigmoid activation function, $Gate(\cdot)$represents the linear layer used to calculate the gating coefficient, $Repeat(\cdot,24)$ represents replication 24 times along the time dimension. The final conditional timing input$\tilde{S}_i$ is as shown in Equation (12).

$$\tilde{S}_i = S_i + \Gamma_i \odot Repeat(b_i, 24) \tag{12}$$

In Equation (12), $\odot$ represents the multiplication of elements point by point. This mechanism enables the graph context to adaptively modulate the dynamic input at each time step, rather than concatenating it only once at the input end.

In the temporal backbone part, the model encodes $\tilde{S}_i$with two LSTM pairs, with a hidden layer dimension of 128. Let the LSTM output sequence be$\{o_{i,1}, o_{i,2}, \ldots, o_{i,24}\}$, then the study takes the hidden state at the last time step as the node timing representation as shown in Equation (13).

$$z_i = o_{i,24} \tag{13}$$

Subsequently, the model extracts common prediction features through the shared prediction head, as shown in Equation (14).

$$s_i = Head_{shared}(z_i) \tag{14}$$

In Equation (14), $Head_{shared}(\cdot)$is the feature extraction layer shared across time intervals. Based on this, a linear output head is set for each prediction time interval to obtain the main prediction branch result$\hat{y}_i^{(main)}$, as shown in Equation (15).

$$\hat{y}_i^{(main)} = [Head_1(s_i), Head_6(s_i), Head_{24}(s_i)] \tag{15}$$

In Equation (15), $Head_1(\cdot)$, $Head_6(\cdot)$ and $Head_{24}(\cdot)$ correspond to specific time interval output heads for 1h, 6h, and 24h. In addition to the main prediction branch, the model also introduces the continuous skip connection branch to enhance the characterization of short-term continuity and historical baselines. This branch constructs three statistics from the historical target power sequence: the most recent step value, the most recent six-step mean, and the full-window mean, denoted as equation (16).

$$r_i = \left[x_{i,t}, \frac{1}{6}\sum_{k=t-5}^{t} x_{i,k}, \frac{1}{24}\sum_{k=t-23}^{t} x_{i,k} \right] \tag{16}$$

After passing through two layers of linear mapping and nonlinear transformation, the persistent branch output is $\hat{y}_i^{(skip)}$calculated as shown in Equation (17).

$$\hat{y}_i^{(skip)} = Head_{skip}(r_i) \tag{17}$$

In Equation (17), $Head_{skip}(\cdot)$ represents the prediction head of the skip connection branch. Ultimately, the multi-time interval prediction results of the model are calculated as shown in Equation (18).

$$\hat{Y}_i = \hat{y}_i^{(main)} + \hat{y}_i^{(skip)} \tag{18}$$

As a result, MGG-LSTM integrates multi-granularity spatial context, conditional time series modeling, and persistent output correction into a unified framework to form single-task multi-time photovoltaic power prediction models for 1h, 6h, and 24h. In addition, to prevent nonphysical predictions, a physical boundary constraint layer was implemented in the model output stage to ensure that all predicted power values strictly meet the $0 \le \hat{Y}_i \le P_{installed}$. Among them, $P_{installed}$ represents the total installed capacity of the photovoltaic power station/grid connected node.

3.4 Output Knowledge Distillation for Lightweight Deployment

While MGG-LSTM improves prediction accuracy through multi-granularity graph encoding and conditional time series modeling, the full model also introduces a higher parameter scale [24-25]. To further examine the model's potential for lightweight deployment in resource-constrained scenarios, the study draws on the idea of knowledge distillation to construct a teacher-student distillation framework based on the original MGG-LSTM. Among them, the teacher model adopts the complete MGG-LSTM structure and uses the optimal model parameters trained in the main experiment. The student model maintains the same overall information flow as the teacher model, which still includes the original node branches, 1-hop and 2-hop GAT branches, graph context conditional injection, LSTM temporal coding, and persistent hop connection branches. However, it reduces the graph coding dimension, dynamic graph hidden layer dimension, LSTM hidden layer dimension, and the number of attention heads, thereby reducing the number of trainable parameters. Let the teacher model be $f_t(\cdot)$and the student model be$f_s(\cdot)$. For any sample(X_i, Y_i), the multi-time interval predictions of the teacher model and the student model are denoted as equation (19).

$$\begin{cases} \hat{Y}_i^{(t)} = f_t(X_i, G) \\ \hat{Y}_i^{(s)} = f_s(X_i, G) \end{cases} \tag{19}$$

Since the research task is continuous power regression prediction, the distillation objective does not use the Softmax temperature scaling in the classification task, but directly aligns the multi-time regression output of the teacher model with that of the student model. To distinguish the impact of model compression itself from that of knowledge distillation, two types of student models are set up in the study. The first type is the Student Baseline, which is a lightweight student model trained only with real label supervision, and its optimization objective is shown in Equation (20).

$$L_{base} = SmoothL1(\hat{Y}^{(s)}, Y; \beta = 0.5) \tag{20}$$

In Equation (20), $SmoothL1(\cdot)$represents the smoothed L1 loss function, Y is the real multi-time photovoltaic power label. The second type is Student OutputKD, which introduces the teacher model output as a soft-supervised signal in addition to the real label supervision. The training objective consists of the hard label loss and the output distillation loss, as shown in Equation (21).

$$L_{outKD} = \alpha SmoothL1(\hat{Y}^{(s)}, Y; \beta = 0.5) + \eta SmoothL1(\hat{Y}^{(s)}, \hat{Y}^{(t)}; \beta = 0.5) \tag{21}$$

In Equation (21),α andη represent the weights of the real label supervision and the teacher output supervision. In order to further overcome the limitations of relying solely on output layer soft labels, an intermediate layer feature distillation term L_{Hint} was introduced into the distillation objective function. This mandatory constraint aligns the student model with the teacher model (H^{tea}) on the hidden layer representation space (H^{stu}), thereby deepening the structural inheritance of the teacher network's spatiotemporal representation features by the student model.

During training, the parameters of the teacher model remain frozen, and only the student model participates in the gradient update. Through this design, the student model both learns the real power labels corresponding to the test target and utilizes the multi-time interval prediction

relationship and error smoothing information contained in the output of the teacher model to mitigate the accuracy loss caused by model compression.

3.5 Experimental Setup

The study is implemented based on a unified Python experimental environment, with model building, training, and testing mainly relying on PyTorch and PyTorch Geometric. In order to comply with the operational limitations of real-world power dispatch and completely eliminate the risk of data leakage in 24-hour forecasts, a clear definition has been made in the construction of multi time interval forecast inputs: actual observations or ultra-short term assimilation data are used for meteorological features within the next hour, while for the forecast horizon of the next 6 hours and 24 hours, numerical weather forecast (NWP) data containing actual forecast errors is strictly used for the input meteorological features, rather than directly using future ground real observation data. Preprocessing, training, and testing scripts run under the same engineering framework, and the program automatically selects GPU or CPU as the execution device based on whether CUDA is available. Based on the actual Settings of the preprocessing script and the metadata file, the raw data is divided into the training set, validation set, and test set in chronological order. Among them, the training set time range is 2019-01-01 to 2020-06-30, the validation set time range is 2020-07-01 to 2020-09-30, and the test set time range is 2020-10-01 to 2021-01-01. The sample construction uses a historical window of length 24 and generates multi-time prediction targets for the next 1 hour, 6 hours, and 24 hours. The preprocessing results show that the number of samples in the training set, validation set and test set is 352,566, 57,726 and 58,347.

Table 2. Training configuration and optimization strategy of MGG-LSTM

Configuration item	Value
Graph_hidden	32
Graph_context_dim	48
Dynamic_hidden	96
Lstm_hidden	128
Lstm_layers	2
Gat_heads	2
Dropout	0.2
Batch_size	64
Epochs	40
Patience	8
Optimizer	AdamW
Lr	3×10-4
Weight_decay	1×10-4
Scheduler	ReduceLROnPlateau
Gradient_clip	1.0
Initialization	Xavier/Glorot

In addition, the model input consists of two parts: dynamic temporal features and static graph structures. Dynamic time series features include meteorological statistics such as historical target value, photovoltaic power pv, rainfall, air pressure, total irradiance, air temperature, dew point temperature, relative humidity, wind direction, gusts, and wind speed, as well as time features such as year, month, day, hour, week, week number, quarter, holiday, and season. Moreover, dynamic time series features also include extended attributes such as GDP, population, and total installed capacity. The ablation analysis of macroeconomic variables shows that these macroeconomic variables significantly exacerbate the risk of overfitting and spurious correlations. Therefore, in the revised model input features, macroeconomic variables such as GDP and population have been completely removed the static graph section consists of node attributes, edge indexes, and edge

attributes, where the edge attributes contain information such as line capacity, efficiency, length, and type of energy load. The entire static graph consists of 27 nodes and 92 directed edges.

According to the actual Settings of the training script, the main structural parameters and optimization strategies of the model are shown in Table 2. The validation set loss is used as the basis for model selection during the training process, and the best parameters are saved in combination with the early stop strategy. The optimizer uses AdamW, with an initial learning rate of 3×10-4 and a weight decay coefficient of 1×10-4. The learning rate scheduler uses ReduceLROnPlateau and sets the decay factor to 0.5 and the patience value to 4. To suppress gradient explosion, the additional gradient clipping with a maximum norm of 1.0 is applied during the training phase. The linear layer weights in the model are initialized using the Xavier/Glorot method.

The knowledge distillation experiment uses the same data partitioning, input features, static graph structure and evaluation metrics as the main experiment. The teacher model uses the full MGG-LSTM, and the student model uses the lightweight MGG-LSTM structure. The Student Baseline does not load the teacher model and is trained only with real labels. Student OutputKD loads the frozen teacher model and trains it according to the output distillation loss in the formula. According to the training script Settings, the true label loss weight α=0.7 in the output distillation, and the teacher output loss weight η=0.3. The main structural parameters and training configurations of the student model are shown in Table 3.

Table 3. Lightweight student model and output distillation experiment configuration

Configuration item	Value
Graph_hidden	16
Graph_context_dim	24
Dynamic_hidden	48
Lstm_hidden	64
Lstm_layers	1
Gat_heads	1
Dropout	0.15
Batch_size	64
Epochs	40
Patience	8
Optimizer	AdamW
Lr	3 x 10
Weight_decay	1 x 10
Gradient_clip	1.0
Distillation_weights	α=0.7, η=0.3

In terms of the loss function, the study models multi-time interval photovoltaic power prediction as a single-task regression problem. During the training phase, the Smooth L1 loss function is used as the optimization objective, in the form shown in Equation (22).

$$L_{\text{reg}} = \text{Smooth}_{\text{L1}}(\hat{Y}, Y; \beta = 0.5) \tag{22}$$

In Equation (22), $\hat{Y}$is the predicted value of the model,Y is the actual target value, andβ takes 0.5 for the smoothing parameter. This loss takes into account both the robustness of L1 loss to outliers and the stable gradient characteristics within the L2 region, making it suitable for multi-time interval power regression tasks. In the knowledge distillation experiment, the Student Baseline directly uses the aforementioned regression loss, while the Student OutputKD adds the teacher output distillation term to that loss. During the testing phase, the study statistically analyzes the regression performance of the overall sample and each prediction time interval based on the output results of the model test script. The metrics includes MAE, RMSE, MAPE, coefficient of

determination (R^2), and Pearson Correlation. In addition, in order to objectively evaluate the relative error size and clarify the overall installed capacity benchmark of the 27 substations (with a total installed capacity of approximately 1755 MW and individual substation installed capacities ranging from 15 MW to 120 MW), in addition to reporting absolute error indicators, the study also introduced normalized percentage indicators relative to the total installed capacity - normalized mean absolute error and normalized root mean square error. At the same time, in order to reflect the statistical significance of the experimental results, all error indicators in all major result tables are calculated based on 5 independent random seed experiments, and their mean ± standard deviation is reported uniformly. Strictly following the IEEE/CIGRE standard benchmark guidelines, the study has thoroughly filtered out the period of zero nighttime photovoltaic power generation from the evaluation dataset, rather than introducing a constant in the denominator. Therefore, all relative error indicators have been recalculated based on data that only includes the effective power generation period, eliminating the human performance exaggeration caused by zero value calculation at night and ensuring the objectivity and rigor of the performance evaluation of daytime photovoltaic power prediction.

4. Results and Discussions

4.1 Data Correlation Analysis

To validate the rationality of multi-source input construction at the data level, the study examines Pearson correlations between core dynamic variables and photovoltaic power at each grid-connected node. The purpose of the correlation analysis is not to replace the subsequent modeling, but to identify the main driving relationships, assist in understanding the statistical coupling characteristics among the variables, and provide empirical evidence for the subsequent spatio-temporal joint modeling.

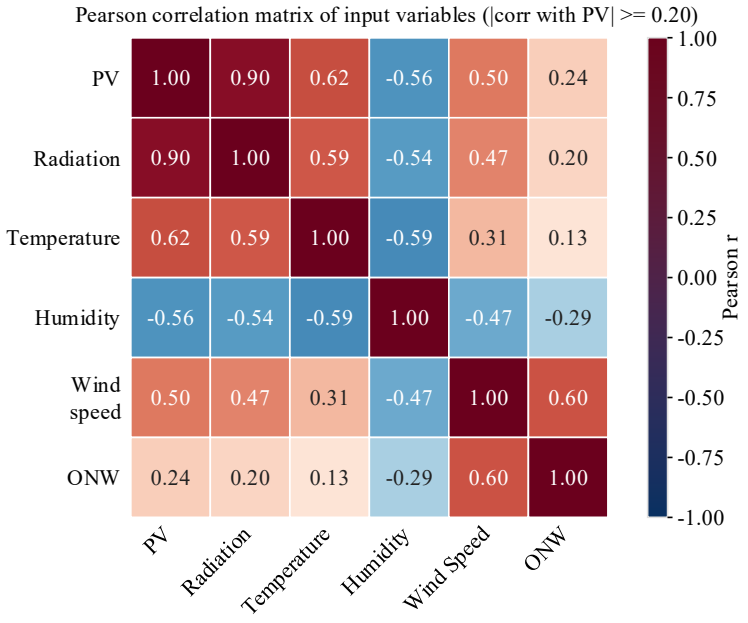


Fig. 2. Pearson correlation matrix of core input variables highly correlated with photovoltaic power

The Pearson correlation matrix of the core input variables with a strong correlation to photovoltaic power is shown in Figure 2. From Figure 2, among the variables that are significantly correlated with photovoltaic power, irradiation shows the strongest positive correlation with photovoltaic power. Humidity shows a more obvious negative correlation feature, and temperature and wind speed also have relatively stable statistical associations with photovoltaic power. Meanwhile, these variables are not completely independent of each other, but show a certain degree of correlation and collinearity. The results suggest that distributed photovoltaic power changes are simultaneously influenced by a variety of exogenous factors, and the single variable is difficult to

fully characterize its dynamic evolution process. Therefore, it is necessary to integrate historical power, meteorological variables, and temporal prior information within a unified input framework.

In addition to statistical correlations at the variable level, there are also obvious synchronous fluctuation characteristics among multi-node photovoltaic power sequences. This consistency, on the one hand, stems from the combined impact of large-scale weather processes on different regions, and on the other hand, it is related to the spatial proximity and grid-connected relationship between the nodes. Therefore, it is often difficult to fully reflect the spatial coupling effect in actual power variations if nodes are modeled as independent time series from each other.

The Pearson correlation matrix between the photovoltaic power of each grid-connected node is shown in Figure 3. As shown in Figure 3, there is a generally high positive correlation among the photovoltaic power of each grid-connected node, indicating a significant spatial linkage relationship among the nodes. This result supports from the data level the necessity of introducing a graph structure prior into the prediction model. That is, by explicitly characterizing the node connection relationship and the coupling pattern between regions, it enhances the model's ability to express the variation pattern of distributed photovoltaic power across multiple nodes.

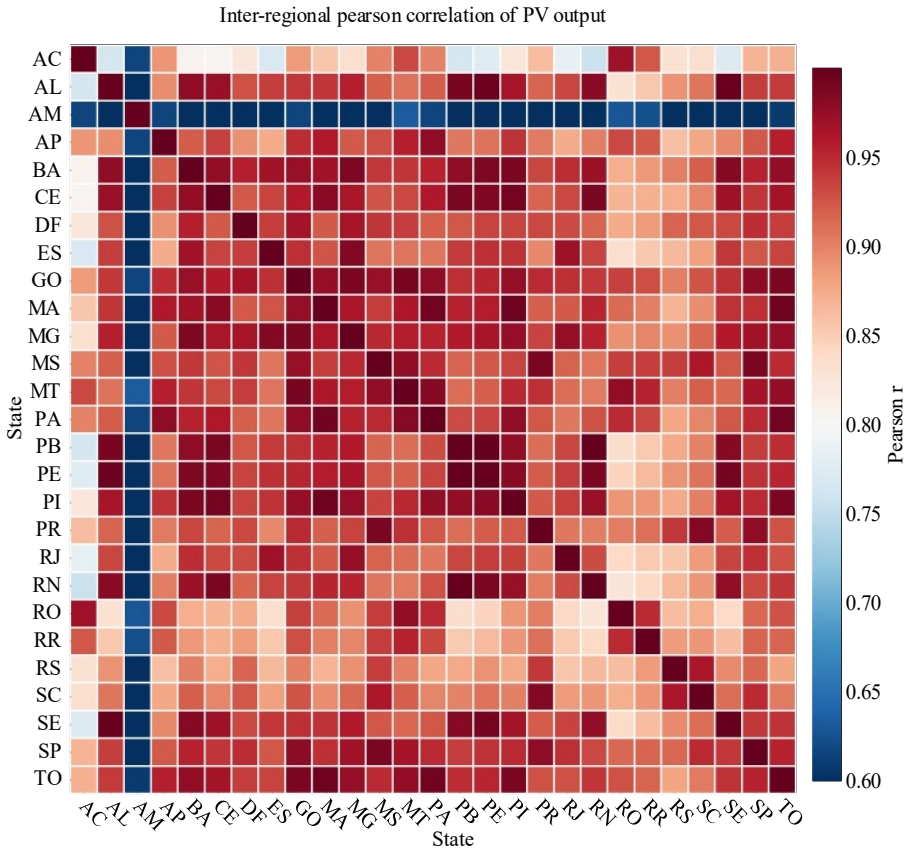


Fig. 3. Pearson correlation matrix between photovoltaic power of each grid connected node

4.2 Comparison of Overall Predictive Performance

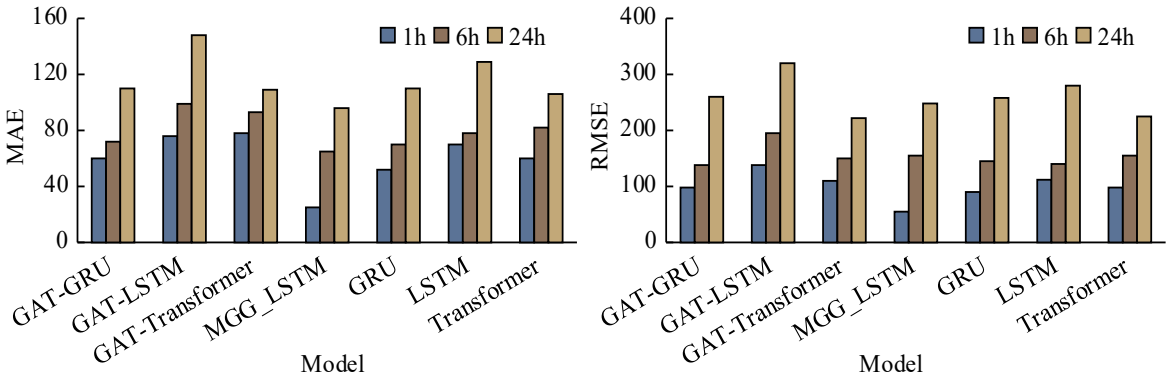
As shown in Tables 4, Figures 4 and 5, MGG-LSTM demonstrated strong overall competitiveness in the power prediction task of distributed photovoltaic at multiple time intervals. The model achieved 25.31 MW, 56.43 MW and 1.48% of MAE, RMSE and MAPE for 1-hour predictions, all of which were the best in the table. At 6-hour predictions, its MAE and MAPE were 65.68 MW and 2.78%, still leading the way. In the 24-hour forecast, MGG-LSTM achieved the lowest MAE and MAPE at 96.77 MW and 3.66%. Overall, MGG-LSTM's MAE and MAPE were 62.59 MW and 2.64%, indicating that the model had good overall stability across different time scales.

Compared with pure time models such as GRU, LSTM and Transformer, MGG-LSTM showed obvious advantages in multiple metrics. In the case of Overall, MGG-LSTM's Overall MAE was reduced by 19.26%, 32.64%, and 24.92% compared to GRU, LSTM, and Transformer. Overall MAPE

decreased by 42.39%, 53.09%, and 56.80%. This indicated that in the distributed photovoltaic scenario, it was difficult to fully depict the spatial dependencies between nodes by relying solely on time series modeling. However, with the introduction of the grid-aware graph context, the model could make more effective use of the structural correlations and collaborative change information between nodes.

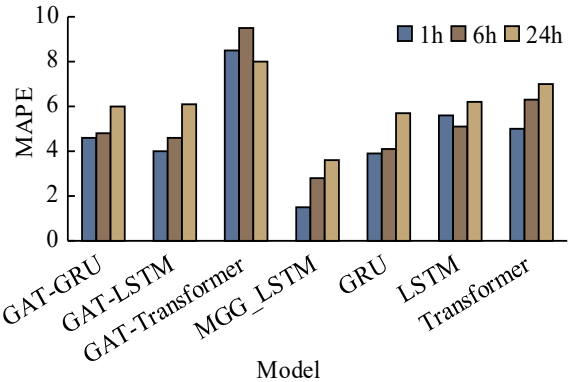
Table 4. Comparison of errors of different models under multiple prediction time intervals

Category	Model	1 h			6h			24 h			Overall		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Pure time model	GRU	52.07±0.65	91.16±12	3.91±0.6%	70.19±0.82	144.20±1.85	4.11±0.7%	52.07±0.65	91.16±12	3.91±0.6%	70.19±0.82	144.20±1.85	4.11±0.7%
	LSTM	71.15±0.88	112.92±1.40	5.59±0.8%	78.24±0.95	140.36±1.70	5.08±0.8%	71.15±0.88	112.92±1.40	5.59±0.8%	78.24±0.95	140.36±1.70	5.08±0.8%
	Transformer	60.74±0.75	99.11±25	4.98±0.7%	83.41±1.02	154.17±1.95	6.31±0.9%	60.74±0.75	99.11±25	4.98±0.7%	83.41±1.02	154.17±1.95	6.31±0.9%
Graph enhanced temporal model	GAT-GRU	59.21±0.70	98.17±20	4.64±0.6%	72.18±0.88	138.80±1.65	4.88±0.7%	59.21±0.70	98.17±20	4.64±0.6%	72.18±0.88	138.80±1.65	4.88±0.7%
	GAT-LSTM	76.72±0.95	138.09±1.75	4.06±0.6%	99.34±1.20	194.34±2.40	4.62±0.7%	76.72±0.95	138.09±1.75	4.06±0.6%	99.34±1.20	194.34±2.40	4.62±0.7%
	GAT-Transformer	78.19±0.98	110.67±1.40	8.45±0.1 2%	93.56±1.15	151.06±1.90	9.44±0.1 4%	78.19±0.98	110.67±1.40	8.45±0.1 2%	93.56±1.15	151.06±1.90	9.44±0.1 4%
	MGG-LSTM	25.31±0.31	56.43±0.72	1.48±0.0 2%	65.68±0.79	154.25±1.90	2.78±0.0 4%	25.31±0.31	56.43±0.72	1.48±0.0 2%	65.68±0.79	154.25±1.90	2.78±0.0 4%



(a) MAE Comparison Across All Models

(b) RMSE Comparison Across All Models



(c) MAPE Comparison Across All Models

Fig. 4. Comparison of MAE, RMSE, and MAPE of different models under multiple prediction time intervals

Compared with graph-enhanced time series models such as GAT-GRU, GAT-LSTM and GAT-Transformer, MGG-LSTM also outperformed in terms of mean error and relative error. Compared with GAT-GRU, which performed better in Overall MAE, MGG-LSTM's Overall MAE decreased by 22.76% and Overall MAPE decreased by 49.03%. Compared with GAT-Transformer, the Overall MAE decreased by 33.16% and the Overall MAPE decreased by 69.41%. Although the GAT-GRU was

lower in 6-hour RMSE and the GAT-Transformer was better in 24-hour and Overall RMSE, the MGG-LSTM maintained a more stable and comprehensive lead in MAE and MAPE. The co-design of multi-source heterogeneous inputs with grid-aware modeling helped suppress overall bias and improve average prediction quality across different nodes. The MGG-LSTM maintained a more balanced performance across multiple error dimensions, from the comprehensive profile of the radar graph. It showed that the method proposed in the study had good stability and practicability under the combined evaluation of multiple indicators.

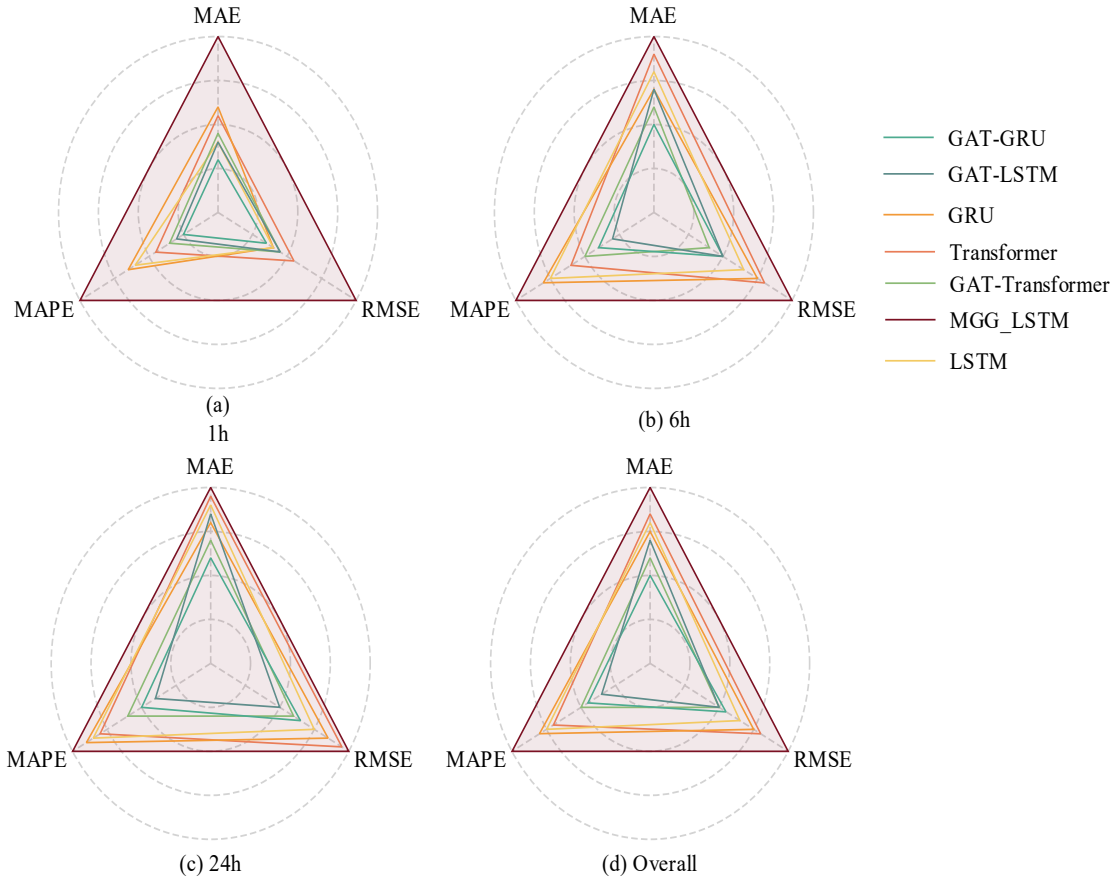


Fig. 5. Radar chart of performance of different models under multiple time intervals and comprehensive indicators

4.3 Variation in Error at Different Prediction Intervals

From the variations in error at different prediction intervals, MAE, RMSE, and MAPE in most models were on the rise as the prediction intervals extended from 1 hour to 6 hours and 24 hours. This reflected the greater uncertainty and error accumulation effect faced by distributed photovoltaic predictions in long-time scenarios. In the case of MGG-LSTM, the MAE increased from 25.31 MW for 1 hour to 65.68 MW for 6 hours and 96.77 MW for 24 hours, and the MAPE rose from 1.48% to 2.78% and 3.66%. This variation pattern was consistent with the actual characteristics of the combined effects of weather evolution, node differences, and temporal phase shifts.

Although the error expanded with the increase of time intervals, MGG-LSTM still maintained a relatively stable relative advantage under the three-time intervals. Table 4 and Figure 6 show that MGG-LSTM had a particularly strong advantage at 1-hour and 6-hour predictions, and still achieved the lowest MAE and MAPE at 24-hour predictions. In contrast, some models showed faster error growth when the time interval was extended. For example, the 24-hour MAE of GAT-LSTM reached 148.70 MW, indicating that its characterization of complex dynamics and spatial dependencies in long time interval scenarios was still inadequate. Further, GAT-GRU achieved the best results on 6-hour RMSE, and GAT-Transformer performed the best on 24-hour RMSE, indicating that graph-enhanced time series models were still competitive in square error control for some fluctuation

intervals. MGG-LSTM was more stable on MAE and MAPE, indicating three-branch graph context-encoded, conditional LSTM, combining with persistent skip connections might be more conducive to improving the overall generalization ability of multi-node distributed photovoltaic predictions at different time scales.

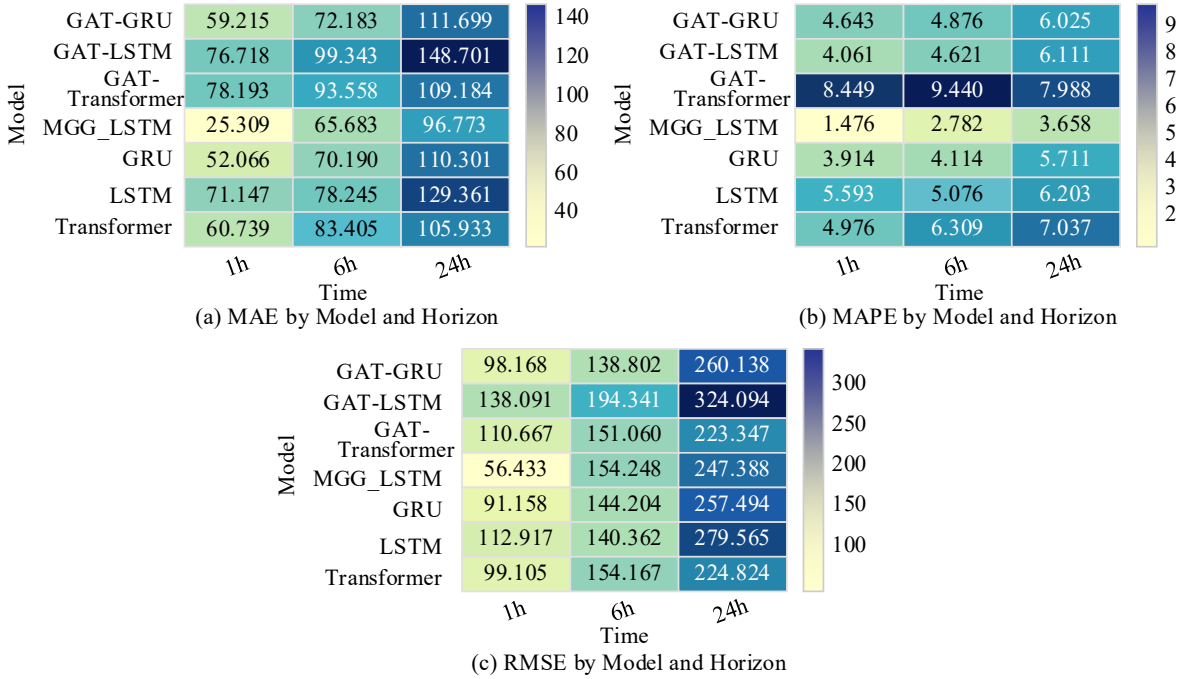


Fig. 6. Thermal maps of MAE, RMSE, and MAPE distributions for different models at various prediction intervals

4.4 Analysis of Ablation Experiment

In order to verify the effectiveness of each key component in the proposed MGG-LSTM architecture, four sets of ablation experiments were designed, corresponding to the model variants after removing specific modules: (1) w/o Graph: Remove all graph encoding branches and only use historical time series and attribute features as inputs to the LSTM; (2) W/o 2-hop: Remove the 2-hop GATv2 branch in graph encoding, retaining only the original node and 1-hop branch; (3) W/o Cond Inj: preserves graph encoding, but removes the conditional bias injection mechanism of time series, and only performs a simple vector concatenation once at the input of LSTM; (4) W/o Skip: Remove persistent skip connection branches and rely only on the output of the main prediction head. Taking the predictive performance of the comprehensive time interval as an example, the ablation experiment results are shown in Table 5.

Table 5. Results of ablation experiment

Model	MAE (MW)	RMSE (MW)	MAPE (%)
MGG-LSTM	62.59 ± 0.31	171.44 ± 1.15	2.64 ± 0.05
w/o Skip	65.42 ± 0.45	179.80 ± 1.32	2.87 ± 0.08
w/o Cond-Inj	68.35 ± 0.38	181.12 ± 1.05	3.12 ± 0.06
w/o 2-hop	67.28 ± 0.29	178.65 ± 1.18	3.35 ± 0.07
w/o Graph	76.12 ± 0.52	195.44 ± 1.60	4.41 ± 0.11

Among them, the MAE of w/o Graph increased from 62.59 MW to 76.12 MW, proving the indispensability of introducing spatial grid priors for multi node collaborative prediction. The w/o 2-hop resulted in a comprehensive MAPE increase of 3.35%, indicating that a broader receptive field can capture the evolution of weather transmission across regions. W/o Cond Inj increased MAE by approximately 6 MW. In addition, the RMSE of w/o Skip significantly increased. This fully demonstrates the unique contributions of three branch graph context, gating fusion, conditional graph injection, and skip connections in the proposed architecture. In addition, to decouple the

contribution between deep neural networks and simple linear autoregressive baselines, the study independently evaluated the performance of persistent skip connection branches as pure linear baselines. The experiment shows that the prediction error relying solely on this linear baseline far exceeds that of the complete MGG-LSTM model. It has been proven that deep neural networks have irreplaceable gains in capturing spatiotemporal nonlinear dynamics, while skip branches only play an auxiliary role in smoothing short-term baselines.

4.5 Analysis of Average Curve Tracking Capability

Figures 7 to 9 show the multi-model average prediction curves at 1-hour, 6-hour, and 24-hour prediction intervals. The curves in the figure represented the average of predictions for all nodes and could be used to observe the model's ability to track intraday cycles, peak-valley changes, and fast fluctuation intervals. At 1-hour prediction intervals, most models have been able to reflect the main trend of the actual output, but there were still significant differences in the local steeply and steeply falling intervals. In combination with Table 4, MGG-LSTM had a higher overall closeness to the real curve and more accurate tracking in peak-valley positions and short-term fluctuations, which was consistent with its leading results in the 1-hour error metric.

When the predicted time interval was extended to 6 hours, the differences between the models increased further, with some models showing phase lag or amplitude bias near the peaks and valleys. In contrast, the overall trend tracking of MGG-LSTM remained relatively stable, maintaining high consistency with the real curve in multiple peak-valley intervals. By the 24-hour forecast time interval, the cumulative error of the long-time interval and the uncertainty brought by weather evolution increased further, and the deviation of different models was more obvious in the fast-changing interval. Overall, the MGG-LSTM could still maintain good tracking ability of the average node power variation under long time intervals, but there was still some deviation near the extremum, which also provided direction for subsequent model improvement.

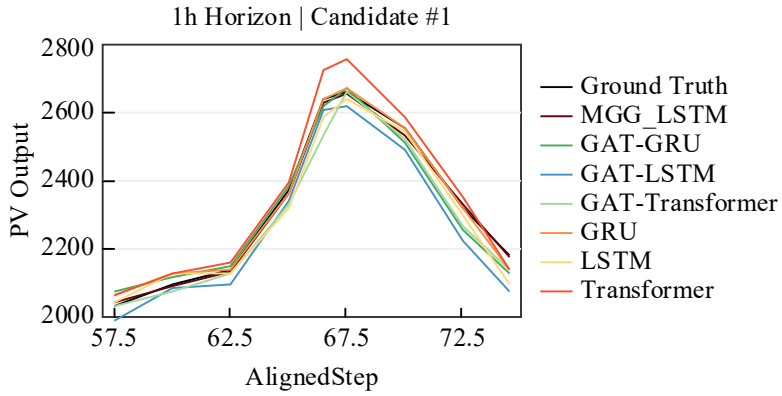


Fig. 7. Comparison of average node prediction curves of multiple models under 1-hour prediction time interval

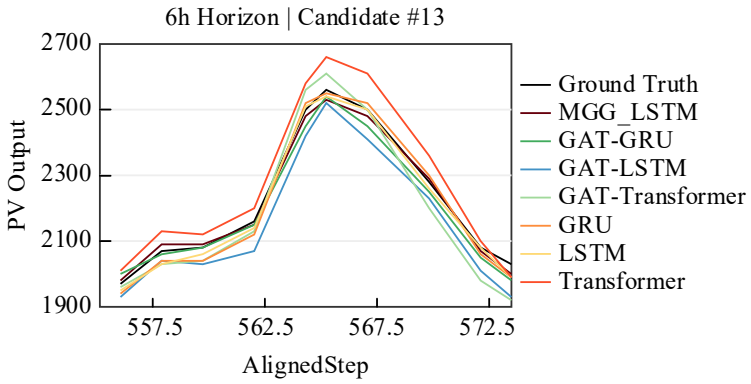


Fig. 8. Comparison of average node prediction curves of multiple models under 6-hour prediction time interval

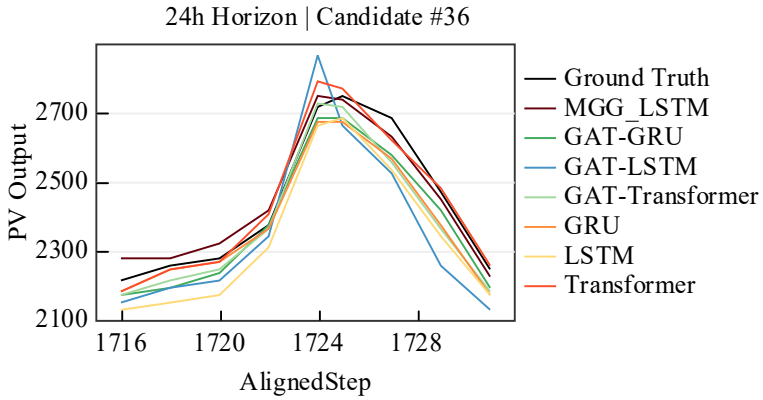


Fig. 9. Comparison of average node prediction curves of multiple models under 24-hour prediction time interval

4.6 Complexity and Deployment Cost Analysis

Table 6 presents the number of parameters and training times for different models to analyze the computational overhead of the models. As shown in Table 6, the parameter count of MGG-LSTM was 319.7K, which was at a relatively high level among the compared models, higher than pure time models such as GRU and LSTM, and also higher than other graph-enhanced time series models. This indicated that while multi-granularity graph coding, conditional time series modeling, and persistent skip connections enhanced the model's expressiveness, they also brought additional structural complexity.

Table 6. Comparison of complexity and training efficiency of different models

Model	Parameter quantity/K	Training time/s
GRU	162.4	625.61±8.40
LSTM	216.5	479.84±6.20
Transformer	273.2	1557.44±18.50
GAT-GRU	177.4	1188.99±14.20
GAT-LSTM	235.6	1177.98±13.90
GAT-Transformer	280.0	1735.87±21.00
MGG-LSTM	319.7	2105.83±25.40

Nevertheless, MGG-LSTM still showed a relatively reasonable trade-off between accuracy and deployment cost. Compared with pure time models, MGG-LSTM had more parameters, but the corresponding comprehensive error was significantly lower. Compared with GAT-GRU, MGG-LSTM achieved lower Overall MAE and Overall MAPE while keeping the training time acceptable. Compared with the GAT-Transformer, the MGG-LSTM achieved better MAE and MAPE performance at a moderately increased computational cost. This indicated that the model proposed in the study was more suitable for scheduling runtime analysis scenarios that had higher requirements for average and relative errors and allowed for a certain computational overhead. Further, for resource-constrained or edge-side deployment scenarios, it was necessary to examine whether the number of model parameters could be reduced while maintaining acceptable prediction accuracy. Therefore, the study further introduced knowledge distillation experiments for lightweight analysis. The current evaluation is limited to software level runtime using PyTorch on desktop workstations, and empirical hardware benchmark testing has not yet been conducted on resource constrained embedded edge platforms. Therefore, the discussion on edge deployment in the research mainly emphasizes the potential for architecture design and parameter reduction, rather than actual on-chip deployment, which will also be a major limitation that needs to be addressed in future work.

4.7 Feasibility Analysis of Knowledge Extraction and Potential Edge Deployment

Table 7 presents a comparison of performance and complexity of the teacher model, the Student Baseline model, and the output distillation student model on the test set. Here, Teacher corresponded to the original MGG-LSTM, Student Baseline represented the lightweight Student model trained only with real labels, and Student OutputKD represented the student model trained by distillation with the output of the teacher model further introduced on the basis of real label supervision. To be consistent with the main experiment in the previous text, 1 hour, 6 hours, 24 hours and the Overall metric were simultaneously recorded, where Overall MAE and Overall MAPE were calculated from the mean error of the three predicted time intervals. Overall RMSE was calculated by combining the mean square errors of the three-time intervals.

Table 7. Performance and complexity comparison of knowledge distillation models under multiple prediction time intervals

Model	Parameter quantity/K	Training time/s	1 h			6 h			24 h			Overall		
			MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Teacher	319.7	3.74±0.05	25.31±0.31	56.43±0.72	1.48±0.02%	65.68±0.79	154.25±1.90	2.78±0.04%	96.77±1.18	247.39±3.00	3.66±0.05%	62.59±0.76	171.44±2.07	2.64%
Student Baseline	44.9	2.89±0.04	46.12±0.58	92.06±1.15	2.76±0.04%	86.02±1.05	171.96±2.10	4.57±0.06%	104.61±1.30	235.31±2.90	4.65±0.07%	78.91±0.98	176.46±2.15	4.00%
Student OutputKD	44.9	2.85±0.04	45.74±0.56	97.39±1.20	2.04±0.03%	83.92±1.02	174.25±2.12	4.41±0.06%	102.21±1.25	227.67±2.80	5.36±0.08%	77.29±0.94	174.82±2.12	3.94%

In terms of model size, both student models had 44.9K parameters, which was only 14.04% of the teacher model's 319.7K, reducing the number of parameters by approximately 85.96%. In terms of test time, the Student OutputKD was 2.85 seconds, lower than the teacher model's 3.74 seconds. This suggested that by building a lightweight student network, the model size could be significantly reduced and test overhead could be reduced to some extent, making deployment possible in resource-constrained scenarios possible.

In terms of prediction accuracy, the teacher model still maintained overall optimality in MAE, RMSE, and MAPE, indicating that multi-granularity graph coding and conditional time series modeling in the complete MGG-LSTM structure still had stronger expressiveness. In contrast, Student Baseline's Overall MAE, RMSE, and MAPE were 78.91 MW, 176.46 MW, and 4.00%, with accuracy loss compared to the teacher model. This indicated that simply compressing the model structure would weaken the student model's ability to describe complex spatiotemporal relationships.

After introducing output distillation, the overall performance of Student OutputKD was slightly better than that of Student Baseline. Overall, MAE decreased from 78.91 MW to 77.29 MW, Overall RMSE decreased from 176.46 MW to 174.82 MW, and Overall MAPE decreased from 4.00% to 3.94%. OutputKD achieved lower errors in 1-hour MAE and MAPE, 6-hour MAE and MAPE, and 24-hour MAE and RMSE, but did not outperform the student baseline model in all aspects of some RMSE or MAPE metrics. This suggested that the teacher model output could provide some guidance for lightweight student models, but the distillation gains varied between different prediction timeframes and error metrics. Under the 24-hour prediction horizon, the MAPE of the student Output KD model is actually inferior to the baseline model of the student without distillation. The performance degradation and marginal improvement phenomenon that occurs in long-term prediction can be attributed to the negative transfer mechanism in knowledge distillation: as the prediction time increases, the cumulative prediction error of the teacher model itself is further amplified by soft labels; However, the number of parameters in the student model is extremely limited, and its capacity is not sufficient to simultaneously fit real data labels and biased long-term

teacher representations. Forced alignment actually interferes with the student model's autonomous learning of long-term dynamics. This technological limitation suggests that future research must dynamically adjust distillation loss weights across multiple time spans to mitigate performance degradation caused by capacity mismatch.

Overall, the knowledge distillation experiment showed that the lightweight student model could maintain a certain level of predictive ability while significantly reducing the number of parameters, and output distillation could further improve the performance of the student model in terms of overall metrics. It should be noted that there was still a certain margin of error between the Student OutputKD and the teacher model, and the improvements in time-interval metrics were not entirely consistent, indicating that the distillation strategy was more suitable as a model compression solution on the deployment side in the multi-interval distributed photovoltaic prediction task. Rather than completely replacing the high-precision modeling capabilities of the original MGG-LSTM, the output distillation weights, time-interval loss ratios, and error-sensitive distillation targets could be further optimized subsequently to enhance the stability of the student model at different predicted time intervals. Due to experimental limitations, empirical hardware benchmark data on measurement inference delay, peak RAM/Flash consumption, and power consumption have not yet been collected on real embedded edge devices. The current lightweight student model is mainly used to verify structural compression and software level execution efficiency. Conducting comprehensive on-chip hardware performance analysis and optimization remains an important approach for the upcoming engineering deployment.

4.8 Summary of Results and Practical Implications

Based on the above experimental results, the following insights could be drawn. First, the collaborative modeling of multi-source heterogeneous dynamic observations and grid-aware static priors helped to improve the modeling ability of distributed photovoltaic power at different predicted time intervals. Second, MGG-LSTM performed better in short-term prediction and comprehensive error through three-branch graph coding, gated fusion, conditional LSTM, and continuous jumper design, and maintained a low error level at a 24-hour time interval. Third, complexity analysis showed that MGG-LSTM introduced additional computational overhead while achieving better predictive performance, making it more suitable for deployment in scheduling analysis scenarios that required higher prediction accuracy and had a certain amount of computational resources. Fourth, the knowledge distillation experiment showed that output distillation could slightly improve the error performance of the lightweight student model in terms of the overall metric, but the time-interval benefits were not completely consistent, providing a feasible direction for model compression deployment in edge-side or resource-constrained environments. Fifth, the prediction results of multi-time distributed photovoltaic could provide quantitative basis for power system dispatching, renewable energy consumption, standby arrangement and operational risk perception, which was of practical significance for safe and stable operation in the context of high proportion of new energy access.

5. Conclusions and Future Work

The study focused on the problem of distributed photovoltaic power prediction driven by multi-source heterogeneous data, presenting grid-aware MGG-LSTM modeling methods and further introducing output knowledge distillation analysis for lightweight deployment. The method took historical photovoltaic power, supplementary dynamic variables, meteorological statistics, time features, and differential features as dynamic inputs, node attributes and edge attributes as static graph inputs, extracted multi-granularity grid-aware graph context through three-branch graph encoding. It simultaneously injected the graph context into the LSTM initial state and time-step dynamic representation, finally combined with shared output heads, time-interval output heads, and persistent jump-link branches. Then, the multi-interval distributed photovoltaic power prediction for the next 1 hour, 6 hours, and 24 hours was completed. Overall, the method achieved the unification of multi-source heterogeneous feature fusion and grid perception modeling.

The experimental results showed that in the existing comparison models, the MAE, RMSE and MAPE of MGG-LSTM in the 1-hour prediction reached 25.31 MW, 56.43 MW and 1.48%. MAE and

MAPE at 6 hours were 65.68 MW and 2.78%, and at 24 hours were 96.77 MW and 3.66%. Overall MAE and Overall MAPE of MGG-LSTM were 62.59 MW and 2.64% in terms of the combined indicators. These results suggested that the combination of three-branch graph context extraction, conditional time series modeling, and continuous jump-connection correction helped improve the overall prediction performance of distributed photovoltaic grid-connected nodes at different prediction time intervals. The knowledge distillation experiment showed that the number of parameters of the lightweight student model decreased from 319.7K of the teacher model to 44.9K, with a significant reduction in parameter size. When comparing Student models, the Overall MAE, RMSE and MAPE of Student OutputKD were 77.29 MW, 174.82 MW and 3.94%, slightly better than the Student Baseline's 78.91 MW, 176.46 MW, and 4.00%. This indicated that output distillation could improve the performance of the lightweight student model in the overall metric, but its benefits were not exactly the same in different prediction time intervals and different error metrics. Therefore, it was more suitable as a compressed deployment solution in resource-constrained scenarios rather than a complete replacement for the high-precision prediction ability of the teacher model.

From an engineering application perspective, the proposed method in the study could provide more reliable predictive support for short-term scheduling, intraday rolling optimization, standby scheduling, photovoltaic consumption analysis, and operational risk perception. Especially in multi-node distributed photovoltaic access scenarios, explicitly incorporating grid connection relationships and their physical attributes into the modeling process helped enhance the model's ability to express coupling relationships between nodes and power propagation characteristics. Output distillation experiments also provided a feasible reference for lightweight deployments on the edge side or in resource-constrained environments. For actual utility deployment, integrating the model into existing EMS and SCADA software relies on containerized inference engines embedded in the application layer. This module regularly obtains telemetry and meteorological inputs, performs multi view forecasting, applies physical boundary constraints, and directly pushes the restricted power curve to the automatic power generation control and economic dispatch module through standard OPC or RESTful APIs.

There are still several limitations in the methodology and practical application of the research. Firstly, although the assumption of static graph structure simplifies operations and ensures the stability of macroscopic spatial coupling, it ignores the objective dynamic switching operations and topology reconstruction in real distribution networks. Secondly, the knowledge distillation framework exposes a significant negative transfer trade-off in the 24-hour long-term forecasting perspective: limited by the parameter capacity of the student model and the amplification of the cumulative error of the teacher model, the relative error of the student model in long-term forecasting has degraded to a certain extent. Finally, although the study employed a seasonal layered rolling cross validation scheme to alleviate time series offset, the framework's universality and generalization ability under different utility grid topologies still require more extensive benchmark testing to ensure academic reproducibility under diverse operating conditions. Future research can be conducted in the following directions: studying dynamic topology or adaptive graph learning mechanisms to enhance the expression of time-varying spatial relationships; Introducing uncertainty and probability prediction frameworks to provide risk information for scheduling decisions; Optimize the weight of knowledge distillation loss to eliminate long-term negative transfer; And conduct empirical measurements of hardware level inference latency and power consumption on actual physical edge devices.

Acknowledgements

The research is supported by: Research on the Impact of Distributed Photovoltaics on Distribution Network Stability under Extreme Climate Conditions (SGXT0000SJJS2600061).

References

[1] Shen Q, Zhang Y, Xi Y. Deep learning-based investment risk assessment model for distributed photovoltaic projects. Journal of Advanced Computing Systems. 2024;4(3):31-46. https://doi.org/10.69987/JACS.2024.40303

[2] Yue H, Ali MM, Lin Y, Liu H. Ultra-short-term forecasting of large distributed solar PV fleets using sparse smart inverter data. IEEE Transactions on Sustainable Energy. 2024;15(3):1968-1980. https://doi.org/10.1109/TSTE.2024.3390578
[3] Abdelsattar M, AbdelMoety A, Emad-Eldeen A. Advanced machine learning techniques for predicting power generation and fault detection in solar photovoltaic systems. Neural Computing and Applications. 2025;37(15):8825-8844. https://doi.org/10.1007/s00521-025-10870-8
[4] Lei J, Gao S, Wei X, Huang T, Shahidehpour M, Contreras J, Tang Q. Peer-to-Peer shareholding strategy for distributed photovoltaic systems to curb energy poverty. IEEE Transactions on Power Systems. 2024;40(4):2832-2844. https://doi.org/10.1109/TPWRS.2024.3506823
[5] Wang Y, Fu W, Chen J, Wang J, Zhen Z, Wang F, Lv Y. Spatiotemporal federated learning based regional distributed PV ultra-short-term power forecasting method. IEEE Transactions on Industry Applications. 2024;60(5):7413-7425. https://doi.org/10.1109/TIA.2024.3403514
[6] Song K, Kim M, Kim H. Graph-based large scale probabilistic pv power forecasting insensitive to space-time missing data. IEEE Transactions on Sustainable Energy. 2024;16(1):160-173. https://doi.org/10.1109/TSTE.2024.3377750
[7] Souhe FGY, Mbey CF, Kakeu VJF, Meyo AE, Boum AT. Optimized forecasting of photovoltaic power generation using hybrid deep learning model based on GRU and SVM. Electrical Engineering. 2024;106(6):7879-7898. https://doi.org/10.1007/s00202-024-02388-7
[8] Wu S, Guo H, Zhang X, Wang F. Short-Term Photovoltaic Power Prediction Based on CEEMDAN and Hybrid Neural Networks. IEEE Journal of Photovoltaics. 2024;14(6):960-969. https://doi.org/10.1109/JPHOTOV.2024.3385746
[9] Boucetta LN, Amrane Y, Arezki S. Wind power forecasting using a GRU attention model for efficient energy management systems. Electrical Engineering. 2025;107(3):2595-2620. https://doi.org/10.1007/s00202-024-02755-4
[10] Ma Y, Li F, Zhang H, Fu G, Yi M. Two-stage photovoltaic power forecasting method with an optimized transformer. Global Energy Interconnection. 2024;7(6):812-824. https://doi.org/10.1016/j.gloei.2024.08.010
[11] Zarghami M, Niknam T, Aghaei J, Nezhad AH. Concurrent PV production and consumption load forecasting using CT-Transformer deep learning to estimate energy system flexibility. IET Renewable Power Generation. 2024;18(13):2139-2161. doi: https://doi.org/10.1049/rpg2.13110
[12] Heng S. Enhanced multi-energy load forecasting via multi-task learning and GRU-attention networks in integrated energy systems. Electrical Engineering. 2025;107(6):7673-7683. https://doi.org/10.1007/s00202-025-02941-8
[13] Liu Z, Qian R, Jin X, Zhao H, Li H, Hu D, Hu H. Multi-source heterogeneous data fusion technology for electric power based on big data mining. Journal of Computational Methods in Sciences and Engineering. 2024;24(6):3366-3380. https://doi.org/10.3233/JCM-247510
[14] Wang Y, Li D, Ma W, Zhang X, Zhu H. A New Modeling Method for Meteorological Information of Regional Distributed Photovoltaic Power Generation Based on Multi-Source Information Fusion. Energy Science & Engineering. 2025;13(8):4211-4229. https://doi.org/10.1002/ese3.2031
[15] Xu Q, Pang Y, Zhou X, Liu Y. PIGAT: Physics-informed graph attention transformer for air traffic state prediction. IEEE Transactions on Intelligent Transportation Systems. 2024;25(9):12561-12577. https://doi.org/10.1109/TITS.2024.3387814
[16] Alomari A, et al. Estimation of daily photovoltaic power one day ahead with hybrid Deep Learning and Machine Learning models. Energy Science & Engineering. 2025;13(4):1478-1491. https://doi.org/10.1002/ese3.1952
[17] Ma J, Zhao M, Shen W, Yang Z, Yu X, Lu S. Photovoltaic Power Prediction Based on NRS-PCC Feature Selection and Multi-Scale CNN-LSTM Network. International Journal of Web Services Research. 2024;21(1):1-15. https://doi.org/10.4018/IJWSR.350325
[18] Zhu H, Zhang X, Wang Y, Ding H. A regional distributed photovoltaic power forecasting method based on cluster division and selection of representative plants. Energy Science & Engineering. 2025;13(9):4314-4329. https://doi.org/10.1002/ese3.2052
[19] Wang Y, Fu W, Zhang X, Zhen Z, Wang F. Dynamic directed graph convolution network based ultra-short-term forecasting method of distributed photovoltaic power to enhance the resilience and flexibility of distribution network. IET Generation, Transmission & Distribution. 2024;18(2):337-352. https://doi.org/10.1049/gtd2.13083
[20] Abdelkader D, Fouzi H, Belkacem K, Ying S. Graph neural networks-based spatiotemporal prediction of photovoltaic power: a comparative study. Neural Computing and Applications. 2025;37(6):4769-4795. https://doi.org/10.1007/s00521-024-10659-1
[21] Jiang Z, Zhang Z, He X, Li Y, Yuan H. Efficient and accurate TEC modeling and prediction approach with random forest and Bi-LSTM for large-scale region. Advances in Space Research. 2024;73(1):650-662. https://doi.org/10.1016/j.asr.2023.09.049

[22] Buratto WG, Muniz RN, Nied A, Barros CFDOD, Cardoso R, Gonzalez GV. Wavelet CNN-LSTM time series forecasting of electricity power generation considering biomass thermal systems. IET Generation, Transmission & Distribution. 2024;18(21):3437-3451. https://doi.org/10.1049/gtd2.13292

[23] Liu ZX, Singh P, Yang WR, Huang WT. LSTM neural network for current prediction and harmonic analysis in solar power system. IET Conference Proceedings CP893. 2024;2024(22):23-24. https://doi.org/10.1049/icp.2024.2810

[24] Fu X, Zhang C, Xu Y, Zhang Y, Sun H. Statistical machine learning for power flow analysis considering the influence of weather factors on photovoltaic power generation. IEEE Transactions on Neural Networks and Learning Systems. 2024;36(3):5348-5362. https://doi.org/10.1109/TNNLS.2024.3414902