EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

A Transition-Aware Power Discriminative Network for Short-Term Photovoltaic Power Forecasting

Li Zhang[1], Tiancheng Shi[1], Xueao Qiu[1,*], Yutong Ye[1] and Ziran Jin[1]

[1]Economic and Technical Research Institute, State Grid Anhui Electric Power Co., Ltd., Hefei 230022, China

Abstract

INTRODUCTION: Short-term photovoltaic (PV) power forecasting is essential for grid dispatching, electricity market trading, energy storage configuration, and microgrid energy management. However, transition events and nonlinear meteorological-power trajectory patterns make accurate multi-step forecasting challenging. OBJECTIVES: This study aims to improve short-term PV power forecasting by enhancing sparse transition periods without manual weather labels and by constraining the global shape of future power trajectories. METHODS: A transition-aware power discriminative network (TAPD) is proposed. TAPD consists of a transition-aware enhancement module (TAE) and a predictor-discriminator joint modeling module (PDJM). TAE estimates time-wise transition scores from power-irradiance response features and adaptively enhances encoded representations. PDJM combines a coarse-to-fine horizon-aware predictor with a conditional power-sequence discriminator to jointly constrain point-wise accuracy, frequency-domain structure, adjacent-step relations, and overall trajectory shape. RESULTS: Experiments on three real PV station datasets released by the State Grid show that TAPD outperforms state-of-the-art forecasting models. Averaged across 1 h, 3 h, 6 h, and 12 h prediction horizons, TAPD reduces mean absolute error (MAE) by 10.09%, 6.74%, and 10.31% on Stations A, B, and C, respectively. CONCLUSION: TAPD effectively captures rapid local power changes while maintaining globally reasonable trajectory shapes, thereby improving the accuracy and robustness of short-term PV power forecasting.

Keywords: photovoltaic power forecasting, transition-aware learning, power discriminative, multi-step time series forecasting, deep learning

Received on 27 June 2026, accepted on 21 September 2026, published on 24 November

Copyright © 2026 Li Zhang *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/ew.15645

1. Introduction

Driven by both the global low-carbon transformation of the energy structure and the "dual-carbon" strategic goals, photovoltaic (PV) power generation, with its significant advantages of cleanliness, safety, and easy distributed deployment, has become one of the fastest-growing renewable energy sources in new power systems [1]. Accurate short-term PV power prediction is an indispensable part of grid dispatching, electricity market trading, energy storage capacity configuration, and microgrid energy management. However, PV power output is strongly coupled with solar irradiance, cloud motion, aerosol distribution, and atmospheric conditions, leading to pronounced intermittency, volatility, and uncertainty [2, 3]. Although physical models [4–6] and statistical forecasting methods [7–9] offer a certain degree of interpretability, they often fail to fully characterize the non-stationary mapping between meteorological variables and PV power under complex weather conditions, resulting in degraded short-term forecasting accuracy.

In this context, deep learning methods have gradually become an important technical route for short-term PV power forecasting, including weather-aware hybrid deep learning methods [10–12] and end-to-end deep learning methods [13–15]. However, the non-stationarity of the environment in PV power generation scenarios and the consistency of the prediction horizon cause the above methods to suffer reduced power forecasting accuracy in practical applications, with the following challenges. First, PV power sequences

*Corresponding author. Email: qxa9939@163.com

frequently contain transition events during sunrise and sunset, cloud occlusion, and rapid irradiance variation, where power and irradiance fluctuate sharply over a short period [12, 16]. The occurrence times of transition events are usually sparse and unevenly distributed, and public labels are lacking, making it difficult to directly train event detectors in a supervised manner [17]. Applying the same nonlinear enhancement to all time steps may introduce additional noise during stationary periods and impair overall performance. Second, short-term forecasting needs to provide future multi-step power trajectories while ensuring that the trajectory shape satisfies the physical rationality required for dispatching decisions. Most existing methods use point-wise regression loss as the single optimization objective, focusing only on point-value errors while ignoring sequence trends, spectral structure, and overall shape constraints.

To address these challenges, transition-aware power discriminative network (TAPD) is proposed for short-term PV power forecasting. To address the first challenge, a transition-aware enhancement (TAE) module is designed to characterize event signals according to the instantaneous response relationship between PV power and irradiance. The module enables TAPD to selectively enhance sparse transition periods without manually annotated weather labels and to maintain prediction stability during stationary periods. To address the second challenge, a predictor-discriminator joint modeling (PDJM) module is proposed to incorporate global trajectory-shape assessment into the multi-step forecasting process. This module preserves reasonable trends, spectral structures, and adjacent-step relations, thereby reducing the impact of long-horizon error accumulation on downstream grid dispatch and reserve-capacity decisions.

Specifically, we propose a TAPD which consists of two key modules: TAE and PDJM. TAPD maps multivariate historical observations to future multi-step power forecasts in an end-to-end manner. The main contributions of this study are summarized as follows:

(i) TAE is proposed to identify and selectively enhance transition events without any manual labels. By introducing power-irradiance differences and instantaneous response-ratio derived features as inputs for transition-score estimation, TAE adaptively modulates encoded representations through stationary/transition dual-branch projection and temporal gating, and uses pseudo-label auxiliary supervision to achieve end-to-end training of unsupervised estimation.

(ii) PDJM is developed to suppress error accumulation and trajectory-shape distortion in long-horizon multi-step forecasting. By constructing a coarse-to-fine horizon-aware predictor and a conditional power-sequence discriminator as an adversarial structure, PDJM jointly constrains future trajectories from four dimensions: point-wise accuracy, frequency-domain structure, adjacent-step relations, and overall shape.

(iii) Comprehensive experiments are conducted on three real PV station datasets from the State Grid dispatching AI innovation competition. Compared with state-of-the-art (SOTA) methods, TAPD reduces the average MAE by 10.09%, 6.74%, and 10.31% on Stations A, B, and C, respectively, averaged across four prediction horizons (1 h, 3 h, 6 h, and 12 h).

2. Related Work

2.1. Weather-Aware Hybrid Deep Learning Methods

Weather-aware hybrid deep learning methods explicitly incorporate meteorological priors into data preprocessing or model structures to better characterize weather-power coupling and improve forecasting accuracy under complex meteorological conditions. Existing methods can be broadly divided into weather-classification-based methods [18–21] and multimodal-fusion-based methods [2, 3, 22, 23]. The former usually cluster or manually classify historical samples according to weather characteristics and then train dedicated predictors for different categories. These methods are intuitive and structurally controllable, but their performance depends heavily on classification quality and cluster-number selection. They also implicitly require sufficient samples for each weather type. When transition weather or rare weather patterns occur, the generalization capability of submodels can decline markedly, and the classification process often depends on external weather interpretation or manual labels. By contrast, multimodal fusion methods jointly feed heterogeneous meteorological information and historical power sequences into neural networks, and may use sky images to characterize high-frequency perturbations caused by cloud occlusion. However, when multimodal data contain time delays, missing values, or numerical weather prediction errors, learning stable and reliable fusion weights remains challenging.

2.2. End-to-End Deep Learning Methods

Unlike weather-aware hybrid methods that embed meteorological priors into preprocessing stages, end-to-end deep learning methods directly learn the mapping from historical multivariate observations to future trajectories, thereby reducing dependence on weather labels and external multimodal data. Existing studies mainly focus on data augmentation [24, 25] and long-range dependency modeling [13–15]. Considering the sparsity of transition events, data augmentation methods generate weather data and corresponding labels for sparse data categories through adversarial data generation, thereby increasing sample diversity and model

adaptability under different weather conditions. However, such methods are often an expansion of existing transition events. When transition patterns not covered by the dataset are encountered, model performance may decline. Long-range dependency modeling methods are usually based on Transformer to improve modeling capability for long-range dependencies and multivariate sparse coupling [26]. In PV scenarios, this class of methods further suppresses low-signal-to-noise frequency bands through spectral filtering, or reshapes irradiance-temperature-power physical-variable combinations at the input side to enhance robustness against meteorology-dominated disturbances [14, 15]. Although these methods substantially improve long-horizon modeling capability, most still rely primarily on point-wise losses and lack direct constraints on the global shape of future trajectories, leaving multi-step error accumulation unresolved.

Compared with these methods, the proposed TAPD differs in two main aspects. First, it constructs a label-free adaptive transition-event awareness mechanism to reduce dependence on manually annotated weather labels. Second, it introduces a conditional power-sequence discriminator to constrain the global morphological rationality of multi-step prediction trajectories.

3. Problem Formulation

This section formally describes the short-term PV power forecasting (ST-PVPF) task and formulates it as a conditional multi-step time-series regression problem. A forecasting model receives a historical multivariate observation sequence $\mathbf{X} \in \mathbb{R}^{L\times D}$, where L denotes the historical window length and D denotes the number of variables. At each time step, the observation vector contains measured PV power and five meteorological variables: global horizontal irradiance (GHI), direct normal irradiance (DNI), total solar irradiance (TSI), air temperature (AT), and atmospheric pressure (AP). The model outputs a future power sequence $\hat{\mathbf{y}} \in \mathbb{R}^{H}$, where H is the forecasting horizon. Therefore, ST-PVPF can be expressed as

$$\hat{\mathbf{y}} = f_\theta(\mathbf{X}), \quad \mathbf{X} \in \mathbb{R}^{L\times D}, \quad \hat{\mathbf{y}} \in \mathbb{R}^{H}. \tag{1}$$

Given a multivariate observation stream recorded by a PV station at a fixed sampling interval, supervised samples are constructed using a sliding window. At a prediction time t, the model looks back L steps and predicts H future steps:

$$(\mathbf{X}_t, \mathbf{y}_t) = ([\mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_t], [p_{t+1}, \ldots, p_{t+H}]). \tag{2}$$

All valid sliding windows form the training set $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^{N}$. In this study, the input length is set to 96 time points, corresponding to 24 h under a 15 min sampling interval. The model uses the meteorological features and historical generation data from the previous 96 time points to predict PV power over future horizons of 1 h, 3 h, 6 h, and 12 h.

The objective of ST-PVPF is to learn a forecasting model $f_\theta(\cdot)$ parameterized by θ that minimizes empirical risk on $\mathcal{D}$:

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} L\left(f_\theta(\mathbf{X}_i), \mathbf{y}_i\right), \tag{3}$$

where $L(\cdot,\cdot)$ denotes the loss function. However, the above objective is not directly tractable for optimization. In practice, we approximate θ^* with a set of differentiable surrogate losses, including a time–frequency hybrid prediction loss L_{pred}, a coarse-prediction loss L_{coarse}, a first-order relation loss L_{rel}, a transition-classification auxiliary loss L_{cls}, and a conditional adversarial loss L_{adv}. The overall optimization objective is

$$\begin{aligned}\theta^* = \arg\min_\theta \big(&L_{\text{pred}} + \lambda_{\text{coarse}} L_{\text{coarse}} + \lambda_{\text{r}} L_{\text{rel}} \\ &+ \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{adv}} L_{\text{adv}}\big),\end{aligned} \tag{4}$$

where λ_{coarse}, λ_{r}, λ_{cls}, and λ_{adv} denote the weighting coefficients.

4. Methodology

The overall architecture of TAPD is shown in Figure 1. TAPD consists of three stages: First, the data encoding stage maps historical multivariate observations into a latent representation space, where a Transformer and a unidirectional LSTM are connected in series to capture both long-range dependencies and local temporal dynamics. Second, the TAE explicitly constructs power-irradiance difference and instantaneous response-ratio derived variables in the raw data space, estimates time-wise transition scores, and selectively enhances encoded representations via a gating mechanism without requiring manually annotated event labels. Third, the PDJM builds a horizon-aware predictor and a conditional power-sequence discriminator on top of the enhanced sequence and fused context.

4.1. Encoding into latent representation space

The data encoder aims to map the raw multichannel observation sequence into a latent representation space that preserves both long-range dependencies and short-term temporal details. Since a single architecture is difficult to use for capturing global interactions and causal temporal dynamics simultaneously, this module adopts a serial backbone consisting of linear projection, sinusoidal positional encoding, a pre-layer-normalization Transformer encoder, and a multilayer unidirectional LSTM. The raw input $\mathbf{X}$ is first projected

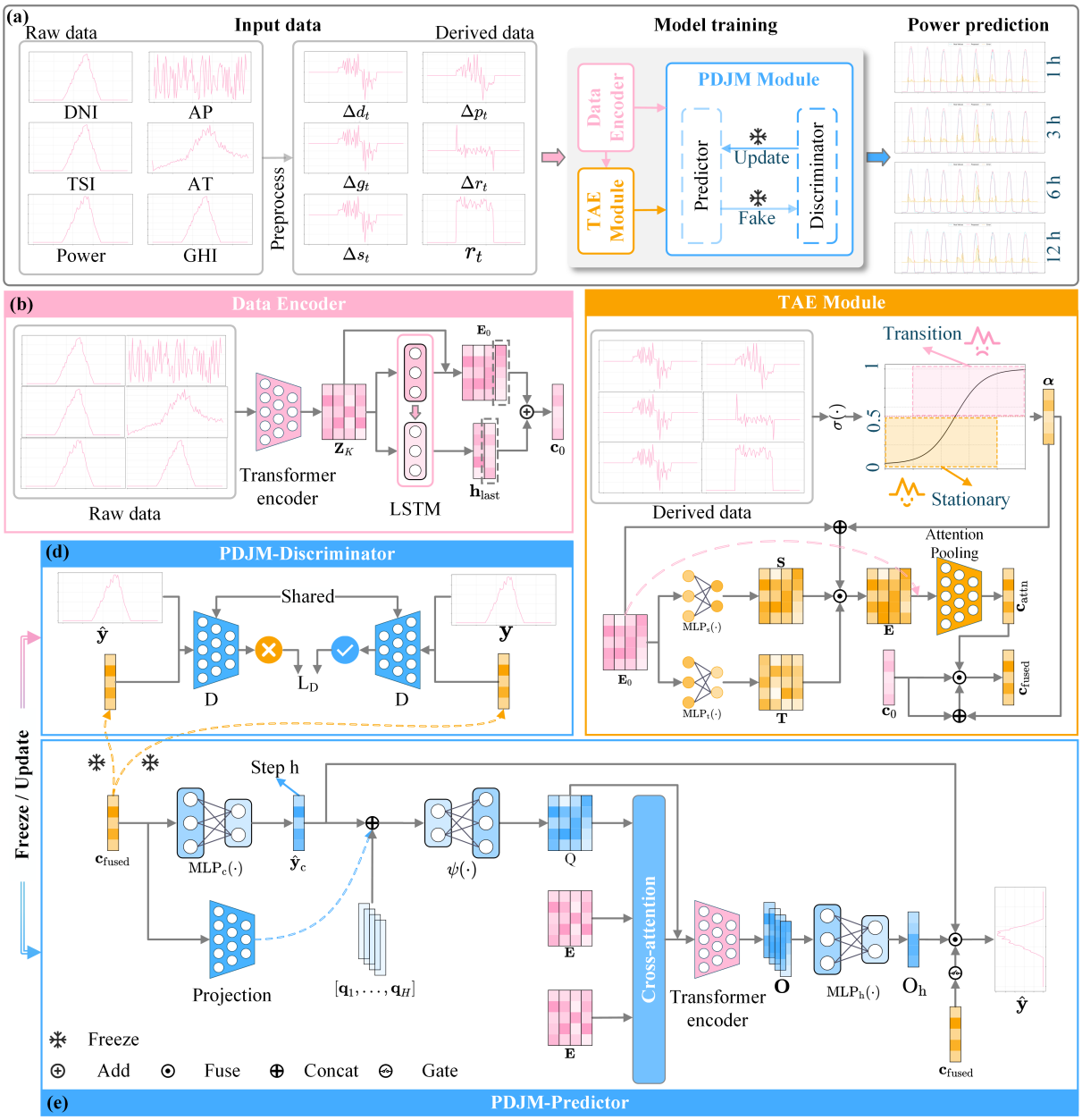

Figure 1. Overall architecture of TAPD
(a) Flow chart of TAPD; (b) Encoding into latent representation space; (c) TAE module; (d) PDJM-discriminator; and (e) PDJM-predictor

into a hidden space to obtain the initial embedded sequence $\mathbf{Z}_0$:

$$\mathbf{Z}_0 = \mathrm{LN}\left(\mathbf{X}\mathbf{W}_e + \mathbf{b}_e\right) + \mathbf{P}, \tag{5}$$

where $\mathbf{W}_e$ and $\mathbf{b}_e$ are the learnable projection weight and bias, $\mathrm{LN}(\cdot)$ denotes layer normalization, and $\mathbf{P}$ is a fixed sinusoidal positional encoding. For the ℓ-th pre-LN Transformer layer, multi-head self-attention $\mathrm{MHA}(\cdot)$ and feed-forward transformation $\mathrm{FFN}(\cdot)$ are computed as:

$$\begin{aligned} \tilde{\mathbf{Z}}_\ell &= \mathbf{Z}_{\ell-1} + \mathrm{MHA}\left(\mathrm{LN}(\mathbf{Z}_{\ell-1})\right), \\ \mathbf{Z}_\ell &= \tilde{\mathbf{Z}}_\ell + \mathrm{FFN}\left(\mathrm{LN}(\tilde{\mathbf{Z}}_\ell)\right), \end{aligned} \tag{6}$$

where $\tilde{\mathbf{Z}}_\ell$ is the intermediate representation after self-attention, and $\mathbf{Z}_\ell$ is the output of the ℓ-th layer.

After stacking K Transformer layers, the output $\mathbf{Z}_K$ is passed through a multilayer unidirectional LSTM to strengthen temporal causality:

$$\mathbf{H}_{\mathrm{lstm}}, \mathbf{h}_{\mathrm{last}} = \mathrm{LSTM}(\mathbf{Z}_K), \tag{7}$$

where $\mathbf{H}_{\mathrm{lstm}}$ is the full sequence of LSTM hidden states and $\mathbf{h}_{\mathrm{last}}$ is the hidden state at the final time step.

The Transformer representation and LSTM output are fused by residual addition and layer normalization:

$$\mathbf{E}_0 = \mathrm{LN}\left(\mathbf{Z}_K + \mathbf{H}_{\mathrm{lstm}}\right), \quad \mathbf{c}_0 = \phi\left([\mathbf{E}_0^{(L)}; \mathbf{h}_{\mathrm{last}}]\right), \tag{8}$$

where $\mathbf{E}_0$ denotes the encoded sequence, $\mathbf{E}_0^{(L)}$ is the last-step representation, $\mathbf{c}_0$ is the global context vector, $\phi(\cdot)$ a multilayer perceptron, and $[\cdot;\cdot]$ the concatenation operation along the feature dimension.

4.2. Transition-Aware Enhancement Module

To identify and enhance transition events without manual labels, the TAE module follows a "physical signature plus gated amplification" design. The core motivation is that the instantaneous response ratio between power and irradiance is relatively stable during stationary periods, whereas power differences and response ratios deviate sharply around transition events such as sunrise/sunset, cloud shadowing, and rapid irradiance changes. The module therefore constructs interpretable derived variables in the raw data space and estimates a time-wise transition score using a shallow multilayer perceptron (MLP).

Let p_t, g_t, d_t, and s_t denote power, GHI, DNI, and TSI at time step t, respectively. First-order differences are defined as:

$$\begin{aligned} \Delta p_t &= p_t - p_{t-1}, \quad \Delta g_t = g_t - g_{t-1}, \\ \Delta d_t &= d_t - d_{t-1}, \quad \Delta s_t = s_t - s_{t-1}, \\ \Delta r_t &= r_t - r_{t-1}. \end{aligned} \tag{9}$$

The instantaneous response ratio between power and irradiance is defined by

$$r_t = \frac{p_t}{s_t + \epsilon}, \tag{10}$$

where ϵ is a small constant for numerical stability, and Δr_t in Equation (9) is the first-order difference of r_t. The derived feature vector is

$$\mathbf{u}_t = [\Delta p_t, \Delta g_t, \Delta d_t, \Delta s_t, r_t, \Delta r_t]. \tag{11}$$

A shallow MLP followed by a sigmoid function $\sigma(\cdot)$ estimates the transition score:

$$\alpha_t = \sigma\left(\mathrm{MLP}(\mathbf{u}_t)\right), \quad \alpha_t \in [0,1], \tag{12}$$

where $\alpha_t \in [0,1]$ is the transition score at step t.

The encoded sequence is projected into stationary and transition branches to obtain the stationary-branch and transition-branch representations $\mathbf{S}$ and $\mathbf{T}$:

$$\mathbf{S} = \mathrm{MLP}_{\mathrm{s}}(\mathbf{E}_0), \quad \mathbf{T} = \mathrm{MLP}_{\mathrm{t}}(\mathbf{E}_0), \tag{13}$$

where $\mathrm{MLP}_{\mathrm{s}}(\cdot)$ and $\mathrm{MLP}_{\mathrm{t}}(\cdot)$ are their respective projection networks.

The temporal gate is computed by concatenating encoded features with transition scores:

$$\mathbf{G} = \sigma\left(\mathrm{MLP}_{\mathrm{g}}([\mathbf{E}_0; \boldsymbol{\alpha}])\right), \tag{14}$$

where $\mathbf{G}$ is the temporal gate, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]$ is the sequence of transition scores, and $\mathrm{MLP}_{\mathrm{g}}(\cdot)$ is the gating network.

The enhanced sequence is then obtained through residual gated modulation:

$$\mathbf{E} = \mathbf{E}_0 + (1 - \boldsymbol{\alpha}) \odot (1 - \mathbf{G}) \odot \mathbf{S} + \boldsymbol{\alpha} \odot \mathbf{G} \odot \mathbf{T}, \tag{15}$$

where $\mathbf{E}$ is the enhanced sequence and $\odot$ denotes element-wise (Hadamard) multiplication.

This design keeps both regimes controllable. When $\alpha_t \to 0$, the stationary branch dominates and avoids over-amplification during stationary periods. When $\alpha_t \to 1$, the transition branch is activated to amplify rapidly changing segments.

To obtain a length-independent pooled context, additive attention pooling is applied to $\mathbf{E}$:

$$\beta_t = \frac{\exp(\mathbf{v}^\top \tanh(\mathbf{W}\mathbf{E}_t))}{\sum_{j=1}^{L} \exp(\mathbf{v}^\top \tanh(\mathbf{W}\mathbf{E}_j))}, \quad \mathbf{c}_{\mathrm{att}} = \sum_{t=1}^{L} \beta_t \mathbf{E}_t, \tag{16}$$

where $\mathbf{v}$ and $\mathbf{W}$ are learnable attention parameters, $\mathbf{E}_t$ is the enhanced representation at step t, β_t is the normalized attention weight, and $\mathbf{c}_{\mathrm{att}}$ is the resulting pooled context vector.

The final fused context $\mathbf{c}_{\mathrm{fused}}$ is generated as

$$\eta = \sigma(\mathbf{w}_\eta^\top [\mathbf{c}_0; \bar{\alpha}] + b_\eta), \quad \mathbf{c}_{\mathrm{fused}} = \eta \mathbf{c}_0 + (1 - \eta)\mathbf{c}_{\mathrm{att}}, \tag{17}$$

where $\bar{\alpha}$ is the mean transition score over the window, $\mathbf{w}_\eta$ and b_η are learnable parameters, $\eta \in [0,1]$ is the fusion gate, and $\mathbf{c}_{\mathrm{fused}}$ is the fused context vector.

4.3. Predictor-Discriminator Joint Modeling Module

To address trajectory-shape distortion in multi-step forecasting, PDJM constructs a predictor-discriminator joint modeling architecture. Directly minimizing point-wise error often leads to over-smoothed long-horizon prediction curves and loss of high-frequency details. Therefore, two mutually constrained subnetworks are built on top of $\mathbf{E}$ and $\mathbf{c}_{\mathrm{fused}}$: a horizon-aware predictor G generates future power values in a coarse-to-fine manner, while a power-sequence discriminator D distinguishes real future trajectories from generated trajectories under the same context. The two networks form a conditional adversarial game, imposing global constraints on future trajectory shapes beyond point-wise accuracy.

The predictor first produces a coarse prediction from the fused context:

$$\hat{\mathbf{y}}_{\mathrm{c}} = \mathrm{MLP}_{\mathrm{c}}(\mathbf{c}_{\mathrm{fused}}), \tag{18}$$

where $\hat{\mathbf{y}}_c \in \mathbb{R}^H$ is the coarse prediction and $\mathrm{MLP}_c(\cdot)$ is the coarse-prediction head.

Learnable horizon query embeddings are then used to inject context priors and coarse prediction priors into each future step. Multi-head cross-attention is performed with the horizon queries as queries and the enhanced sequence as keys and values:

$$\mathbf{Q} = \psi([\mathbf{q}_1, \ldots, \mathbf{q}_H], \mathbf{c}_{\text{fused}}, \hat{\mathbf{y}}_c), \quad \mathbf{O} = \mathrm{MHA}(\mathbf{Q}, \mathbf{E}, \mathbf{E}), \tag{19}$$

where $\mathbf{q}_1, \ldots, \mathbf{q}_H$ are learnable horizon query embeddings, $\psi(\cdot)$ is the mapping that injects the context and coarse-prediction priors into the queries, $\mathbf{Q}$ is the resulting query matrix, and $\mathbf{O}$ is the cross-attention output obtained with $\mathbf{E}$ serving as keys and values.

The refined prediction is generated by adding a gated residual offset to the coarse prediction:

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_c + \mathbf{g}_h \odot \mathbf{o}_h, \quad \mathbf{g}_h = \sigma(\mathrm{MLP}_h(\mathbf{c}_{\text{fused}})), \tag{20}$$

where $\mathbf{o}_h$ is the residual offset mapped from the cross-attention output $\mathbf{O}$, $\mathbf{g}_h \in [0,1]^H$ is the horizon-wise gate produced by $\mathrm{MLP}_h(\cdot)$, and $\hat{\mathbf{y}}$ is the final refined prediction.

For the discriminator, either the real future sequence $\mathbf{y}$ or the generated sequence $\hat{\mathbf{y}}$ is combined with the conditional context. Step-wise tokens explicitly integrate values, first-order differences, context, and horizon embeddings:

$$\mathbf{z}_h = \phi_D([y_h, \Delta y_h, \mathbf{c}_{\text{fused}}, \mathbf{q}_h]), \quad h = 1, \ldots, H, \tag{21}$$

where y_h and Δy_h are the value and its first-order difference at step h of the input sequence (either the real $\mathbf{y}$ or the generated $\hat{\mathbf{y}}$), $\phi_D(\cdot)$ is the token-embedding function, and $\mathbf{z}_h$ is the resulting step-wise token.

After prepending a learnable embedding, the sequence is encoded by a Transformer, and the embedding is mapped to a scalar logit, then, the conditional discriminator score $D(\cdot \mid \mathbf{c}_{\text{fused}})$ can be obtained as:

$$\begin{aligned} \mathbf{Z}_D &= [\mathbf{z}_{\text{cls}}, \mathbf{z}_1, \ldots, \mathbf{z}_H], \\ D(\mathbf{y} \mid \mathbf{c}_{\text{fused}}) &= \mathbf{w}_D^\top \mathrm{Transformer}_D(\mathbf{Z}_D)_{\text{cls}} + b_D, \end{aligned} \tag{22}$$

where $\mathbf{z}_{\text{cls}}$ is the learnable embedding, $(\cdot)_{\text{cls}}$ denotes its encoded output representation, $\mathbf{w}_D$ and b_D are the output-layer weight and bias.

The prediction loss combines a time-domain Huber loss and a frequency-domain loss computed on the real-valued fast Fourier transform spectrum:

$$L_{\text{pred}} = \lambda_{\text{time}} \mathrm{Huber}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_{\text{freq}} \|\mathcal{F}(\hat{\mathbf{y}}) - \mathcal{F}(\mathbf{y})\|_1, \tag{23}$$

where $\mathrm{Huber}(\cdot,\cdot)$ is the Huber loss, $\mathcal{F}(\cdot)$ denotes the real-valued fast Fourier transform, $\|\cdot\|_1$ is the ℓ_1 norm, and λ_{time} and λ_{freq} are the time- and frequency-domain weighting coefficients.

To provide an effective prior for the refinement stage, the coarse prediction is additionally supervised by a coarse-prediction loss L_{coarse}:

$$L_{\text{coarse}} = \mathrm{MSE}(\hat{\mathbf{y}}_c, \mathbf{y}), \tag{24}$$

where $\hat{\mathbf{y}}_c$ is the coarse prediction.

A first-order relation loss is used to constrain adjacent-step differences:

$$L_{\text{rel}} = \|\Delta\hat{\mathbf{y}} - \Delta\mathbf{y}\|_1, \tag{25}$$

where $\Delta\hat{\mathbf{y}}$ and $\Delta\mathbf{y}$ denote the first-order differences of the predicted and ground-truth sequences, respectively.

The transition-classification auxiliary loss L_{cls} supervises the mean transition score using a pseudo-label z derived from historical power, historical irradiance, and future power differences:

$$L_{\text{cls}} = -z \log(\bar{\alpha}) - (1 - z) \log(1 - \bar{\alpha}), \tag{26}$$

where $z \in \{0, 1\}$ is the transition pseudo-label and $\bar{\alpha}$ is the mean transition score.

The conditional adversarial loss L_D follows the standard conditional GAN objective:

$$L_D = -\mathbb{E}[\log D(\mathbf{y} \mid \mathbf{c}_{\text{fused}})] - \mathbb{E}[\log(1 - D(\hat{\mathbf{y}} \mid \mathbf{c}_{\text{fused}}))], \tag{27}$$

where $\mathbb{E}[\cdot]$ denotes the expectation over the sample distribution.

The generator adversarial loss L_{adv} encourages the generated trajectory $\hat{\mathbf{y}}$ to be judged as real by the discriminator.

$$L_{\text{adv}} = -\mathbb{E}[\log D(\hat{\mathbf{y}} \mid \mathbf{c}_{\text{fused}})]. \tag{28}$$

TAPD is trained by alternating optimization between the discriminator and the predictor side. In the discriminator step, only the discriminator parameters are updated by minimizing L_D in Equation (27), while the generated trajectory $\hat{\mathbf{y}}$ and the conditional context $\mathbf{c}_{\text{fused}}$ are detached from the gradient graph so that the predictor is not affected. In the predictor step, the discriminator parameters are frozen, and the encoder, the TAE module, and the predictor are updated by minimizing the overall objective in Equation (4), where L_{adv} is the only term that backpropagates gradients through the discriminator. To mitigate instability at the early stage of GAN training, TAPD further adopts a two-stage training strategy: the predictor is first optimized using only the non-adversarial losses, and the adversarial term is then gradually introduced through warm-up so that adversarial supervision remains a weak auxiliary signal rather than dominating the forecasting task.

Table 1. Descriptive statistics of meteorological variables and power output at the three solar stations

Solar Station	Statistics	TSI (W/m^2)	DNI (W/m^2)	GHI (W/m^2)	AT (°C)	AP (hPa)	Power (MW)
Station A	Mean	266.21	93.26	67.69	13.15	913.37	9.67
	Std.	367.89	200.78	111.20	14.34	8.74	13.71
Station B	Mean	169.30	122.15	78.30	13.70	861.04	19.57
	Std.	248.08	178.99	117.59	12.04	6.15	27.94
Station C	Mean	163.24	142.02	21.22	18.01	956.42	4.23
	Std.	245.39	213.49	31.90	8.56	30.54	6.52

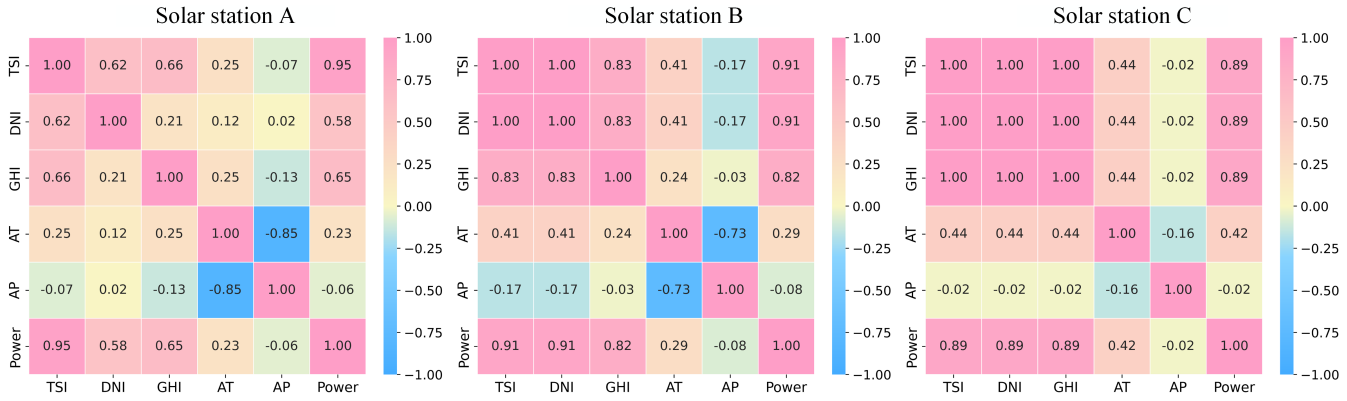

Figure 2. Pearson correlation heatmaps of the three PV stations
Irradiance-related features show strong positive correlations with PV power, whereas AT and AP have weaker linear correlations

5. Experimental Results and Discussion

5.1. Dataset

The experiments use a public dataset from the State Grid dispatching AI innovation competition [27]. The data were collected from three real grid-connected PV stations operated by the State Grid Corporation of China, denoted as Station A, Station B, and Station C. The data span from 1 January 2019 to 31 December 2020 with a sampling interval of 15 min, resulting in 70,176 records for each station.

Each record contains measured power and five meteorological variables: TSI, DNI, GHI, AT, and AP. TSI characterizes solar radiation received at the top of the atmosphere and is mainly driven by Sun-Earth distance and solar activity. DNI represents the radiation flux received per unit area perpendicular to the solar beam, while GHI denotes the short-wave radiation received on a horizontal surface. AT and AP are measured at 1.5 m above ground. The nominal generation output capacities of the three solar stations are 50 MW, 130 MW, and 30 MW, respectively. Table 1 provides the statistical information for the three solar stations, including the mean, standard deviation.

Figure 2 shows the Pearson correlation heatmaps of the three stations. The irradiance-related features (TSI, DNI, and GHI) exhibit strong positive correlations with PV power. For example, the correlation between TSI and power at Station C reaches 0.98. By contrast, AT and AP show weaker linear correlations with power, and the AP correlation is only -0.06. It should be noted that Pearson correlation only characterizes linear relationships, while AT and AP may indirectly affect PV output by modulating irradiance conditions. Therefore, all five meteorological variables are retained as forecasting factors, allowing the model to automatically learn nonlinear interactions between these variables and PV power.

5.2. Evaluation Metrics

To quantitatively evaluate forecasting performance, three standard metrics are introduced from the perspectives of error magnitude, fluctuation stability, and goodness of fit: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). Their definitions are

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|, \tag{29}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{30}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{31}$$

Table 2. Performance comparison of different models at Station A. In each row, the best result is shown in **bold** and the second-best is underlined. The IMP column reports the relative improvement of TAPD over the second-best baseline.

Metric	Horizon	IMP (%)	TAPD	LSTM	FETN	FARM	FE-iTransformer
MAE	4	6.28	**0.0812**	0.0965	0.1084	0.0867	0.1096
	12	10.55	**0.1245**	0.1570	0.1430	0.1392	0.1579
	24	11.86	**0.1600**	0.1895	0.1815	0.2638	0.1871
	48	11.65	**0.1801**	0.2064	0.2039	0.2999	0.2082
RMSE	4	-2.05	0.1944	0.2016	0.2131	**0.1905**	0.2283
	12	1.60	**0.2807**	0.2994	0.2957	0.2853	0.3078
	24	7.11	**0.3239**	0.3487	0.3492	0.4618	0.3579
	48	1.43	**0.3616**	0.3668	0.3885	0.5248	0.3811
R^2	4	-0.16	0.9621	0.9592	0.9544	**0.9636**	0.9477
	12	0.28	**0.9209**	0.9101	0.9123	0.9184	0.9050
	24	1.91	**0.8948**	0.8781	0.8778	0.7862	0.8716
	48	0.44	**0.8690**	0.8652	0.8487	0.7240	0.8545

Table 3. Performance comparison of different models at Station B. In each row, the best result is shown in **bold** and the second-best is underlined. The IMP column reports the relative improvement of TAPD over the second-best baseline.

Metric	Horizon	IMP (%)	TAPD	LSTM	FETN	FARM	FE-iTransformer
MAE	4	15.55	**0.0876**	0.1096	0.1037	0.1068	0.1123
	12	5.22	**0.1315**	0.1709	0.1387	0.1524	0.1555
	24	5.30	**0.1653**	0.2247	0.1745	0.3439	0.1903
	48	0.90	**0.1919**	0.2109	0.1937	0.3199	0.2073
RMSE	4	4.63	**0.2082**	0.2183	0.2259	0.2232	0.2327
	12	3.98	**0.2845**	0.3207	0.2963	0.2990	0.3042
	24	3.36	**0.3373**	0.3839	0.3490	0.5376	0.3552
	48	-0.27	0.3716	**0.3706**	0.3752	0.5603	0.3755
R^2	4	0.49	**0.9529**	0.9482	0.9445	0.9458	0.9411
	12	0.82	**0.9120**	0.8882	0.9046	0.9028	0.8994
	24	1.01	**0.8764**	0.8399	0.8676	0.6859	0.8629
	48	-0.10	0.8498	**0.8506**	0.8469	0.6585	0.8466

Here, N is the number of samples, $\hat{y}_i$ is the predicted value, y_i is the ground truth, and $\bar{y}$ is the mean of the ground-truth sequence. MAE is relatively insensitive to outliers and intuitively reflects average absolute deviation. RMSE assigns larger weights to large errors and is suitable for evaluating error-control capability under strongly fluctuating scenarios. R^2 evaluates the ability of a model to explain variations in the true sequence. Using these three metrics together provides a comprehensive assessment of prediction performance.

5.3. Performance Comparison

To systematically evaluate the multi-step forecasting capability of TAPD, the proposed model is compared with four representative time-series forecasting models: LSTM, FETN [14], FARM [13], and FE-iTransformer [15]. Experiments are conducted on Stations A, B, and C under four horizons $H \in \{4, 12, 24, 48\}$, corresponding to 1 h, 3 h, 6 h, and 12 h. The evaluation metrics are MAE, RMSE, and R^2. The complete results are summarized in Tables 2–4, where the IMP column reports the average relative improvement of TAPD over the second-best baseline.

Overall, TAPD achieves the best results in most of the 36 configurations formed by three datasets, four horizons, and three metrics, and its advantage is particularly evident in MAE. For the longest horizon $H = 48$, TAPD obtains MAE = 0.1801, RMSE = 0.3616, and $R^2 = 0.8690$ at Station A, outperforming all baselines on all three metrics. At Station B, TAPD achieves the best MAE = 0.1919 and remains very close to the best baseline in RMSE and R^2. At Station C, TAPD obtains MAE = 0.1929, RMSE = 0.3893, and $R^2 = 0.8436$, outperforming all baselines. These results indicate that the synergistic design of the horizon-aware predictor and conditional adversarial game effectively suppresses long-horizon error accumulation and generates more stable multi-step trajectories.

Under the short horizon $H = 4$, performance gaps between models become relatively small. For example, at Station A, TAPD obtains RMSE = 0.1944 and $R^2 = 0.9621$, which are close to FARM (0.1905 / 0.9636), while still achieving the best MAE (0.0812

Table 4. Performance comparison of different models at Station C. In each row, the best result is shown in **bold** and the second-best is underlined. The IMP column reports the relative improvement of TAPD over the second-best baseline.

Metric	Horizon	IMP (%)	TAPD	LSTM	FETN	FARM	FE-iTransformer
MAE	4	18.12	**0.0637**	0.0794	0.0845	0.0778	0.1007
	12	9.58	**0.0977**	0.1108	0.1204	0.1081	0.1231
	24	6.84	**0.1503**	0.1677	0.1613	0.1708	0.1789
	48	6.70	**0.1929**	0.2087	0.2067	0.3134	0.2120
RMSE	4	8.04	**0.1258**	0.1368	0.1723	0.1465	0.1763
	12	0.71	**0.2031**	0.2068	0.2431	0.2046	0.2352
	24	0.47	**0.3182**	0.3197	0.3274	0.3356	0.3411
	48	0.29	**0.3893**	0.3904	0.4256	0.5311	0.4009
R^2	4	0.30	**0.9836**	0.9807	0.9693	0.9778	0.9679
	12	0.06	**0.9574**	0.9558	0.9389	0.9567	0.9428
	24	0.11	**0.8954**	0.8944	0.8893	0.8836	0.8798
	48	0.11	**0.8436**	0.8427	0.8130	0.7089	0.8342

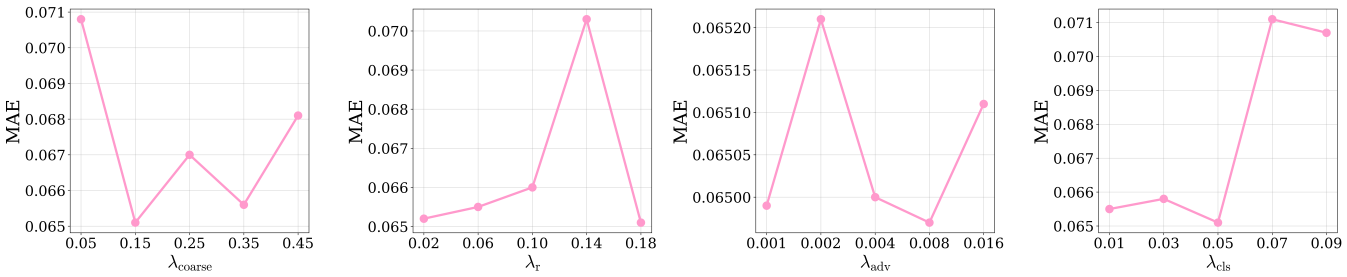

Figure 3. Sensitivity analysis of loss-weight coefficients under $H = 4$ at Station C
From top-left to bottom-right: λ_{coarse}, λ_r, λ_{adv}, and λ_{cls}

vs. 0.0867). When the horizon expands to $H \in \{12, 24, 48\}$, TAPD consistently surpasses FARM. This pattern—small gaps at short horizons and larger advantages at long horizons—suggests that short-horizon forecasting is comparatively easier and tends to saturate model performance, whereas the coarse-to-fine and adversarial constraint mechanisms in TAPD gradually release their advantage as the horizon grows.

From the perspective of metrics, TAPD shows the most stable advantage in MAE and achieves the best or second-best R^2 in all 12 configurations formed by three datasets and four horizons. This indicates that its prediction sequences are generally consistent with the global trend of the true sequences. The advantages in MAE and RMSE further demonstrate that TAPD reduces both point-wise errors and large abnormal errors, exhibiting robust performance across datasets, horizons, and metrics.

5.4. Sensitivity Analysis

To investigate the influence of loss-weight coefficients on model performance, sensitivity analyses are performed for λ_{coarse}, λ_r, λ_{adv}, and λ_{cls} on the challenging Station C dataset under $H = 4$. During each scan, the remaining coefficients are fixed at their optimal values, and the MAE curve is recorded, as shown in Figure 3.

For λ_{coarse}, the scan range is [0.05, 0.45], and the curve shows a typical decrease-then-increase trend. The minimum MAE of approximately 0.0651 appears near $\lambda_{coarse} = 0.15$. If λ_{coarse} is too small, the coarse prediction cannot provide an effective prior; if it is too large, the refinement offset is suppressed and the model tends to degenerate into coarse prediction. For λ_r, the scan range is [0.02, 0.18]. The curve is generally stable but has an isolated spike at $\lambda_r = 0.14$, while the endpoints $\lambda_r = 0.02$ and 0.18 achieve nearly identical optimal values. This weak bimodal pattern suggests that the first-order relation loss has a limited impact on final accuracy, but inappropriate weighting may still conflict with the main-task gradient.

For λ_{adv}, the scan range is [0.001, 0.016], and the MAE fluctuates only at the scale of 1×10^{-4} (0.0649–0.0652), with the optimum near $\lambda_{adv} = 0.008$. This result is consistent with the common empirical observation that adversarial signals should act as weak auxiliary supervision in GAN training. For λ_{cls}, the curve remains stable and reaches the lowest MAE when $\lambda_{cls} \leq 0.05$, but increases rapidly when $\lambda_{cls} \geq 0.07$. This indicates that overly strong transition-classification supervision can dominate the main-task gradient and impair multi-step forecasting accuracy. Therefore, λ_{cls} is recommended to be controlled within [0.01, 0.05].

Table 5. Ablation results for auxiliary loss terms under $H = 4$ using MAE as the evaluation metric
In each column, the best result is shown in **bold** and the second-best is underlined

Configuration	Station A	Station B	Station C
w/o L_{adv}	0.0819	0.0891	0.6478
w/o L_{rel}	0.0817	0.0886	0.0653
w/o L_{cls}	**0.0760**	0.0887	0.0651
Full TAPD	0.0812	**0.0876**	**0.0637**

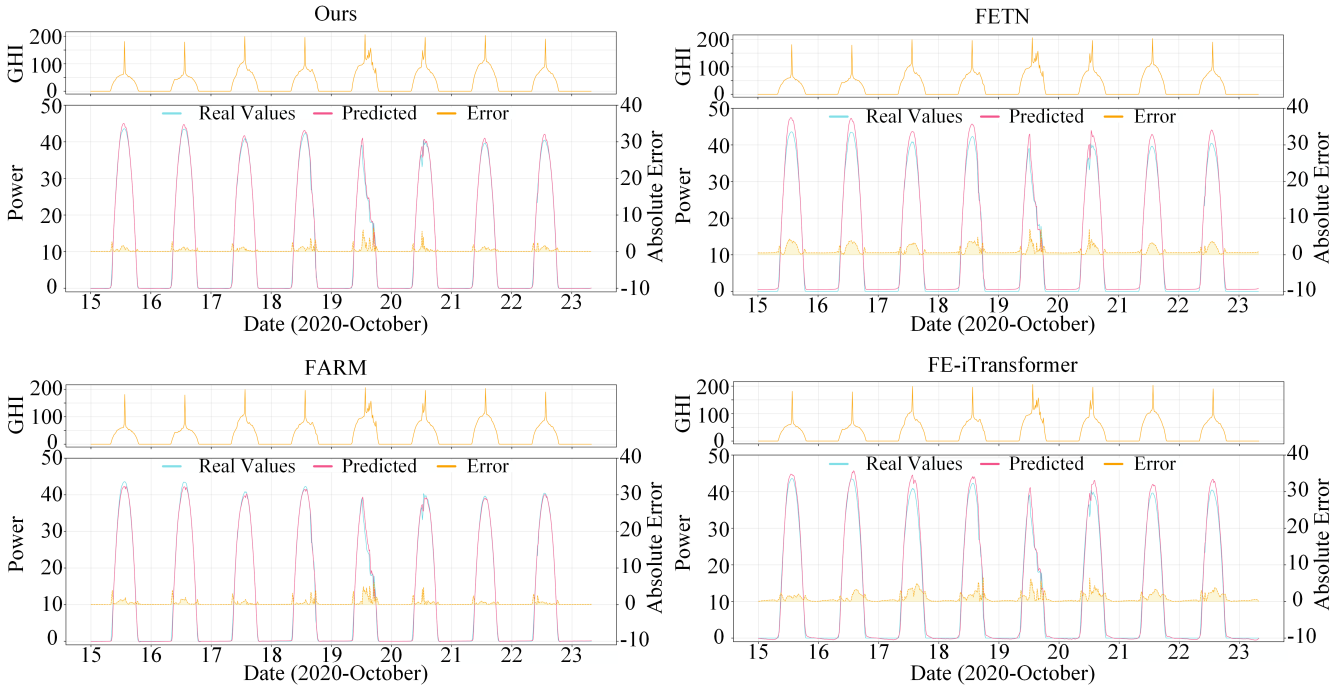

Figure 4. Prediction curves and absolute errors of different models on a continuous segment of the Station C test set from 15 October to 23 October 2020 under $H = 4$

In summary, TAPD is relatively robust to λ_{adv} and λ_r, while being more sensitive to λ_{coarse} and λ_{cls}. The recommended coefficient ranges are $\lambda_{coarse} \in [0.10, 0.20]$, $\lambda_r \in [0.02, 0.10]$, $\lambda_{adv} \in [0.004, 0.012]$, and $\lambda_{cls} \in [0.01, 0.05]$. Within these ranges, MAE remains below 0.066 and can serve as a default reference for subsequent experiments and cross-station transfer settings.

5.5. Ablation Study

To quantitatively analyze the contribution of auxiliary loss terms, three ablation configurations are designed by removing the conditional adversarial loss L_{adv}, the first-order relation loss L_{rel}, and the transition-classification auxiliary loss L_{cls}, respectively. These configurations are compared with the full model on Stations A, B, and C under $H = 4$, using MAE as the evaluation metric. The results are summarized in Table 5.

The results show that removing any auxiliary loss affects model performance, although the magnitude varies across datasets. At Stations B and C, the full model achieves the lowest MAE (0.0876 and 0.0637, respectively) compared with all ablation variants, indicating the synergistic effect of adversarial, relation, and transition-classification losses. Particularly at Station C, removing the adversarial loss sharply increases MAE to 0.6478, nearly one order of magnitude larger than other configurations. This phenomenon suggests that the conditional adversarial game plays a key role in stabilizing predicted trajectories at stations with strong fluctuations, and that point-wise losses alone cannot prevent shape collapse under challenging scenarios.

5.6. Comparative Analysis of Different Models

To visually compare prediction shapes across models, a continuous segment from 15 October to 23 October 2020 in the Station C test set is selected. Under the

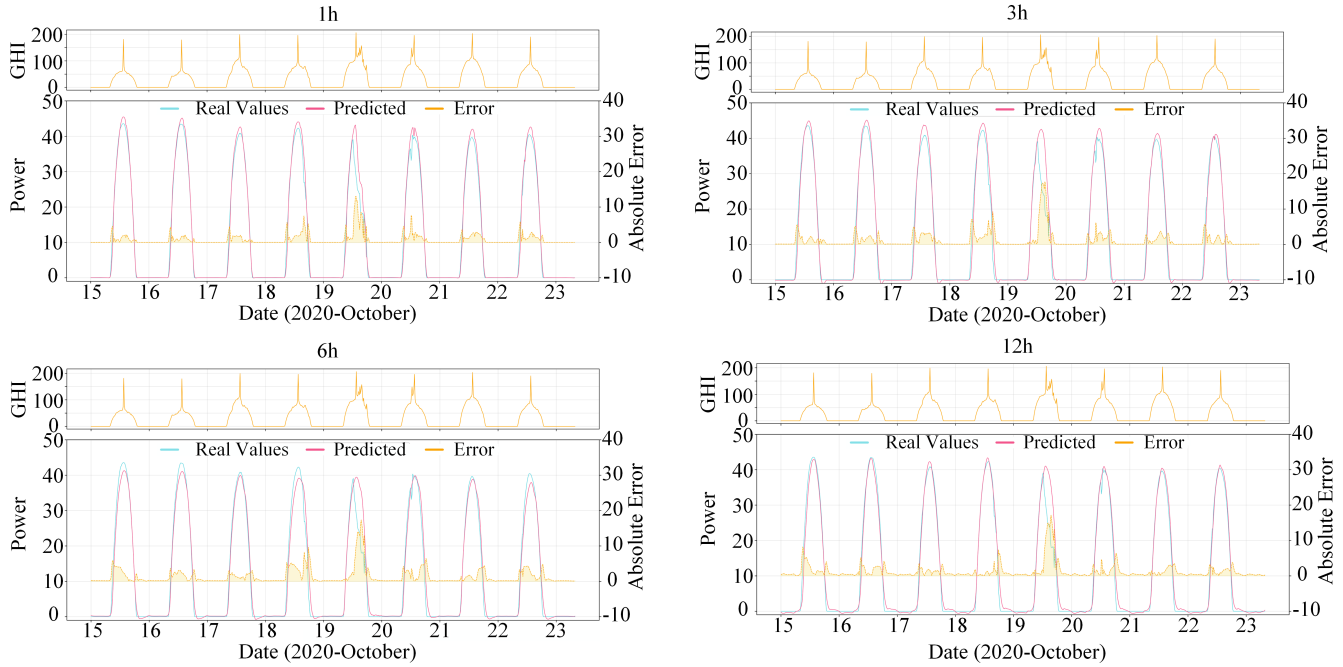

Figure 5. Prediction curves and absolute errors of TAPD under $H = 4$, 12, 24, and 48 on the same continuous Station C test segment

forecasting horizon $H = 4$ (1 h), prediction curves of TAPD, FETN, FARM, and FE-iTransformer are plotted. Each subplot overlays the GHI irradiance sequence at the top and uses orange shading to represent absolute errors at the bottom, as shown in Figure 4. This segment contains several typical sunny days and two clear transition days, making it suitable for evaluating model stability under different weather states.

During sunny periods, all four models fit the bell-shaped daily cycle well, and their prediction curves highly overlap with the ground-truth curve. Absolute errors remain low, and differences between models are not significant. The most evident differences appear on the two transition days with high-frequency GHI fluctuations. FETN and FE-iTransformer show slight lag and peak underestimation, while FARM exhibits larger peak deviations. By contrast, TAPD produces the smallest error-shaded area and the closest trajectory shape to the ground truth. This result highlights the collaborative effect between explicit enhancement of transition events by TAE and global trajectory-shape constraints imposed by the discriminator.

At night, the predictions of all models remain close to zero power and consistent with the true sequence, without false spikes. This indicates that TAPD maintains stability in low-irradiance and low-output periods. Meanwhile, adversarial training does not introduce additional abnormal outputs, avoiding the mode hallucination problem that GAN-like methods may encounter in low-energy intervals.

5.7. Comparative Analysis of Different Prediction Horizons

To analyze the degradation characteristics of TAPD under different forecasting horizons, the same Station C test segment from 15 October to 23 October 2020 is used to plot prediction curves and error shading under $H = 4$ (1 h), $H = 12$ (3 h), $H = 24$ (6 h), and $H = 48$ (12 h), as shown in Figure 5.

During sunny periods, prediction curves under all four horizons remain highly consistent with the true sequence, with only slight deviations near peaks and absolute errors generally remaining below 5. This indicates that under relatively stable background irradiance patterns, the horizon-aware predictor can stably extrapolate to a 12 h horizon using the coarse prediction prior, without introducing obvious trend drift.

During transition days, errors show clear horizon dependence. When $H = 4$, the error shading only locally increases around GHI fluctuation moments. When $H = 12$, the error magnitude expands moderately, but peak localization remains relatively accurate. When $H = 24$ and $H = 48$, both the width and height of error shading increase, and peak underestimation on 19 October is approximately 5 units. This pattern indicates that the main error source of TAPD under long horizons lies in high-frequency details around transition events rather than global trend bias. The marginal effect of transition scores and difference-derived variables extracted by

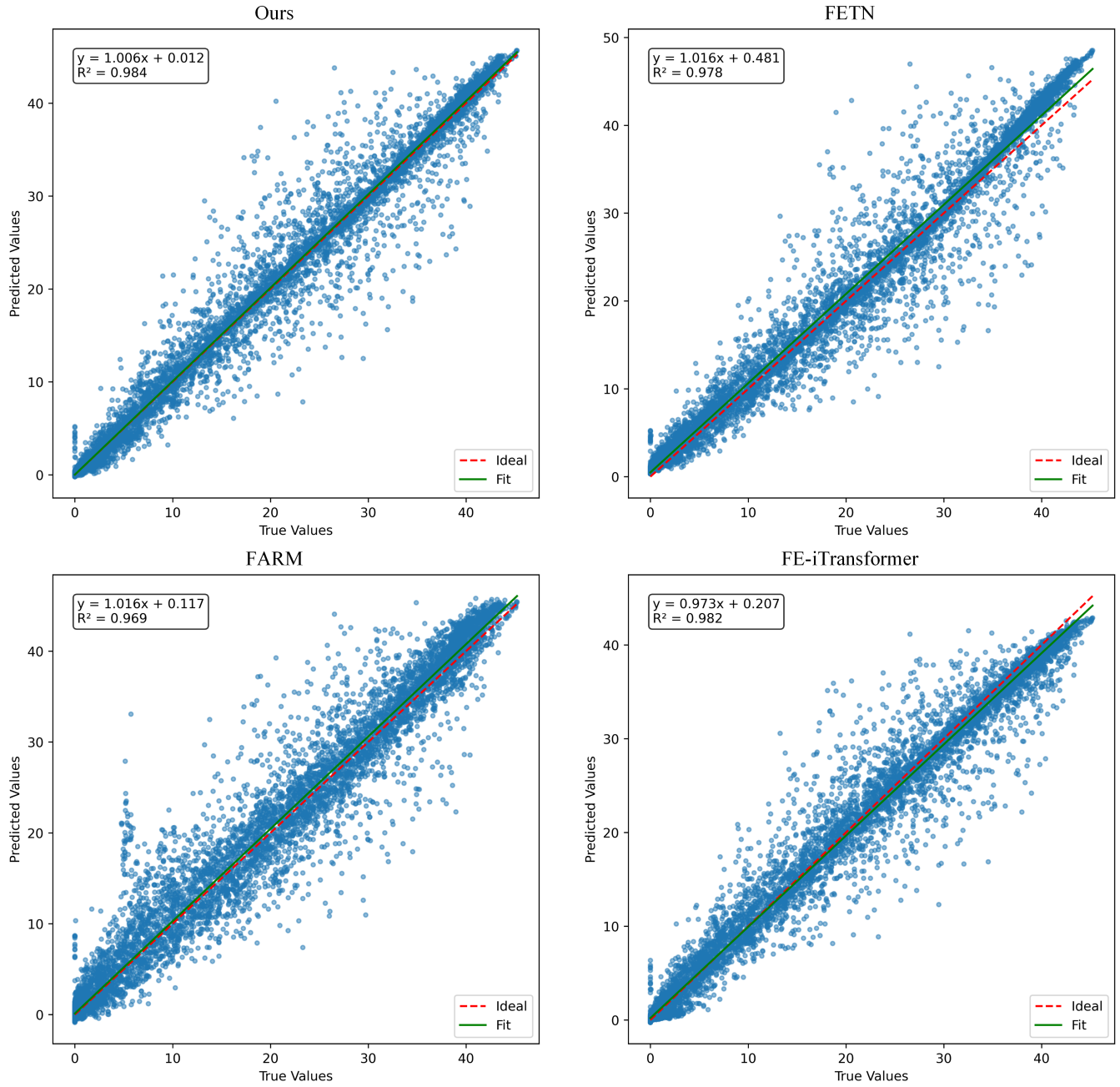

Figure 6. Predicted-versus-observed scatter plots and linear fits of four models on the Station C test set under $H = 4$

TAE gradually weakens when $H \geq 24$, suggesting that future work could introduce multi-scale frequency-domain modeling to further enhance high-frequency detail capture.

To further evaluate prediction quality at the distribution level, predicted-versus-observed scatter plots in are drawn for TAPD, FETN, FARM, and FE-iTransformer under $H = 4$ on the Station C test set. For each model, a linear regression line $y = ax + b$ is fitted, and the fitted parameters and R^2 values are annotated in Figure 6. Ideally, points should be close to the red diagonal line $y = x$, namely $a = 1$ and $b = 0$, while R^2 should approach 1.

The fitted line of TAPD is $y = 1.006x + 0.012$ with $R^2 = 0.984$, showing the slope closest to 1, the intercept closest to 0, and the highest goodness of fit. Its points are symmetrically distributed along the diagonal without obvious systematic bias. FETN has a fitted line of $y = 1.016x + 0.481$; although its slope is close to 1, the high intercept indicates an upward bias in the low-power range. FARM has a fitted line of $y = 1.016x + 0.117$ with $R^2 = 0.969$, and its middle-to-high power

points exhibit the largest vertical dispersion, indicating weaker stability near peaks. FE-iTransformer has a fitted line of $y = 0.973x + 0.207$, suggesting systematic underestimation in the high-power range.

Considering slope, intercept, and R^2 together, TAPD is closest to the theoretical optimum in all dimensions and is the most unbiased and highly correlated method among the four models. This scatter-level analysis corroborates the aggregated MAE/RMSE/R^2 results in Section 5.3 and further validates the effectiveness and stability of TAPD for multi-step PV power forecasting.

6. Conclusions

This study focuses on short-term PV power forecasting, a key task for grid dispatch, electricity market operation, energy storage configuration, and microgrid energy management. To address the non-stationarity of meteorological-power relationships and the insufficient trajectory-shape constraints in multi-step forecasting, TAPD is proposed. Through the collaborative design of the TAE and PDJM modules, TAPD explicitly models non-stationary transition segments without relying on external weather labels and imposes global constraints on future prediction trajectories. Experimental results show that TAPD performs stably for short-term PV power forecasting within a 12 h horizon, effectively improving forecasting accuracy and enhancing trajectory-shape rationality. Nevertheless, there remains room for improvement under ultra-long horizons beyond 12 h, frequent extreme weather events, or cold-start scenarios at newly built PV stations where data distributions shift substantially. Future research will focus on multimodal weak-supervision fusion, ultra-long-horizon multi-scale modeling, and cross-station adaptive transfer to further extend the applicability and engineering practicality of TAPD.

Author Contributions

Li Zhang: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing - review & editing, Writing - original draft. Tiancheng Shi: Resources, Data curation. Xueao Qiu: Conceptualization, Methodology, Supervision, Formal analysis, Project administration, Writing - review & editing, Writing - original draft. Yutong Ye: Supervision, Writing - review. Ziran Jin: Investigation, Instrumentation, Experimental Setup, Validation, Writing - original draft. All authors have read and agreed to the final version of the manuscript.

Acknowledgements

This research was funded by the Science and Technology Project of State Grid Anhui Electric Power Co., Ltd., grant number 521209250009.

Data Availability Statement

The data analyzed in this study are available from the public State Grid Renewable Energy Generation Forecasting Competition dataset cited in this article.

References

[1] IEA. Renewables 2024: Analysis and Forecasts to 2030. Paris: International Energy Agency; 2024.

[2] Gaboitaolelwe J, Zungeru AM, Yahya A, Lebekwe CK, Vinod DN, Salau AO. Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison. *IEEE Access*. 2023;11:40820-40845.

[3] Sabadus A, Blaga R, Hategan SM, Calinoiu D, Paulescu E, Mares O, Boata R, Stefu N, Paulescu M, Badescu V. A Cross-Sectional Survey of Deterministic PV Power Forecasting: Progress and Limitations in Current Approaches. *Renew. Energy*. 2024;226:120385.

[4] Mayer MJ, Gróf G. Extensive Comparison of Physical Models for Photovoltaic Power Forecasting. *Appl. Energy*. 2021;283:116239.

[5] Mayer MJ, Yang D. Pairing Ensemble Numerical Weather Prediction with Ensemble Physical Model Chain for Probabilistic Photovoltaic Power Forecasting. *Renew. Sustain. Energy Rev.*. 2023;175:113171.

[6] Mayer MJ, Yang D. Optimal Place to Apply Post-Processing in the Deterministic Photovoltaic Power Forecasting Workflow. *Appl. Energy*. 2024;371:123681.

[7] Markovics D, Mayer MJ. Comparison of Machine Learning Methods for Photovoltaic Power Forecasting Based on Numerical Weather Prediction. *Renew. Sustain. Energy Rev.*. 2022;161:112364.

[8] Pombo DV, Bindner HW, Spataru SV, Sørensen PE, Bacher P. Increasing the Accuracy of Hourly Multi-Output Solar Power Forecast with Physics-Informed Machine Learning. *Sensors*. 2022;22(3):749.

[9] Hou Z, Zhang Y, Liu Q, Ye X. A Hybrid Machine Learning Forecasting Model for Photovoltaic Power. *Energy Rep.*. 2024;11:5125-5138.

[10] He F, Fan J, Deng Y, Zhang X, Lau KT, Wang D. Weather Foundation Model Enhanced Decentralized Photovoltaic Power Forecasting Through Spatio-temporal Knowledge Distillation. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2025)*; Montreal, Canada; 2025. p. 9665-9673.

[11] Cui S, Lyu S, Ma Y, Wang K. Improved Informer PV Power Short-Term Prediction Model Based on Weather Typing and AHA-VMD-MPE. *Energy*. 2024;307:132766.

[12] Zhang M, Han Y, Wang C, Yang P, Wang C, Zalhaf AS. Ultra-Short-Term Photovoltaic Power Prediction Based on Similar Day Clustering and Temporal Convolutional Network with Bidirectional Long Short-Term Memory Model: A Case Study Using DKASC Data. *Appl. Energy*. 2024;375:124085.

[13] Yu S, He B, Fang L. Multi-Step Short-Term Forecasting of Photovoltaic Power Utilizing TimesNet with Enhanced Feature Extraction and a Novel Loss Function. *Appl. Energy*. 2025;388:125645.

[14] Si G, Xie M, Luo F, Xia T, Xi L. Frequency Domain Enhanced Transformer Network for Robust Photovoltaic Power Prediction. *Sol. Energy*. 2026;307:114358.

[15] Tang H, Kang F, Li X, Sun Y. Short-Term Photovoltaic Power Prediction Model Based on Feature Construction and Improved Transformer. *Energy*. 2025;320:135213.
[16] Inman RH, Pedro HTC, Coimbra CFM. Solar Forecasting Methods for Renewable Energy Integration. *Prog. Energy Combust. Sci.*. 2013;39(6):535-576.
[17] Zhang J, Zhu X, Xie Y, Chen G, Liu S. Detection and Prediction of Wind and Solar Photovoltaic Power Ramp Events Based on Data-Driven Methods: A Critical Review. *Energies*. 2025;18(13):3290.
[18] Xie G, Zhang Z, Xie S, Yuan Z, Liu H. CPWformer-DEC: improved Transformer with class-priority weather attention and dynamic error compensation for photovoltaic power forecasting. *Expert Syst. Appl.*. 2026;301:130580.
[19] Jiang X, Gou Y, Jiang M, Luo L, Zhou Q. Photovoltaic Power Forecasting with Weather Conditioned Attention Mechanism. *Big Data Min. Anal.*. 2025;8(2):326-345.
[20] Dai H, Zhen Z, Wang F, Lin Y, Xu F, Duić N. A Short-Term PV Power Forecasting Method Based on Weather Type Credibility Prediction and Multi-Model Dynamic Combination. *Energy Convers. Manag.*. 2025;326:119501.
[21] Lyu J, Zhu G, He C. Data-Driven Short-Term Photovoltaic Power Forecasting under Extreme Weather Conditions Using GRU-KAN Model. *Renew. Energy*. 2026;257:124756.
[22] Hua Q, Chu H, Zhang F, Zhang Y, Dong C. TFDN: A reliable hybrid time and frequency domain-based model for photovoltaic power generation time series forecasting. *Int. J. Mach. Learn. Cybern.*. 2025;16(7-8):4347-4365.
[23] Liu D, Zhu J, Lin X, Niu L, Tao Z. Channel-correlation aware photovoltaic power forecasting framework based on multi-perspective modeling. *Expert Syst. Appl.*. 2025;288:128171.
[24] Peng T, Ma C, Zhang Z, He R, Nazir MS, Zhang C. An integrative approach to enhance photovoltaic power forecasting via TimeGAN-augmented data balancing and DES-improved autoformer model. *Comput. Electr. Eng.*. 2026;130:110858.
[25] Park S, Moon J, Hwang E. Data Generation Scheme for Photovoltaic Power Forecasting Using Wasserstein GAN with Gradient Penalty Combined with Autoencoder and Regression Models. *Expert Syst. Appl.*. 2024;257:125012.
[26] Gao P, Li Y, Song W. Lightweight Multi-Modal Transformer for Photovoltaic Power Forecasting With Sky Imagery and Meteorological Data. *IEEE Access*. 2026;14:95289-95304.
[27] Chen Y, Xu J. Solar and Wind Power Data from the Chinese State Grid Renewable Energy Generation Forecasting Competition. *Sci. Data*. 2022;9(1):577.