Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multimodal Transformer Framework with State-Aware Adaptive Fusion for Ultra-Short-Term Solar Irradiance Forecasting with Sky Cloud Images

Houhang Zhao
Dongbei University of Finance and Economics
Aiyi Song
Dalian Neusoft University of Information
Yu Jin
jinyudc@dufe.edu.cn

Dongbei University of Finance and Economics
Haoxiang Yu
Education University of Hong Kong

Article

Keywords: Solar irradiance prediction, Heterogeneous modality representation, Multimodal fusion, State-aware adaptive fusion

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10464830/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multimodal Transformer Framework with State-Aware Adaptive Fusion for Ultra-Short-Term Solar Irradiance Forecasting with Sky Cloud Images

Houhang Zhao [a], Aiyi Song [b], Yu Jin[, a], Haoxiang Yu[*, c],*

a: School of Statistics, Dongbei University of Finance and Economics, Dalian, China

b: Dalian Neusoft University of Information, School of Health and Medical Technology, China

c: Academy for Educational Development and Innovation, The Education University of Hong Kong

: **Corresponding authors: Yu Jin: E-mail addresses: jinyudc@dufe.edu.cn; Telephone number: +86 13842870866, Fax numbers: 0086-0411-84710451; Haoxiang Yu: E-mail addresses: yuhaoxiang423@163.com; Telephone number: +86 17604090600.*

E-mail addresses: zhaohouhang1999@mail.163.com (Houhang Zhao); songaiyi@neusoft.edu.cn (Aiyi Song); jinyudc@dufe.edu.cn (Yu Jin); yuhaoxiang423@163.com (Haoxiang Yu)

ABSTRACT

As renewable energy penetration continues to increase, accurate ultra-short-term solar irradiance forecasting is essential for improving energy utilization and supporting power system operation. However, irradiance variations are jointly influenced by historical trends and rapid cloud changes, making traditional single-modal methods insufficient for capturing complex dynamics under varying weather conditions. This paper proposes a multimodal solar irradiance forecasting framework based on the collaborative representation of sky cloud images and numerical sequences. The framework introduces a heterogeneous modal state encoding mechanism and a state-aware adaptive selective attention fusion mechanism to enhance interactions between visual weather information and historical physical patterns. Specifically, the numerical modality captures irradiance trends and temporal correlations through dependency modeling, while the image modality extracts cloud structures and evolution features through spatiotemporal representation. The proposed fusion mechanism dynamically adjusts modality contributions according to forecasting conditions, enabling effective integration of complementary information. Experimental results demonstrate that the proposed framework outperforms benchmark models in forecasting accuracy and robustness, validating the effectiveness of combining visual cloud information with numerical evolution patterns for ultra-short-term solar irradiance forecasting.

Keywords: Solar irradiance prediction; Heterogeneous modality representation; Multimodal fusion; State-aware adaptive fusion

1. INTRODUCTION

With the ongoing advancement of the “dual carbon” goals and the accelerated development of a new power system, the share of solar photovoltaic (PV) power in the energy mix continues to rise. Compared with traditional controllable power sources, solar irradiance exhibits significant fluctuations and uncertainty; its variations are influenced not only by regular factors such as diurnal cycles and seasonal changes but also by random factors such as sudden changes in local weather [1]. Particularly in ultra-short-term forecasting scenarios, the rapid formation or dissipation of clouds over observation stations or solar power plants can cause sharp increases or decreases in solar irradiance, which in turn affect estimates of solar power output and pose challenges to renewable energy integration and the safe and stable operation of the power system [2]. Therefore, developing an ultra-short-term solar irradiance prediction model capable of accurately detecting cloud-induced fluctuations and rapidly responding to changes in irradiance holds significant theoretical and engineering value. In recent years, with the rapid penetration of wind and solar energy into energy systems, research on renewable energy forecasting based on time-series feature engineering, ensemble learning, and artificial intelligence methods has continued to advance, further driving the development and optimization of photovoltaic forecasting models under complex conditions [3, 4].

Most existing methods for predicting solar irradiance rely on structured numerical data, such as historical irradiance records or numerical weather forecasts, and use these continuous numerical variables to build predictive models. Given that solar irradiance time series are inherently highly stochastic and exhibit distinct seasonal and weather-related fluctuations, researchers have conducted extensive studies using various approaches, including statistical methods, machine learning, and deep learning [5, 6]. Bacher et al. [7] proposed an online short-term solar power forecasting method that combines numerical weather prediction with recursive least squares to forecast the short-term solar power output of a rooftop photovoltaic system in Denmark. Pedro and Coimbra[8] systematically compared various solar power forecasting techniques without exogenous variables, demonstrating that forecasts based solely on historical solar irradiance or power time series can achieve a certain level of accuracy under relatively stable weather conditions. Sharadga et al. [9] further compared the performance of statistical and artificial intelligence models in solar power forecasting, validating the feasibility of modeling based on historical time series. Although these methods can effectively capture the continuous patterns of solar irradiance, historical irradiance values primarily reflect past solar radiation conditions and thus fail to fully characterize leading indicators such as sudden local weather changes and future irradiance variations. Consequently, under complex weather conditions, these methods are prone to prediction lags or amplified errors [10].

To address the limitations of relying solely on historical irradiance data, subsequent studies incorporated exogenous meteorological variables and numerical weather forecast information to characterize the dynamic behavior of solar irradiance under varying climatic conditions. For example, Mellit et al. [11] compared deep learning networks,

including LSTM, GRU, and one-dimensional CNNs, for short-term solar irradiance forecasting; du Plessis et al. [12] investigated the ability of deep learning models to capture operational behavior in utility-scale photovoltaic systems; and Transformer-based architectures and attention mechanisms have been introduced to enhance the representation of weather categories and multiscale features [13]. By expanding input variables, these methods provide models with additional weather-related information, improving prediction accuracy under different seasons and weather conditions. However, historical irradiance time series still struggle to provide timely information on cloud shading due to the rapid evolution of cloud cover. Moreover, meteorological observations and numerical weather forecasts are limited by spatial resolution and update frequency, making it difficult to accurately characterize local cloud evolution [14, 15]. Therefore, numerical models alone remain insufficient for ultra-short-term forecasting under local occlusion and sudden weather variations. With the development of multimodal deep learning, image information has been further incorporated to improve spatial perception of cloud fields [16, 17].

To overcome the spatial perception limitations of numerical models, researchers have introduced satellite imagery to characterize cloud field distributions [18-20]. Deep learning methods, including convolutional neural networks, residual networks, and attention mechanisms, enable satellite images to extract cloud spatial distributions and evolution patterns. Combined with irradiance or meteorological variables, these features improve the ability of forecasting models to capture cloud-induced irradiance changes [21]. Si et al. [22] integrated cloud motion prediction with an XGBoost model considering solar position, while Xie et al. [23] utilized satellite image texture features and transfer learning for regional solar radiation forecasting. Compared with numerical models, satellite imagery provides more intuitive cloud information; however, its large coverage area and limited update frequency restrict its application in minute-level ultra-short-term forecasting of local cloud variations [19, 24]. In contrast, ground-based sky images provide higher spatiotemporal resolution and can directly capture local cloud cover and solar occlusion, making them suitable for minute-to-hour-scale cloud dynamics analysis [25]. Kamadinata et al. [26] used sky images and artificial neural networks for solar irradiance prediction, while Feng et al. [27] proposed a CNN-based model using sky image sequences for minute-to-hourly solar energy forecasting. These studies demonstrate that sky images provide valuable visual weather information, particularly under complex conditions such as overcast skies and rapidly moving clouds [28].

However, existing studies based on satellite imagery or sky images still mainly rely on a single image modality for solar irradiance forecasting. Although images provide rich visual information on weather conditions, they cannot directly reflect historical irradiance variations. Meanwhile, image quality and insufficient information in individual frames may affect feature stability, causing models to overlook temporal evolution patterns, especially under complex weather conditions. Paletta et al. [29] showed that although spatiotemporal image features improve short-term forecasting performance, some models remain limited in detecting abrupt events, such as cloud-induced solar occlusion, in advance. Therefore, image-based and numerical time-series approaches exhibit natural complementarity in solar irradiance forecasting: images provide external visual

information on cloud evolution and occlusion events, whereas numerical sequences provide historical irradiance trends and temporal constraints. Their fusion can improve the characterization of irradiance variations under complex weather conditions [30].

Regarding the fusion of image and numerical modalities, existing methods can generally be classified into output-level fusion, feature-level fusion, and deep learning-based joint fusion [31]. Some studies separately construct models using image and numerical information and then combine predictions through weighted averaging or ensemble learning. However, such post-fusion strategies only integrate model outputs and fail to exploit cross-modal interactions during feature learning, limiting their ability to capture correlations between cloud variations and irradiance responses [32]. Another approach extracts image features, such as weather categories, and incorporates them as external inputs into forecasting models. For example, Xie et al. [31] performed short-term solar irradiance forecasting by extracting features from sky images and combining them with an SSA-BiLSTM; Ajith and Martínez-Ramón [25] fused infrared cloud images with irradiance data for microscale solar radiation prediction. Although these methods improve the utilization of visual information, compressing images into limited features may result in the loss of cloud spatial structures and motion characteristics. Furthermore, many multimodal methods remain based on feature concatenation or shallow fusion, insufficiently exploiting complementary information, redundancy, and dynamic relationships among modalities [33] . Therefore, developing effective deep multimodal fusion mechanisms that fully exploit the complementary characteristics of image and numerical modalities remains essential for improving ultra-short-term solar irradiance forecasting performance.

Given the limitations of existing methods in terms of multimodal information interaction mechanisms and adaptability to complex weather scenarios, this paper proposes TSGAN (Transformer Selective Gated Attention Network), a Transformer-based, state-aware adaptive fusion forecasting framework. By synergistically fusing visual information from sky-cloud images with structured numerical sequence data, this framework enhances the model's ability to capture complex dynamic changes in ultra-short-term solar irradiance. The framework first constructs a Unified State Encoder (USE) with heterogeneous representation paths tailored to the data characteristics of different modalities. For the numerical modality, a Transformer encoding structure is utilized to model deep temporal relationships among historical irradiance and meteorological variables, and a global dependency capture mechanism is employed to extract dynamic evolutionary patterns and latent correlation features associated with irradiance changes. For the image modality, a visual encoding branch integrating spatial perception and temporal evolution is constructed. A convolutional feature extraction mechanism captures local spatial structures in sky-cloud patterns, while a temporal modeling mechanism further characterizes dynamic information such as cloud movement and visual state changes, thereby forming a visual state representation with spatiotemporal consistency. After obtaining high-level state representations from different modalities, this paper further designs a selective gated attention network as the core cross-modal fusion module. By introducing a modality-state-aware mechanism, the network dynamically adjusts the relative contributions of the image and numerical modalities across different forecasting

scenarios. Unlike traditional feature concatenation or fixed-weight fusion methods, TSGAN does not combine all modal information indiscriminately. Instead, by integrating state awareness with a gated weight-generation mechanism, it adaptively learns the relative importance of features from different modalities and performs information filtering and complementary enhancement in the latent space. This mechanism dynamically adjusts the contributions of the two modalities based on the weather state of the current sample, thereby enhancing the model's responsiveness to irradiance changes under complex weather conditions and improving forecasting stability. The main contributions of this paper are as follows:

(1) We propose TSGAN, a multimodal solar irradiance forecasting framework based on a Transformer and a selective gated attention network. By constructing a collaborative representation system for sky-cloud images and numerical time series, TSGAN enables the joint modeling of visual weather information and historical physical patterns, effectively addressing the information gaps inherent in single-modality approaches regarding spatial perception or temporal evolution.

(2) We construct a dual-branch state encoding structure tailored to heterogeneous inputs. The Transformer temporal encoder extracts long-term dependencies and dynamic patterns from the numerical modality, while a visual encoding mechanism captures cloud structures and motion characteristics from sky-cloud images, thereby yielding complementary multimodal state representations.

(3) We propose a selective gated attention network that transforms the multimodal fusion process from static feature combination into dynamic information selection based on sample-state awareness. This network can adaptively adjust the fusion weights of image and numerical features according to changes in modality reliability under different weather conditions, enabling more refined cross-modal information interaction.

2. METHODOLOGY

To validate the effectiveness and superiority of the proposed method, this study constructs an overall research framework, as shown in Figure 1. The framework primarily consists of four steps: (1) multimodal state encoding and feature alignment; (2) state-aware adaptive modality fusion; (3) multimodal fusion forecasting; and (4) model performance evaluation.

The TSGAN proposed in this paper is designed for ultra-short-term GHI forecasting tasks, and its overall architecture consists of a unified state encoder and a modality-selective predictive decoder. The unified state encoder extracts heterogeneous state representations from numerical sequences and sky images, while the modality-selective predictive decoder employs a modality-selective attention gating mechanism as its core fusion module to dynamically filter and adaptively fuse the two types of representations, ultimately outputting GHI forecasts for future time steps. Unlike traditional multimodal models, which directly concatenate features from different modalities and feed them into the prediction layer, the core idea of TSGAN is to explicitly model the multimodal fusion process as a dynamic decision-making process involving "modality-state identification—

modality-importance allocation—complementary information fusion," enabling the model to autonomously determine the relative contributions of the numerical and image modalities to the forecast based on the current sample state. Specifically, the numerical modality consists of historical GHI, DNI, DHI, and meteorological variables, primarily describing the historical inertia and short-term evolution trends of irradiance; the image modality consists of continuous sky-cloud images, primarily providing visual priors such as cloud-cover information and changes in sky brightness. Although the two modalities are complementary, their relative importance varies across different weather scenarios. Therefore, this paper does not adopt a fusion strategy based on fixed weights or simple concatenation; instead, it incorporates a modality-selective attention gating mechanism as the core design of TSGAN, enabling the model to adaptively determine the effective contributions of the two modalities at both the sample and feature-dimension levels.

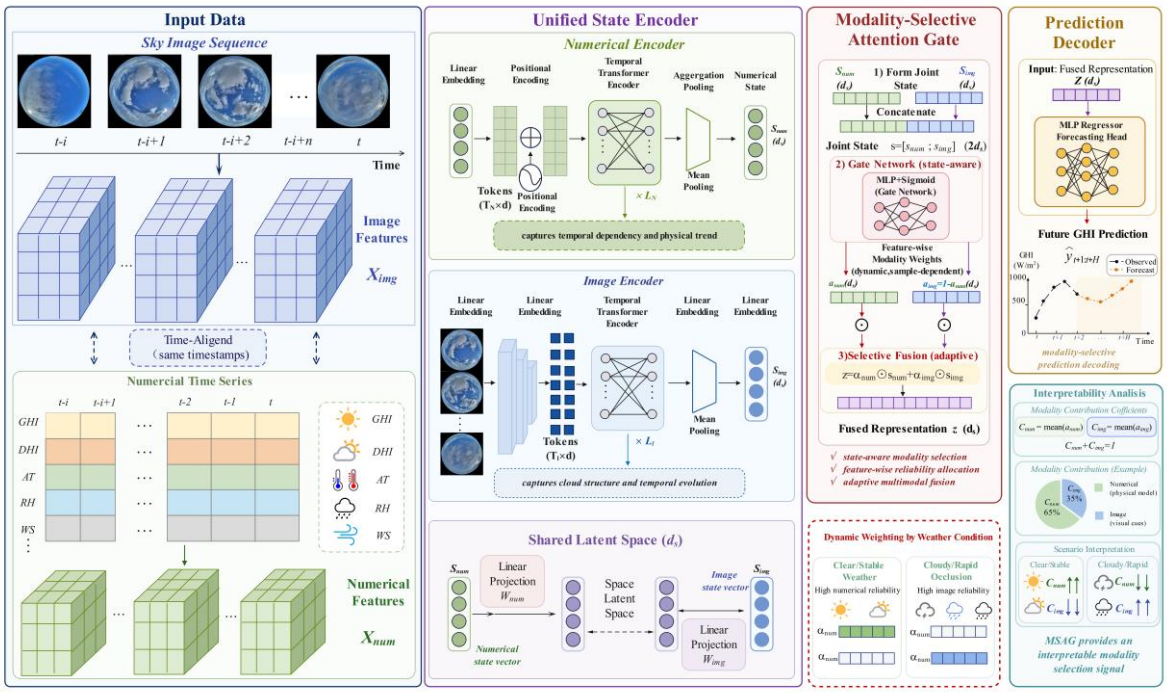


Figure 1. Forecasting framework.

2.1 Unified State Encoder

The goal of a unified state encoder is to map numerical sequences and sky images onto the same latent space, providing comparable high-level representations for subsequent modal selection and fusion. Given a history observation window of length L , the numerical modal input is denoted as $\boldsymbol{X}^{num} \in \mathbb{R}^{L\times F}$, where F is the dimension of the numerical variables; the time-aligned sequence of sky images is denoted as $\boldsymbol{X}^{img} \in \mathbb{R}^{L\times C\times H\times W}$. The prediction process of TSGAN can be formalized as:

$$\hat{y}_{t+\tau} = F_\theta(X^{num}, X^{img}) \tag{1}$$

Here, $\hat{\boldsymbol{y}}_{t+\tau}$ represents the GHI prediction values for the next τ steps, and F_θ is a nonlinear mapping function composed of a state encoder, a gated fusion module, and a prediction decoder.

For numerical modalities, this paper employs a temporal Transformer encoder to extract dynamic dependencies among historical physical variables. First, the raw numerical sequences are mapped to a unified latent space, with positional encoding added to preserve temporal order; subsequently, the Transformer encoder models global dependencies across different time steps to obtain contextual representations of the numerical modalities:

$$Z_{num}^0 = X^{num} W_n + b_n + PE_{num} \tag{2}$$

$$Z_{num} = TransformerEncoder_{num}(Z_{num}^0) \tag{3}$$

$$h_{num} = \text{Pool}(Z_{num}) \tag{4}$$

Here, $\boldsymbol{Z}_{num}$ represents the high-level temporal representation of the numerical sequence, and $\boldsymbol{h}_{num} \in \mathbb{R}^d$ is the aggregated numerical modal state vector. This branch captures the nonlinear relationship between GHI and meteorological variables while preserving the overall trend of change within the historical window.

For the image modality, this paper adopts an encoding architecture that combines a CNN with a temporal Transformer. The CNN first extracts local spatial features from single-frame sky images, including changes such as areas where the sun is obscured and the edges of cloud formations; it then organizes the visual features of consecutive image frames into a temporally ordered sequence of visual tokens and uses the temporal Transformer to model the dynamic evolution of the cloud sequence:

$$e_t^{img} = f_{cnn}(I_t), \quad t = 1, 2, ..., L \tag{5}$$

$$Z_{img}^0 = [e_1^{img}, e_2^{img}, ..., e_L^{img}] + PE_{img} \tag{6}$$

$$Z_{img} = TransformerEncoder_{img}(Z_{img}^0) \tag{7}$$

$$h_{img} = \text{Pool}(Z_{num}) \tag{8}$$

Here, $\boldsymbol{h}_{img}$ represents the image modal state vector. This branch not only describes the spatial structure of a single-frame cloud image but also captures the continuous trends in cloud cover within the historical window, thereby providing visual prior information for short-term GHI fluctuations.

2.2 State-Aware Adaptive Fusion Mechanism

After obtaining the numerical state vector h_{num} and the image state vector h_{img},

the model further enables dynamic interaction between the two types of modal information through an adaptive attention fusion module. Since the effectiveness of image and numerical information varies under different weather conditions, this paper constructs a state-aware weight-allocation mechanism that adaptively adjusts the contributions of features from different modalities based on the joint state of the current sample, enabling the model to fully leverage the historical evolution patterns in numerical time series and the visual prior information in sky-cloud images. This module first jointly models the states of the two modalities to generate the corresponding modality-selection weights and then dynamically filters and fuses different feature dimensions based on these weights. Compared with fixed fusion strategies, this method can flexibly adjust modality dependencies in response to changing weather conditions, effectively utilizing historical trends under stable weather conditions while enhancing the role of visual information in scenarios such as rapid cloud movement or solar occlusion, thereby obtaining a more discriminative fused state representation.

Specifically, the state vectors of the two modalities are first concatenated to form $\boldsymbol{r = [h_{num}; h_{img}]}$ joint state representation. The joint state vector r contains both the numerical modality's representation of historical physical trends and the image modality's representation of the visual state of the sky; therefore, it can be regarded as a multimodal state summary of the current forecasting sample. Next, modality-selection weights are generated using a lightweight attention-gating network. In a two-modality scenario, the sigmoid gate can be viewed as a form of binary soft-attention allocation:

$$s = W_g r + b_g \tag{9}$$

$$g = \sigma(s), \quad g \in (0,1)^d \tag{10}$$

$$a_{num} = g, \quad a_{img} = 1 - g \tag{11}$$

Here, s represents the modal selection score derived from the joint state mapping, g denotes the attention weight of the numerical modality on each latent feature dimension, and $1-g$ denotes the attention weight of the image modality on the corresponding dimension. Since g is determined by the current sample's joint state r, this weight is not a fixed parameter but rather a modality-reliability response that changes dynamically with the input sample. If the historical numerical sequence in a given sample exhibits a stable trend, the gating network can increase the weight of the numerical features; if the sky image shows obvious cloud cover or sudden changes in brightness, the gating network can enhance the contribution of the image features in the relevant dimensions. Ultimately, the model obtains a fused representation through dimension-wise weighting and feeds it into the forecasting decoder to predict solar irradiance at future time steps:

$$h_{\text{fuse}} = a_{num} \odot h_{num} + a_{img} \odot h_{img} = g \odot h_{num} + (1-g) \odot h_{img} \tag{12}$$

$$\hat{y}_{t+\tau} = f_{dec}(h_{\text{fuse}}) \tag{13}$$

Here, $\odot$ denotes element-wise multiplication, and $f_{dec}(\cdot)$ represents the prediction decoder. Unlike global scalar weights, $g \in (0,1)^d$ is a weight vector at the feature dimension level; therefore, TSGAN does not merely decide "whether to choose numerical values or images as a whole," but performs fine-grained information filtering on each hidden feature dimension, enabling the model to simultaneously preserve the temporal evolution patterns in the numerical modality and the cloud variation information in the image modality. To further explain the model's overall dependence on different modalities at the sample level, we can define a modal contribution coefficient:

$$a_{num} = \frac{1}{d}\sum_{k=1}^{d} g_k, \quad a_{img} = 1 - a_{num} \tag{14}$$

When $\boldsymbol{a_{num}}$ is large, it indicates that the model considers the current sample to be more dependent on the physical trends in the historical data sequence; when $\boldsymbol{a_{img}}$ is large, it indicates that the model considers the visual information of cloud layers in the sky images to be more critical for the prediction. Thus, TSGAN not only performs feature fusion but also provides an interpretable modal selection signal that can be used to analyze how the model's reliance on the two modalities changes under different weather conditions.

2.3Evaluation Metrics

To comprehensively evaluate the performance of the proposed model in the solar irradiance forecasting task, this paper uses mean absolute error (MAE), symmetric mean absolute percentage error (SMAPE), normalized root mean square error (NRMSE), and the coefficient of determination (R^2) as evaluation metrics. Among these, MAE measures the average absolute deviation between predicted and observed values and provides an intuitive measure of the model's overall forecasting error; SMAPE evaluates the deviation between predicted and observed values in terms of relative error, thereby mitigating the impact of differences in data scale across varying irradiance levels; NRMSE emphasizes larger errors and is therefore better suited to evaluating the model's ability to capture irradiance fluctuations under sudden weather changes, such as cloud cover; and R^2 measures how well the predicted values fit the observed trend, reflecting the model's ability to capture the overall evolution of solar irradiance. The formulas for these metrics are as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{15}$$

$$SMAPE = \frac{100\%}{N}\sum \frac{\left|y_i - \hat{y}_i\right|}{(\left|y_i\right| + \left|\hat{y}_i\right|)/2} \tag{16}$$

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}} \tag{17}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{18}$$

3 EMPIRICAL RESEARCH

3.1 Data Sources and Sample Construction

This paper conducts experiments using an open dataset published on the Open Energy Data Initiative (OEDI) platform. The data are sourced from the *abq_data* subset available on its public webpage (https://data.openei.org/submissions/8414） [34]. This dataset is designed for the joint modeling of solar irradiance and sky images. It includes ground-based meteorological observations, solar irradiance variables, and sky images corresponding to the observation times in the Albuquerque, New Mexico, area, thereby providing both numerical state information and visual information on cloud conditions. Given that this paper focuses on short-term forecasting performance under continuous time-series conditions, and considering that the original data for 2020 and 2022 are incomplete, directly including these years in the experiments could lead to incomplete seasonal distributions and an imbalance between the training and test samples. Therefore, this paper selects the relatively complete data for the entire year of 2021 as the study sample and divides them into four subsets—Q1, Q2, Q3, and Q4—based on calendar quarters, thereby enabling a more thorough evaluation of the model's forecasting stability and generalization ability across different temporal contexts.

This study uses data with a 10-minute resolution to ensure temporal consistency across different modalities. Numerical variables and corresponding sky images are aligned according to the same time scale and used as inputs for the forecasting model.. After this processing, each sample contains both a historical numerical sequence and a sky-image sequence with the same time index, thereby ensuring temporal consistency across modalities. For missing values and invalid samples, this study first excludes records that cannot be matched to valid image paths and interpolates isolated missing values in the numerical variables. Additionally, considering that the forecasting task has limited practical significance when nighttime irradiance is zero, this study retains only valid daytime samples for which ghi_0 > 0, preventing a large number of nighttime zero-value samples from weakening the model's ability to learn cloud-induced irradiance fluctuations.

Regarding variable selection, this study uses global horizontal irradiance, ghi_0, as the forecasting target and employs its historical observations as key input information for the numerical modality. The numerical modality primarily includes variables related to irradiance, meteorological conditions, and local wind fields, which are used to characterize variations in solar irradiance and their external weather drivers. The image

modality consists of sky images corresponding to the observation times, which provide visual information such as cloud cover, cloud type, solar occlusion, and cloud texture. These two modalities characterize irradiance variation processes from the perspectives of structured meteorological observations and unstructured sky images, respectively, laying the foundation for subsequent multimodal fusion forecasting. The specific variable names and definitions are presented in **Table 1.**

During the sample construction phase, this paper uses a sliding-window approach to generate supervised learning samples. The aggregated data have a time interval of 10 minutes, and the model uses observations from the previous hour as input; thus, each sample contains numerical features from six consecutive historical time steps and the corresponding sky images. The prediction targets are global horizontal irradiance values for the next 10, 20, and 30 minutes, corresponding to forecasting steps h = 1, 2, and 3, respectively.

Table 1. Description of numerical meteorological and irradiance variables

Variable	The full name of the variable
ghi_0	Global Horizontal Irradiance at the current time step
Dni_0	Direct Normal Irradiance at the current time step
Dhi_0	Diffuse Horizontal Irradiance at the current time step
Bp_0	Barometric Pressure at the current time step
Wind_speed_0	Wind speed at the current time step
Wind_direction_0	Wind direction at the current time step
Temp_air_0	Air temperature at the current time step
Relative_humidity_0	Relative humidity at the current time step
Wind_direction_std	Standard deviation of wind direction at the current time step
Rain_mm_tot	The average daily length of each piece of Dianping text

To avoid information leakage in time-series forecasting, all datasets in this paper are partitioned chronologically rather than randomly shuffled, thereby preserving the temporal structure of the samples. Within each quarterly dataset, the first 70% of the samples are used for model training, the subsequent 15% for validation and parameter selection, and the final 15% for testing and performance evaluation.

To validate the effectiveness of the proposed multimodal gated fusion model, this paper establishes three comparison groups. Comparison Group (i) compares the proposed method with baseline models that use only the numerical modality, including traditional statistical models, machine learning models, and deep time-series models, to assess whether introducing sky images can further improve forecasting accuracy at different forecasting horizons. Comparison Group (ii) compares the proposed method with visual forecasting models that use only sky images to examine whether combining numerical sequence information with image information improves the ability to characterize absolute irradiance levels and capture cloud-induced fluctuations. Comparison Group (iii) compares the proposed method with multimodal fusion strategies, such as direct concatenation, standard gated fusion, and cross-attention, to verify the effectiveness of TSGAN in adaptive modality selection and forecasting performance enhancement.

3.2 Comparison with the Baseline Model

To evaluate the effectiveness of the proposed method compared with conventional numerical forecasting frameworks, this paper selects XGBoost, RF, MLP, GRU, LSTM, Transformer, and PatchTST as baseline models [35-38]. Among these, RF, XGBoost, and MLP represent traditional machine learning and ensemble learning methods; GRU and LSTM focus on extracting recurrent dependencies within sequences; and Transformer and PatchTST further characterize time series from the perspectives of global dependencies and patch-based modeling, respectively. All models employ the same data partitioning, input window, and forecasting settings; the relevant results are presented in **Table 2** and **Figure 2.**

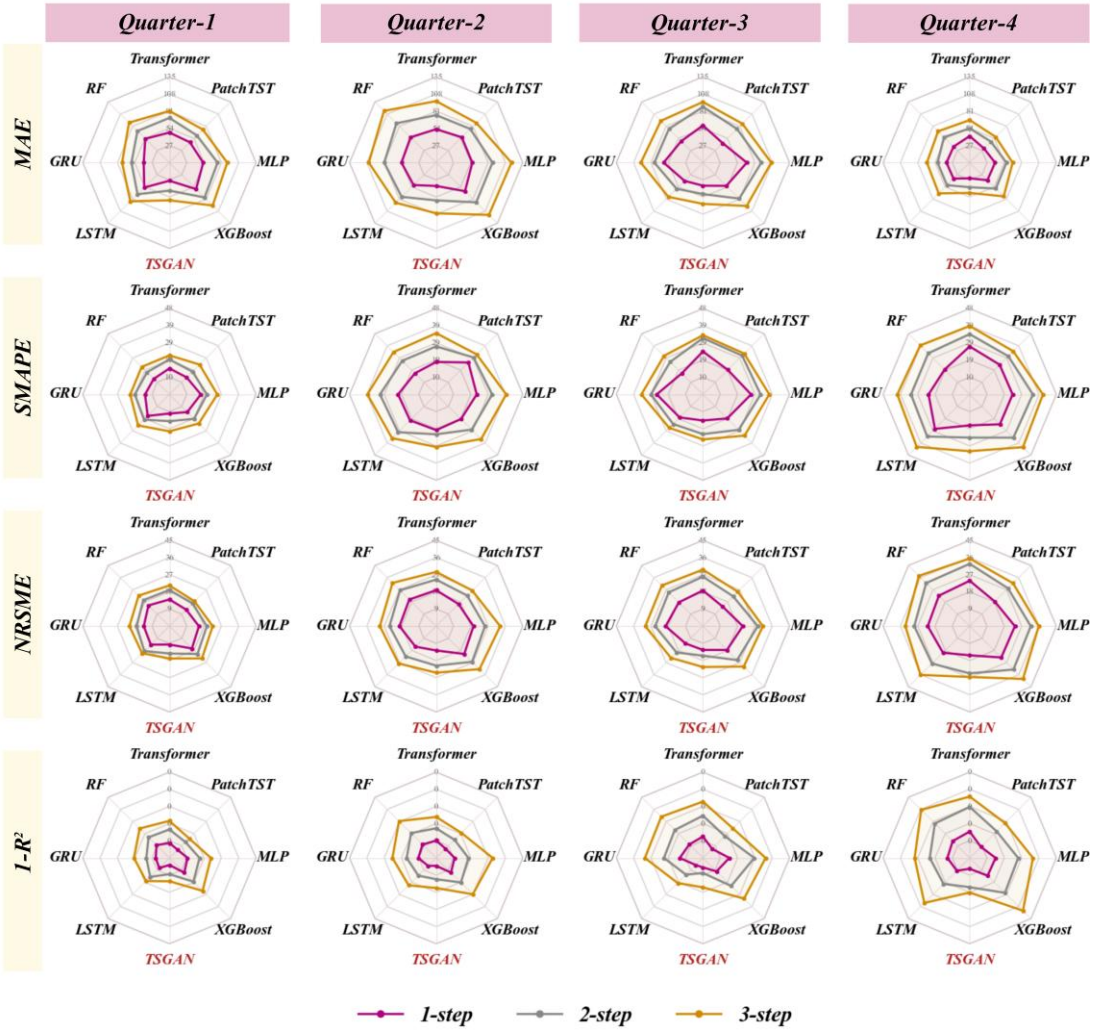


Figure 2 Performance comparison of numerical forecasting models

As shown by the performance distributions of the various models across different quarters and forecasting horizons in the figure, all types of numerical models are capable of learning certain variation patterns from historical irradiance and meteorological variables; however, their forecasting capabilities vary significantly across seasons and as the forecasting horizon increases. Compared with the various numerical baseline models, the method proposed in this paper maintained superior overall performance across all four

quarters and three forecasting steps, with average MAE, NRMSE, and SMAPE values of 47.312, 17.813, and 21.245, respectively, and an average R^2 of 0.9194. Further examination of performance variations across different forecasting horizons reveals that the proposed method demonstrates a more pronounced advantage in one-step-ahead forecasts. This is because irradiance changes over short timescales are often directly influenced by local cloud movement and shading variations, whereas visual information provides prior knowledge of the spatial state in the immediate vicinity, enabling the model to capture potential irradiance trends in advance. Compared with models that rely solely on historical numerical responses, the proposed method can utilize image information to promptly correct for sudden irradiance fluctuations, thereby achieving higher accuracy in short-term forecasting.

As the forecasting horizon extends from Step 1 to Step 3, performance generally declines. This is because irradiance variations are influenced not only by historical trends but also by short-term dynamic factors such as cloud movement and changes in shaded areas. For models that rely solely on numerical information, forecasts are primarily based on observed irradiance responses. When rapid cloud changes occur in the future, these models struggle to obtain timely information about changes in the external environment, making them prone to forecasting lags and cumulative errors. In contrast, the proposed method maintains relatively stable performance as the forecasting horizon increases. For example, the average R^2 decreases from 0.9636 at Step 1 to 0.8790 at Step 3, representing a relatively limited decline. This indicates that multimodal information fusion can effectively enhance the model's ability to anticipate future irradiance trends. Specifically, the numerical modality provides historical patterns of irradiance variation and meteorological constraints, while the image modality extracts early-warning information from cloud structures and potential occlusion trends. The two modalities complement each other, enabling the model not only to learn "changes that have already occurred" but also to capture "changes that are about to occur," thereby reducing the uncertainty associated with extrapolation based solely on historical time series.

Table 2. Performance comparison between the proposed model and single numerical modality models for solar irradiance forecasting

STEP	Model	Q1				Q2				Q3				Q4			
		MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2
Step-1	RF	52.9897	12.3797	15.3156	0.9256	56.3470	16.8027	19.6662	0.9216	47.7288	16.7416	17.5993	0.9212	35.4693	19.8229	22.5299	0.9017
	XGBoost	58.7660	14.0933	16.4614	0.9140	63.4248	19.6014	20.5291	0.9146	51.5024	19.1569	17.5571	0.9216	38.9485	23.8233	23.1045	0.8966
	MLP	52.5641	17.5571	15.2730	0.9260	56.1962	22.6262	19.2813	0.9246	68.2113	26.8714	20.5997	0.8921	39.4765	23.9402	23.3823	0.8941
	GRU	40.0042	13.5983	13.2609	0.9442	53.9849	21.2596	18.9924	0.9269	61.1701	25.7836	19.3013	0.9052	36.1166	22.9691	21.6926	0.9089
	LSTM	46.8536	15.4118	13.6882	0.9405	53.5547	19.9720	18.8822	0.9277	54.3577	20.0967	18.7509	0.9106	38.8054	26.6381	22.4317	0.9025
	Transformer	47.2497	14.6467	13.9456	0.9383	52.4942	18.2871	18.7854	0.9285	58.2745	24.1689	18.6239	0.9118	41.7129	27.0208	23.6987	0.8912
	PatchTST	46.2377	13.8455	14.3517	0.9346	53.1004	21.6201	18.9871	0.9269	50.9296	20.3758	17.7784	0.9196	36.2664	24.5423	22.3933	0.9029
	TSGAN	**28.4211**	**10.7657**	**9.5866**	**0.9708**	**36.6041**	**20.1119**	**12.6762**	**0.9674**	**36.1998**	**14.6895**	**12.3918**	**0.9609**	**24.5830**	**17.4334**	**15.2121**	**0.9552**
Step-2	RF	70.2633	17.6600	19.1439	0.8796	87.3365	26.7932	26.4605	0.8552	74.8056	25.8801	24.9931	0.8371	56.4977	32.7474	31.8323	0.7988
	XGBoost	77.2864	19.4473	20.5321	0.8615	88.1219	28.0325	26.2850	0.8571	79.7950	28.2082	25.0105	0.8369	57.5884	34.4438	32.0023	0.7967
	MLP	74.9282	20.9336	19.3706	0.8768	87.5695	31.2376	25.2535	0.8681	90.8654	32.1935	28.3509	0.7904	57.1429	35.2786	31.5829	0.8020
	GRU	59.0432	19.2945	16.9440	0.9057	80.8119	31.3712	24.1437	0.8794	75.5947	29.0572	24.7177	0.8407	50.0470	32.7457	28.6730	0.8368
	LSTM	70.0936	20.0605	18.3628	0.8893	83.3903	29.1944	24.8387	0.8724	87.0310	32.1071	26.6532	0.8147	57.9969	34.8493	30.7947	0.8117
	Transformer	70.2039	19.7134	18.9135	0.8825	74.6050	27.0350	24.3105	0.8778	87.6198	31.7198	25.8468	0.8258	53.7978	33.9544	32.5558	0.7896
	PatchTST	61.6895	20.1016	17.8046	0.8959	80.5617	29.7874	25.2363	0.8683	75.8673	28.7317	23.7668	0.8527	63.0903	38.0866	34.8760	0.7585
	TSGAN	**43.9107**	**15.0788**	**14.2682**	**0.9331**	**59.2346**	**22.4698**	**20.5965**	**0.9123**	**49.3443**	**22.3576**	**15.4098**	**0.9381**	**38.5968**	**24.3409**	**24.6437**	**0.8794**
Step-3	RF	88.8093	21.8896	22.5330	0.8283	114.8431	33.7175	31.9194	0.7854	92.7084	30.7261	30.0340	0.7594	69.5711	39.0298	37.1546	0.7200
	XGBoost	95.1660	23.1660	23.8095	0.8083	115.9998	35.5146	31.7069	0.7883	96.1837	32.8281	29.8207	0.7628	74.8514	42.0306	38.9493	0.6923
	MLP	89.9705	26.7531	22.3870	0.8305	117.1338	39.0895	33.1054	0.7692	106.2680	37.0357	30.9844	0.7439	66.7619	40.8752	35.7077	0.7414
	GRU	73.5781	21.8581	20.6928	0.8552	104.8188	38.3769	29.1331	0.8213	96.5297	34.3541	29.8170	0.7628	66.5768	40.3966	33.3098	0.7750
	LSTM	80.7717	23.0682	21.1303	0.8490	114.9874	38.3904	30.5616	0.8033	104.5444	35.7512	30.9886	0.7438	68.1873	41.8047	35.9558	0.7378
	Transformer	81.1241	22.0003	21.2581	0.8472	96.5460	34.4869	28.3348	0.8309	95.2539	33.3290	29.5062	0.7678	66.8385	38.5422	35.3214	0.7470
	PatchTST	78.5047	23.6689	21.7759	0.8396	84.8937	31.4819	27.9287	0.8357	98.9263	34.3361	29.5281	0.7674	70.7175	38.8915	39.6316	0.6814
	TSGAN	**59.0974**	**20.7958**	**16.8688**	**0.9038**	**79.5132**	**29.5678**	**24.2747**	**0.8759**	**64.6891**	**25.3517**	**21.3197**	**0.8788**	**47.5525**	**31.9790**	**26.5137**	**0.8574**

3.3 Comparison with Single-Modality Image Models

Furthermore, this paper compares the proposed method with single-image-modality models, including CNN, ResNet, and CNN-LSTM [39] . These models directly extract cloud-texture and temporal-evolution information from continuous sky-image sequences, providing visual cues regarding rapid changes in irradiance. All models use the same image sequences and data partitions; the relevant results are presented in **Table 3** and **Figure 3**.

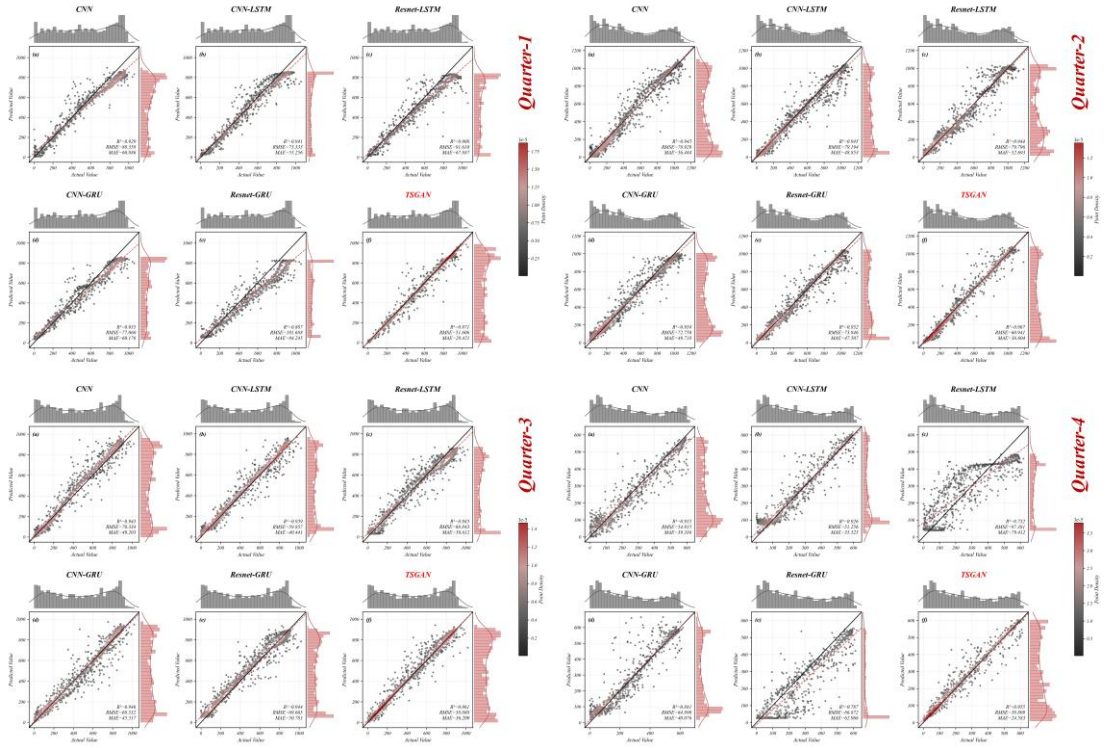


Figure 3. Performance comparison of image-based forecasting models

The figure compares the forecasting results of the proposed method and the single-image-modality models across different quarters. The scatter plots illustrate the agreement between the predicted and actual irradiance values, while the black dashed line represents perfect agreement. The closer the data points are to this line, the more closely the model's predictions match the actual values. The marginal histograms at the top, bottom, and right sides of the figure show the distributions of actual and predicted irradiance values across different ranges. These histograms can be used to assess whether the model captures the variation range and probability distribution of actual irradiance.

As shown in the figure, although single-image models such as CNN, ResNet, and CNN-LSTM can capture certain trends in irradiance variation by utilizing cloud structures and occlusion changes in sky images, their prediction points still exhibit some dispersion, with some deviating significantly from the ideal diagonal line. At the same time, discrepancies exist between the distributions of the predicted and actual values, indicating that when relying solely on visual information, the models' ability to

characterize the overall range of irradiance variation remains limited. This is primarily because image information reflects the current state of the sky, whereas irradiance is determined not only by cloud morphology but also by factors such as the sun's position and historical irradiance conditions. Identical or similar visual features may correspond to different irradiance responses, making single-image models prone to forecasting biases in complex environments. In contrast, the prediction points of the proposed method are generally more concentrated around the ideal diagonal line and exhibit greater consistency across different quarters. Furthermore, the distribution of its predicted values more closely resembles that of the actual values, indicating that the model not only reduces forecasting errors but also more accurately preserves the overall distributional characteristics of irradiance variation. This suggests that incorporating additional state information can mitigate the uncertainty introduced by a single visual representation, resulting in more stable forecasts across different irradiance levels.

The results in the table show that, compared with the image-based baseline models, the proposed method reduces MAE and NRMSE by 32.38% and 23.32%, respectively. These results further demonstrate that multimodal fusion can effectively address the limitations of single-image models in characterizing irradiance states, enabling the model to capture trends more accurately and produce more reasonable forecast distributions. From this perspective, the improvement achieved by the proposed method does not simply stem from an increase in the amount of input information but rather from the complementary constraints formed between different types of information. While single-image models primarily rely on visual patterns to determine changes in irradiance, the proposed method integrates additional information related to irradiance conditions, enabling the model to obtain a more comprehensive understanding of variation patterns under complex weather conditions. Consequently, multimodal fusion enhances the model's adaptability to abnormal fluctuations and different seasonal conditions, thereby improving the reliability of the forecasting results.

Table 3. Comparison of the proposed model with image-based methods

STEP	Model	Q1				Q2				Q3				Q4			
		MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2
Step-1	CNN	60.8863	19.4475	14.9276	0.9293	46.4366	20.0451	15.0090	0.9543	49.2031	18.6164	15.0043	0.9427	39.3357	31.4015	20.9847	0.9147
	ResNet-LSTM	67.8866	19.7962	17.0178	0.9081	52.6929	20.2368	16.5982	0.9441	50.6120	21.8449	14.6862	0.9451	79.4117	43.8879	37.2124	0.7318
	CNN-LSTM	55.2556	17.1375	13.6226	0.9411	48.8529	20.3098	16.4731	0.9450	40.4415	18.3717	12.7693	0.9585	35.5252	27.4414	19.5867	0.9257
	ResNet-GRU	84.2449	23.0642	18.8898	0.8868	47.5070	19.2686	15.3606	0.9522	50.7806	21.0146	14.8617	0.9438	62.9657	47.7757	33.1964	0.7865
	CNN-GRU	60.1757	17.6644	14.3149	0.9350	49.7185	20.6398	15.1326	0.9536	45.5170	18.2212	14.6199	0.9456	40.0758	28.4571	24.7963	0.8809
	TSGAN	**28.4211**	**10.7657**	**9.5866**	**0.9708**	**36.6041**	**20.1119**	**12.6762**	**0.9674**	**36.1998**	**14.6895**	**12.3918**	**0.9609**	**24.5830**	**17.4334**	**15.2121**	**0.9552**
Step-2	CNN	75.3770	20.9309	18.3015	0.8900	71.1133	29.1468	22.1796	0.8982	60.5951	24.4050	20.9891	0.8851	51.6569	35.4326	28.4207	0.8396
	ResNet-LSTM	110.5658	29.9702	25.9004	0.7797	74.5637	27.0218	23.4731	0.8860	74.1428	28.7704	22.7781	0.8647	50.6950	36.7209	28.5991	0.8376
	CNN-LSTM	84.9361	21.5693	19.5844	0.8740	75.9351	30.0893	22.3264	0.8969	63.3491	23.9593	21.0501	0.8844	50.0196	33.5476	27.5993	0.8488
	ResNet-GRU	114.3686	34.1274	30.5537	0.6934	80.7047	28.2466	24.4027	0.8768	71.7284	25.6310	21.5609	0.8788	56.1173	38.5184	31.1703	0.8071
	CNN-GRU	102.9355	28.5014	24.3959	0.8045	73.2372	28.5925	23.1134	0.8895	57.9530	23.5628	19.3614	0.9022	50.1919	32.3101	28.9391	0.8337
	TSGAN	**43.9107**	**15.0788**	**14.2682**	**0.9331**	**59.2346**	**22.4698**	**20.5965**	**0.9123**	**49.3443**	**22.3576**	**15.4098**	**0.9381**	**38.5968**	**24.3409**	**24.6437**	**0.8794**
Step-3	CNN	83.5889	23.2879	20.1045	0.8633	86.1066	33.8267	25.8874	0.8589	77.5286	30.0210	23.3805	0.8542	67.7925	42.3373	33.8331	0.7678
	ResNet-LSTM	90.1304	27.1278	20.7855	0.8539	103.8583	37.2314	30.2723	0.8070	82.9596	30.0554	23.9282	0.8473	67.9225	41.4289	34.9518	0.7522
	CNN-LSTM	85.9227	24.7037	19.9498	0.8654	105.5398	36.7112	29.5436	0.8162	76.0604	26.6169	23.4054	0.8539	71.1775	42.0150	36.2762	0.7331
	ResNet-GRU	109.9737	29.0672	25.5261	0.7796	86.6949	31.7819	26.1368	0.8561	89.3834	32.2895	26.5124	0.8125	67.0376	39.0241	33.6354	0.7705
	CNN-GRU	89.6667	25.8159	21.3057	0.8465	88.9447	34.8601	27.5067	0.8407	73.2801	26.1237	23.7445	0.8496	76.8135	44.3453	37.3280	0.7174
	TSGAN	**59.0974**	**20.7958**	**16.8688**	**0.9038**	**79.5132**	**29.5678**	**24.2747**	**0.8759**	**64.6891**	**25.3517**	**21.3197**	**0.8788**	**47.5525**	**31.9790**	**26.5137**	**0.8574**

3.4 Comparison with Other Multimodal Fusion Methods

To further analyze the impact of the fusion strategy itself, this paper compares the proposed method with approaches such as direct concatenation, gated fusion, and Cross-Attention [40] . Since the only differences among these methods lie in the modalities' interaction and joint representation, this comparison allows us to evaluate the effectiveness of the fusion mechanisms while essentially eliminating differences in the input information. The relevant results are shown in **Table 4**.

Overall, compared with the average performance of the other fusion methods, the proposed method reduced MAE, RMSE, NRMSE, and SMAPE by 18.39%, 12.15%, 13.21%, and 13.91%, respectively, while increasing the average R^2 by 0.0260. Even compared with the best-performing fusion baseline, the proposed method reduced the four error metrics by 15.48%, 10.43%, 11.69%, and 10.82%, respectively.

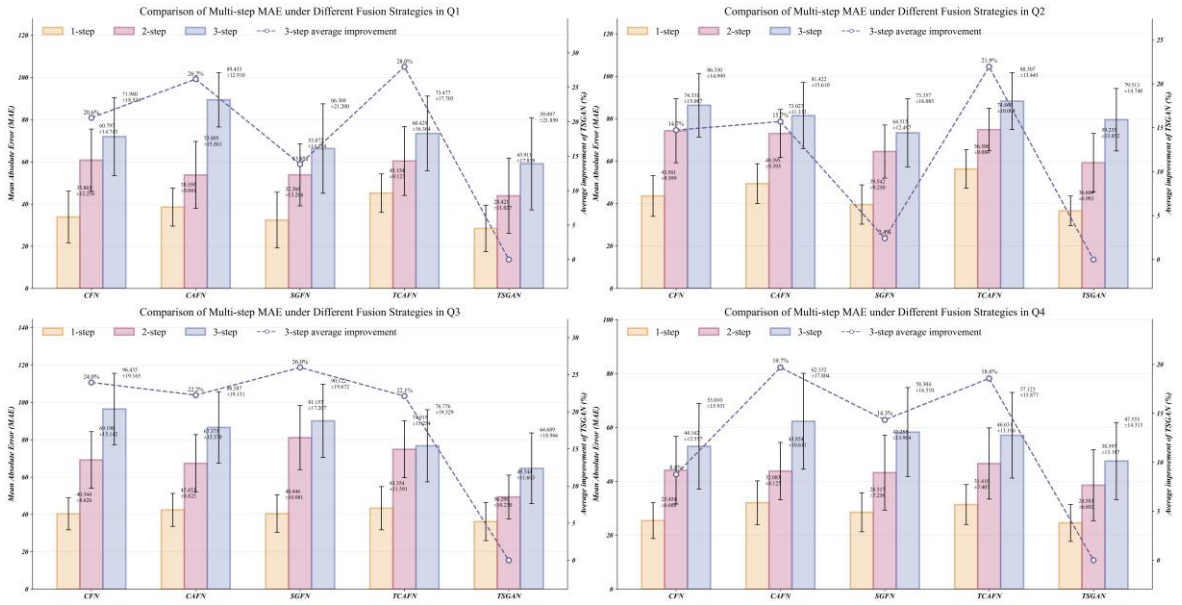


Figure 4 Prediction performance comparison under different fusion strategies

Figure 4 compares MAE of different multimodal fusion strategies across four quarters and three forecasting horizons. The bars represent the average error of each method at different forecasting horizons, the error bars reflect the variation in model forecasting errors across different experimental samples, and the dashed lines indicate the trend in model performance as the forecasting horizon increases. As shown in the figure, compared with the single-modality methods, all multimodal fusion strategies achieved some degree of performance improvement, indicating clear complementarity between image and numerical information. However, significant differences remain among the fusion methods, indicating that multimodal forecasting performance depends not only on the incorporation of multiple information sources but also on how information is exchanged between modalities. Compared with the other fusion strategies, the proposed method maintains lower error levels across all four quarters and

forecasting horizons, while its error bars are generally narrower. This suggests that the method not only improves forecasting accuracy but also reduces performance fluctuations across different scenarios.

Analysis of error variations across different fusion strategies reveals that, although some methods can reduce the average forecasting error by combining image and numerical information, they still exhibit significant fluctuations across different quarters and forecasting horizons. For example, certain fusion strategies achieve good forecasting results at Step 1, but as the forecasting horizon increases, their errors grow markedly, indicating that fixed fusion methods struggle to sustain effective information utilization over longer forecasting horizons. In contrast, the proposed method maintains a low error level even for Step 3 forecasts, with the average R^2 remaining high, indicating that this method can effectively mitigate error accumulation in longer-horizon forecasts. This is primarily because the proposed fusion strategy does not simply concatenate features from different modalities; rather, it adjusts the contributions of different types of information based on the current input state, enabling the model to focus more on effective features associated with current trends during the forecasting process.

Furthermore, the error bars of the proposed method are generally narrow, indicating greater stability across different quarters and weather conditions. Compared with fusion methods that exhibit significant increases in error under complex scenarios, the proposed method reduces the overall MAE by approximately 18.39%, while also exhibiting less performance variation across different experimental settings. This suggests that the dynamic fusion mechanism can reduce information bias caused by a single fixed fusion approach, enabling the model to maintain more consistent forecasting capabilities under different environmental conditions. Particularly under complex irradiance variations, the model prevents any single type of information from excessively dominating the forecasting process, thereby reducing forecasting bias in anomalous scenarios. Traditional multimodal fusion methods typically assume that different modalities have fixed levels of importance across all scenarios; however, in actual forecasting, the reliability of information from different sources varies with weather conditions and temporal scales. When the input information is highly relevant to the current state, a fixed fusion method can yield good forecasting results; however, when the validity of certain information decreases, excessive fusion may introduce additional interference. The proposed fusion strategy dynamically adjusts the information-interaction relationships between modalities, enabling the model to utilize features from different sources more flexibly and thereby enhancing the effectiveness of the joint representation.

Consequently, the experimental results demonstrate that improvements in multimodal forecasting performance stem not only from the joint input of image and numerical information but, more importantly, from effective synergy between the different modalities. Through dynamic information organization, the proposed method reduces forecasting errors while enhancing model stability across different quarters and forecasting horizons, thereby validating the effectiveness of the proposed fusion strategy in complex irradiance forecasting tasks.

Table 4. Comparison of the proposed model with other fusion methods

STEP	Model	Q1				Q2				Q3				Q4			
		MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2	MAE	SMAPE	NRMSE	R^2
Step-1	CFN	33.8629	11.7939	11.2062	0.9602	43.5615	20.4751	13.9736	0.9604	40.3441	18.3517	12.7362	0.9587	25.4541	16.7763	16.3730	0.9481
	CAFN	38.5501	12.2260	11.0842	0.9610	49.3951	22.0236	15.1711	0.9533	42.4327	17.5376	13.5038	0.9536	32.0827	21.2089	20.3811	0.9195
	SGFN	32.3662	12.5895	11.1414	0.9606	39.5424	15.6119	14.1757	0.9593	40.4458	18.2122	13.2014	0.9557	28.5172	19.3743	17.8619	0.9382
	TCAFN	45.1537	13.5854	12.2421	0.9524	56.3061	25.4211	16.3367	0.9459	43.3537	19.5663	14.4369	0.9470	31.4102	23.6310	18.2674	0.9354
	TSGAN	**28.4211**	**10.7657**	**9.5866**	**0.9708**	**36.6041**	**20.1119**	**12.6762**	**0.9674**	**36.1998**	**14.6895**	**12.3918**	**0.9609**	**24.5830**	**17.4334**	**15.2121**	**0.9552**
Step-2	CFN	60.7973	18.0223	16.0775	0.9151	74.3315	27.6418	22.3132	0.8970	69.1981	26.5466	21.5787	0.8786	44.1622	29.3326	26.6717	0.8588
	CAFN	53.8047	17.2609	15.1621	0.9245	73.0267	28.9266	21.4064	0.9052	67.3755	27.6483	21.1848	0.8830	43.8535	28.5826	25.8506	0.8673
	SGFN	53.8722	20.7055	15.0869	0.9252	64.5147	23.6376	21.1750	0.9073	81.1567	31.3069	24.0044	0.8497	43.2551	27.6386	26.7824	0.8576
	TCAFN	60.4293	18.1699	16.7945	0.9074	74.8670	29.4463	22.6267	0.8941	74.9146	30.9109	22.0610	0.8731	46.6308	30.4362	27.9863	0.8445
	TSGAN	**43.9107**	**15.0788**	**14.2682**	**0.9331**	**59.2346**	**22.4698**	**20.5965**	**0.9123**	**49.3443**	**22.3576**	**15.4098**	**0.9381**	**38.5968**	**24.3409**	**24.6437**	**0.8794**
Step-3	CFN	71.9797	20.1198	18.6624	0.8822	86.3303	31.9057	25.1329	0.8670	96.4355	33.7121	27.5757	0.7972	53.0098	37.6579	31.2857	0.8015
	CAFN	89.4325	22.3617	20.9417	0.8517	81.4223	27.3142	25.0953	0.8674	86.5868	32.5640	25.4438	0.8273	62.3322	39.2589	33.5386	0.7719
	SGFN	66.3045	21.0048	17.5596	0.8957	**73.3572**	**27.4594**	**22.5308**	**0.8931**	90.1218	33.1891	25.7172	0.8236	58.3039	35.1412	32.8347	0.7813
	TCAFN	73.4771	23.7718	18.3604	0.8860	88.3074	31.9208	26.3327	0.8540	76.7755	27.8649	24.0623	0.8456	57.1233	34.4008	31.6957	0.7962
	TSGAN	**59.0974**	**20.7958**	**16.8688**	**0.9038**	79.5132	29.5678	24.2747	0.8759	**64.6891**	**25.3517**	**21.3197**	**0.8788**	**47.5525**	**31.9790**	**26.5137**	**0.8574**

4. DISCUSSION

4.1 Seasonal Variation Analysis

Solar irradiance forecasting performance is influenced not only by model structure but also by the combined effects of seasonal variations in the Sun's motion patterns and changes in cloud-evolution characteristics. The experimental results across the four quarters show that different models exhibit significant performance differences during seasonal transitions, indicating that irradiance time series exhibit strong seasonal dependence. Their variation patterns are not determined entirely by historical trends but are dynamically modulated by external environmental conditions. Overall, in Q1 and Q4, because of the lower solar altitude angle, irradiance levels are generally lower, and variation trends are governed more strongly by patterns of solar motion. Historical irradiance time series can therefore provide relatively stable trend information, allowing models that rely partly on time-series characteristics to maintain a certain level of forecasting capability. Accordingly, TSGAN's modality-selection mechanism tends to increase the contribution of historical time-series features. However, in Q2 and Q3, as solar radiation intensifies and local weather activity becomes more frequent, cloud formation and dissipation increasingly influence irradiance variations, causing the irradiance time series to exhibit more pronounced nonstationarity and random fluctuations. Under these conditions, models that rely solely on historical data struggle to promptly detect future changes in environmental conditions and are prone to interpreting sudden decreases or recoveries in irradiance as anomalous disturbances, thereby leading to forecasting lags and error accumulation. **Figure 5** illustrates the variation in forecasting performance across different quarters.

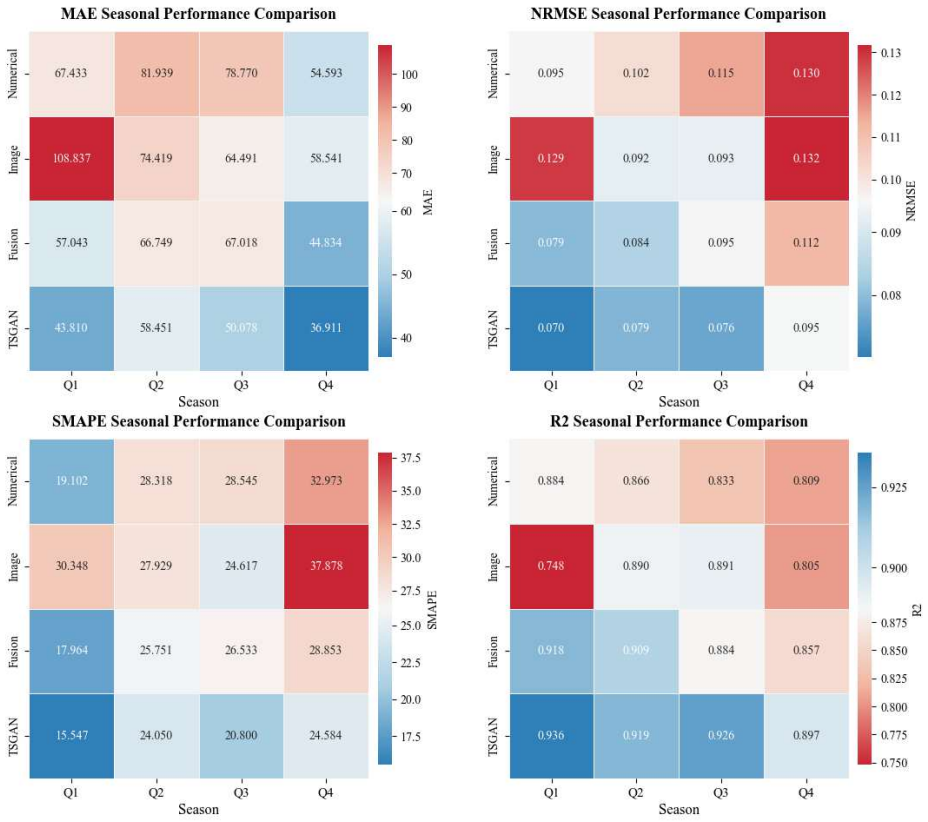


Figure 5 Comparison of the proposed model with different models under seasonal conditions

As shown in **Figure 5**, seasonal variations have a relatively significant impact on the forecasting performance of different types of models; however, the proposed TSGAN consistently maintains optimal or near-optimal performance across all four quarters. Specifically, whether evaluated using error metrics such as MAE, NRMSE, and SMAPE or the goodness-of-fit metric R^2, TSGAN outperforms the numerical, image-based, and general multimodal fusion methods. At the same time, TSGAN exhibits relatively small performance fluctuations across different quarters; for example, its NRMSE consistently remains low, while its R^2 remains high. This indicates that although seasonal variations alter irradiance levels and weather conditions, they do not significantly diminish the model's relative advantage. In contrast, the single-image model exhibits significant fluctuations in Q1 and Q4, indicating that the predictive validity of sky images is easily affected by changes in the Sun's altitude and seasonal cloud conditions. Although the single-numerical-modality model is generally more stable, its error gradually increases in Q2, Q3, and Q4, indicating that when relying solely on historical irradiance and meteorological time series, the model's adaptability to rapid cloud changes and seasonal distributions remains limited. General fusion methods can mitigate the information deficiencies of single-modality approaches to some extent, but their performance still fluctuates significantly with seasonal variations, indicating that simple concatenation or fixed fusion cannot consistently ensure the effective use of information from different modalities.

TSGAN demonstrates greater stability across different seasons, primarily because its modality-selection mechanism dynamically adjusts the contributions of the two modalities based on current weather conditions and irradiance changes. When irradiance variations are relatively stable, the model can rely more heavily on historical numerical sequences to maintain the irradiance baseline and continuity of variation; when rapid irradiance fluctuations occur because of cloud movement or dissipation, the model can enhance the contribution of cloud-structure and spatial-variation information from sky images. Through this state-dependent modality filtering, the model reduces interference from lagging numerical information and conflicts between modalities in the joint representation, enabling multisource information to achieve more effective complementarity and coupling under different seasonal conditions. Therefore, TSGAN's performance improvement does not stem solely from an increase in the amount of input information but more likely from its ability to dynamically identify the reliability of different modalities under seasonal changes and maintain more stable and effective information synergy under complex weather conditions.

This dynamic adjustment capability distinguishes TSGAN from traditional fixed-fusion methods. Traditional fusion methods typically assume that the importance of different modalities remains constant, which may lead to the inclusion of low-reliability information in the forecasting process during seasonal changes. For example, under the low-light conditions observed in Q1, image-texture information may be insufficient; maintaining a high weight for visual information may therefore introduce noise. Conversely, under conditions of intense cloud variation in summer, excessive reliance on historical time series prevents timely responses to sudden decreases in irradiance. Through its modality-selective attention mechanism, TSGAN automatically adjusts modality contributions based on the current environmental state, achieving a transition from "fixed fusion" to "state-aware fusion." Therefore, TSGAN's performance improvement does not stem solely from an increase in the amount of input information but more likely from its ability to dynamically identify the reliability of different modalities under seasonal changes and maintain more stable and effective information synergy under complex weather conditions.

4.2 DM Test Results

To further verify the statistical reliability of the performance improvements observed above, this

paper employs the Diebold-Mariano (DM) test to conduct pairwise comparisons of forecasting errors between TSGAN and the main comparison models under different prediction horizons. The absolute error is used as the loss function, and the results are presented in **Table 5.**

The test results show that more than 98% of the comparisons are statistically significant at both the 10% and 5% levels. Notably, 96.88% of the comparisons remain statistically significant at the 1% level. These results indicate that TSGAN's forecasting performance improvement relative to the comparison models does not stem from individual samples or random variation but is statistically significant. Furthermore, the significance results remain highly consistent across different quarters and forecasting horizons, indicating that seasonal variations and longer forecasting horizons do not diminish the model's relative advantage. Therefore, the DM test results further support the preceding experimental conclusions from a statistical inference perspective and validate the effectiveness of the proposed modality-selective attention-gating mechanism in coordinating image information with numerical states and enhancing forecasting stability and reliability.

Table 5 DM test results of the proposed model compared with benchmark models.

Model	Q1			Q2			Q3			Q4		
	1-step	2-step	3-step	1-step	2-step	3-step	1-step	2-step	3-step	1-step	2-step	3-step
MLP	-12.7788***	-11.7851***	-10.0271***	-9.3577***	-10.5347***	-8.0600***	-18.8208***	-12.5396***	-11.4990***	-10.6755***	-8.1341***	-6.7784***
GRU	-6.4584***	-6.5432***	-5.0078***	-8.5068***	-8.3165***	-6.3110***	-15.5039***	-9.2549***	-9.1397***	-8.8954***	-5.5422***	-8.4642***
LSTM	-10.4720***	-10.9580***	-7.0116***	-8.4164***	-9.1363***	-9.1489***	-11.6417***	-12.8344***	-10.6713***	-11.0516***	-9.0427***	-6.8990***
RF	-11.8516***	-8.6125***	-8.1876***	-9.3329***	-9.3748***	-8.6126***	-7.4249***	-8.1984***	-6.9713***	-7.9269***	-8.1162***	-7.3516***
XGBoost	-14.1962***	-9.9908***	-9.2250***	-12.2544***	-9.5931***	-9.1045***	-10.3229***	-10.7285***	-8.4087***	-10.6161***	-8.4612***	-8.5239***
Transformer	-10.2675***	-8.7142***	-7.1272***	-7.9009***	-5.6173***	-5.3294***	-13.7082***	-13.7975***	-8.7120***	-12.7152***	-6.0573***	-7.0449***
PatchTST	-9.1898***	-6.7949***	-5.7098***	-8.2501***	-7.7571***	-1.6850**	-9.0908***	-9.8981***	-11.3183***	-8.4704***	-10.1418***	-6.9812***
CNN	-17.1114***	-23.1374***	-8.0157***	-12.0363***	-9.2743***	-2.8404***	-9.2300***	-6.8245***	-3.7432***	-11.6160***	-6.3098***	-10.4385***
CNN-LSTM	-17.6320***	-23.5472***	-10.0034***	-6.5151***	-5.9130***	-3.5773***	-3.0690***	-3.5580***	-3.1681***	-9.8048***	-6.3174***	-10.9104***
ResNet	-17.8175***	-24.2792***	-12.3518***	-9.8929***	-7.5286***	-2.5580***	-10.1358***	-8.3019***	-7.8416***	-27.6246***	-7.8710***	-7.3029***
CNN-LSTM_1	-18.9347***	-13.4203***	-8.1658***	-9.1431***	-6.6870***	-6.0934***	-5.7215***	-5.4476***	-3.4185***	-10.3610***	-5.9171***	-8.7862***
ResNet_1	-24.9044***	-16.8316***	-8.4329***	-6.4990***	-5.6567***	-5.4960***	-9.2932***	-8.8533***	-6.3484***	-19.8362***	-5.8865***	-7.9785***
Concat	-4.2220***	-8.4955***	-4.9562***	-4.9931***	-6.5308***	-2.5532***	-4.6039***	-8.5471***	-11.0671***	-0.9781	-3.2117***	-2.1690**
Cross-Attention	-7.6372***	-4.4062***	-8.5703***	-8.4485***	-6.0782***	-0.6136	-5.2414***	-7.8175***	-7.9000***	-7.0235***	-2.9125***	-6.8276***
Gated Fusion	-2.9367***	-4.3878***	-3.5406***	-2.2395**	-2.6934***	2.4160	-3.7449***	-13.9326***	-8.8966***	-3.7641***	-2.7014***	-4.6332***
Patch Fusion	-11.1843***	-6.4616***	-5.6044***	-13.3398***	-7.0923***	-3.3013***	-6.2304***	-10.9694***	-4.7956***	-7.1169***	-4.7417***	-4.2579***

Note: ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The DM statistic is calculated based on absolute error loss and uses a one-tailed test; a negative value indicates that TSGAN has a lower prediction loss than the corresponding comparison model.

5. CONCLUSION

This paper proposes a multimodal adaptive fusion framework for ultra-short-term solar irradiance forecasting. By jointly modeling historical irradiance time series and dynamic cloud information from sky images, the framework effectively overcomes the limitations of traditional single-source forecasting methods, which struggle to capture rapid fluctuations caused by local cloud changes under complex weather conditions. Rather than simply adding input modalities, this method establishes deep correlations between numerical state evolution and changes in the visual environment, enabling the model to simultaneously learn historical patterns of irradiance variation and potential future perturbations, thereby enhancing forecasting accuracy, stability, and adaptability to complex scenarios. Based on systematic experimental validation across different quarters, forecasting horizons, and multiple benchmark models, this paper draws the following main conclusions:

(1) The unified state encoder proposed in this paper effectively addresses the limited ability of traditional numerical baseline models to detect abrupt changes induced by localized clouds. Experimental results show that although models relying solely on historical irradiance and meteorological variables can effectively learn the periodic variation patterns of solar irradiance, they depend primarily on past state information and lack the ability to anticipate sudden irradiance changes caused by factors such as cloud cover and rapid cloud movement. Consequently, when the forecasting horizon is extended or weather conditions become more complex, a single numerical-modality model is prone to forecasting lags and cumulative errors. In contrast, by incorporating additional visual information about the sky, the proposed method enables the model not only to capture temporal dependencies in irradiance sequences but also to utilize external environmental states to anticipate future changes, thereby addressing, at the mechanistic level, the lack of leading information in traditional time-series models.

(2) The multimodal joint representation method proposed in this paper effectively addresses the limitations of single-image models in anchoring absolute irradiance levels and maintaining historical continuity. Although visual models such as CNNs and ResNets can extract dynamic features—such as solar occlusion and cloud movement—from sky images, they struggle to establish a stable mapping between visual states and absolute irradiance levels when relying solely on image information. As a result, their forecasts are susceptible to variations in the Sun's position and image quality. Experimental results indicate that visual features alone cannot fully address the irradiance forecasting problem. Images primarily describe the current state of the sky but lack constraints from the historical evolution of irradiance and patterns of solar motion, potentially leading to inconsistencies between visual states and actual irradiance responses. The proposed model integrates the historical physical patterns provided by the numerical modality with the prior information on future changes provided by the image modality. This transforms the forecasting process from simple state matching into comprehensive modeling of irradiance variation mechanisms, thereby effectively reducing the forecasting bias caused by a single visual representation.

(3) The state-aware adaptive fusion mechanism proposed in this paper effectively overcomes the limitation of traditional fusion mechanisms in responding to changes in modality reliability. Existing multimodal methods typically employ feature concatenation, fixed-weight fusion, or shallow interaction approaches. Although these methods can simultaneously utilize information from multiple sources, they assume that the relative importance of different modalities remains fixed across weather

conditions, making it difficult to adapt to dynamic changes in modality reliability during actual forecasting. Simple fusion methods struggle to fully exploit the coupling relationships between modalities, whereas an effective cross-modal interaction mechanism is essential for improving forecasting performance. The proposed TSGAN further transforms the fusion process into a state-driven information-filtering process, enabling the model to dynamically adjust the contributions of the numerical and image modalities based on current irradiance changes. Under clear-sky or stable conditions, the model can make greater use of continuity information from historical time series, whereas under rapidly changing cloud cover or local shading conditions, it enhances the contribution of visual information, thereby reducing the information bias caused by fixed fusion strategies. Experimental results show that, compared with other multimodal fusion methods, the proposed method maintains lower error levels across all four quarters and different forecasting horizons while exhibiting narrower error ranges, indicating that this mechanism not only improves forecasting accuracy but also enhances cross-scenario stability.

The TSGAN proposed in this paper offers a forecasting approach that shifts from “information superposition” to “state selection” for the collaborative modeling of ground-based sky images and historical numerical time series. It can provide more timely irradiance information for photovoltaic power smoothing, energy-storage control, and short-term grid dispatch. The current findings remain limited by the scope of the data and the forecasting horizons. Future research could incorporate continuous cloud-motion characterization and satellite or infrared imagery and evaluate the model’s applicability under distribution shifts and uncertainty quantification through cross-site validation and probabilistic interval forecasting.

Credit Author Statement

Houhang Zhao: Writing – original draft, Software, Investigation, Formal analysis. Aiyi Song: Writing – review & editing, Resources. Yu Jin: Writing – original draft, Formal analysis, Conceptualization, Funding acquisition. Haoxiang Yu: Writing – review & editing, Supervision, Methodology.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Major Program of National Social Science Foundation of China (Grant No. 22&ZD165)

Data Availability Statement

The dataset used in this study is publicly available from the Open Energy Data Initiative (OEDI) platform. The experimental data are obtained from the publicly released abq_data subset, which can be accessed at https://data.openei.org/submissions/8414. The dataset contains ground-based meteorological observations, solar irradiance measurements, and corresponding sky images, providing multimodal data for solar irradiance forecasting research.

REFERENCES

[1] V. Gayathry, D. Kaliyaperumal, and S. R. Salkuti, "Seasonal solar irradiance forecasting using artificial intelligence techniques with uncertainty analysis," *Scientific Reports,* vol. 14, no. 1, p. 17945, 2024/08/02 2024.

[2] D. Hua, J. Zeng, Q. Lin, S. Cui, Y. Liu, and Y. Liang, "An expectile-based framework for risk-calibrated credible capacity evaluation of virtual power plants under wind and PV forecast uncertainties," *Scientific Reports,* vol. 16, no. 1, p. 15253, 2026/03/31 2026.

[3] N. A. Elmunim *et al.*, "Enhancing wind and solar energy forecasting through time-series feature engineering and ensemble machine learning," *Scientific Reports,* vol. 16, no. 1, p. 15546, 2026/05/19 2026.

[4] H. Li, H. Guo, J. Wang, Y. Wang, and C. Wu, "Tourism combination forecasting with swarm intelligence," *Annals of Tourism Research,* vol. 111, p. 103932, 2025/03/01/ 2025.

[5] J. Tian, R. Ooka, and D. J. J. o. C. P. Lee, "Multi-scale solar radiation and photovoltaic power forecasting with machine learning algorithms in urban environment: A state-of-the-art review," vol. 426, p. 139040, 2023.

[6] H. Guo, S. Bi, Y. Jin, H. Zhao, and Y. Ai, "Carbon Price Prediction With Public Social Media Big Data and an Interpretable Multi-Objective Intelligent Feature Optimization Strategy," *Journal of Forecasting,* vol. n/a, no. n/a, 2026/01/27 2026.

[7] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy,* vol. 83, no. 10, pp. 1772-1783, 2009/10/01/ 2009.

[8] H. T. C. Pedro and C. F. M. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," *Solar Energy,* vol. 86, no. 7, pp. 2017-2028, 2012/07/01/ 2012.

[9] H. Sharadga, S. Hajimirza, and R. Balog, "Time series forecasting of solar power generation for large-scale photovoltaic plants," *Renewable Energy,* vol. 150, 12/01 2019.

[10] T. N. Nguyen and F. Müsgens, "What drives the accuracy of PV output forecasts?," *Applied Energy,* vol. 323, p. 119603, 2022/10/01/ 2022.

[11] A. Mellit, A. M. Pavan, and V. J. R. E. Lughi, "Deep learning neural networks for short-term photovoltaic power forecasting," vol. 172, pp. 276-288, 2021.

[12] A. A. du Plessis, J. M. Strauss, and A. J. Rix, "Short-term solar power forecasting: Investigating the ability of deep learning models to capture low-level utility-scale Photovoltaic system behaviour," *Applied Energy,* vol. 285, p. 116395, 2021/03/01/ 2021.

[13] G. Xie, Z. Zhang, S. Xie, Z. Yuan, and H. Liu, "CPWformer-DEC: improved Transformer with class-priority weather attention and dynamic error compensation for photovoltaic power forecasting," *Expert Systems with Applications,* vol. 301, p. 130580, 2026/03/10/ 2026.

[14] A. L. Jonathan *et al.*, "A multimodal deep learning approach for very short-term solar forecasts using sky images and historical numerical data," *Renewable Energy,* vol. 255, p. 123774, 2025/12/15/ 2025.

[15] B. Yong, Y. Zhang, J. Shen, A. Ren, X. Zhou, and Q. Zhou, "ConvODE-Mixer: A multimodal deep learning model for ultra-short-term PV power forecasting," *Solar Energy,* vol. 300, p. 113777, 2025/11/01/ 2025.
[16] B. Xiong *et al.*, "Multimodal ultra-short-term probabilistic solar power forecasting with generative AI and Transformer," *Advances in Applied Energy,* p. 100250, 2025.
[17] Z. Zhang, X. Huang, C. Li, F. Cheng, and Y. J. E. Tai, "CRAformer: A cross-residual attention transformer for solar irradiation multistep forecasting," vol. 320, p. 135214, 2025.
[18] J. Liu, H. Zang, L. Cheng, T. Ding, Z. Wei, and G. Sun, "A Transformer-based multimodal-learning framework using sky images for ultra-short-term solar irradiance forecasting," *Applied Energy,* vol. 342, p. 121160, 2023.
[19] Y. Nie *et al.*, "Sky image-based solar forecasting using deep learning with heterogeneous multi-location data: Dataset fusion versus transfer learning," *Applied Energy,* vol. 369, p. 123467, 2024.
[20] M. Mouhafid, Y. Zhou, C. Shan, and Z. Xiao, "A Robust Approach to Early Glaucoma Identification from Retinal Fundus Images using Dirichlet-based Weighted Average Ensemble and Bayesian Optimization," *Current Medical Imaging,* vol. 21, 2025/01/01/ 2025.
[21] D. Stefanov and J. Demsar, "Forecasting solar power production by using satellite images," *Solar Energy,* vol. 300, p. 113823, 2025.
[22] Z. Si, M. Yang, Y. Yu, and T. J. A. E. Ding, "Photovoltaic power forecast based on satellite images considering effects of solar position," vol. 302, p. 117514, 2021.
[23] Y. Xie, J. Zheng, F. Mei, G. Taylor, and A. J. A. E. Gao, "An efficient approach for regional photovoltaic power forecasting optimization based on texture features from satellite images and transfer learning," vol. 385, p. 125505, 2025.
[24] W. Kong, Y. Jia, Z. Y. Dong, K. Meng, and S. J. A. E. Chai, "Hybrid approaches based on deep whole-sky-image learning to photovoltaic generation forecasting," vol. 280, p. 115875, 2020.
[25] M. Ajith and M. J. A. E. Martínez-Ramón, "Deep learning based solar radiation micro forecast by fusion of infrared cloud images and radiation data," vol. 294, p. 117014, 2021.
[26] J. O. Kamadinata, T. L. Ken, and T. J. R. E. Suwa, "Sky image-based solar irradiance prediction methodologies using artificial neural networks," vol. 134, pp. 837-845, 2019.
[27] C. Feng, J. Zhang, W. Zhang, and B.-M. J. A. E. Hodge, "Convolutional neural networks for intra-hour solar forecasting based on sky image sequences," vol. 310, p. 118438, 2022.
[28] J. Zhang, R. Verschae, S. Nobuhara, and J.-F. J. S. E. Lalonde, "Deep photovoltaic nowcasting," vol. 176, pp. 267-276, 2018.
[29] Q. Paletta, G. Arbod, and J. J. S. E. Lasenby, "Benchmarking of deep learning irradiance forecasting models from sky images–An in-depth analysis," vol. 224, pp. 855-867, 2021.
[30] Q. Zhang *et al.*, "Multimodal bidirectional MAMBA framework for ultra-short-

term photovoltaic power forecasting using sky images," *Renewable Energy,* vol. 256, p. 124361, 2025.
[31] T. Baltrusaitis, C. Ahuja, L.-P. J. I. t. o. p. a. Morency, and m. intelligence, "Multimodal machine learning: A survey and taxonomy," vol. 41, no. 2, pp. 423-443, 2019.
[32] L. Zhang, R. Wilson, M. Sumner, and Y. J. R. E. Wu, "Advanced multimodal fusion method for very short-term solar irradiance forecasting using sky images and meteorological data: A gate and transformer mechanism approach," vol. 216, p. 118952, 2023.
[33] M. Wang, R. Cao, H. Xie, Y. Jiang, Q. Huang, and X. Ou, "Ultra-short-term solar irradiance forecasting based on cloud-type guidance and multimodal fusion," *Solar Energy,* vol. 305, p. 114244, 2026/02/01/ 2026.
[34] D. M. Riley, C. A. Zapotocky, and B. K. Byford, "A Data Set for Coincident Sky Images, Hemispherical Solar Irradiance, and Weather in Albuquerque, NM and Eugene, OR," Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States)2025.
[35] Y. Xu, S. Zheng, Q. Zhu, K.-c. Wong, X. Wang, and Q. Lin, "A complementary fused method using GRU and XGBoost models for long-term solar energy hourly forecasting," *Expert Systems with Applications,* vol. 254, p. 124286, 2024/11 2024.
[36] X. Xu, L. Guan, Z. Wang, R. Yao, and X. Guan, "A double-layer forecasting model for PV power forecasting based on GRU-Informer-SVR and Blending ensemble learning framework," *Applied Soft Computing,* vol. 172, p. 112768, 2025/03/01/ 2025.
[37] B. Xiong *et al.*, "Multimodal ultra-short-term probabilistic solar power forecasting with generative AI and transformer," p. 100250, 2025.
[38] G. Piantadosi *et al.*, "Photovoltaic power forecasting: A Transformer based framework," vol. 18, p. 100444, 2024.
[39] H. Luo, C. Tong, W. Gu, and Z. Li, "Time-series imaging for improving the accuracy of short-term load forecasting," *Energy,* vol. 333, p. 137282, 2025/10/01/ 2025.
[40] K. Wang, S. Shan, W. Dou, H. Wei, K. J. E. C. Zhang, and Management, "A cross-modal deep learning method for enhancing photovoltaic power forecasting with satellite imagery and time series data," vol. 323, p. 119218, 2025.