FusionSF: Fuse Heterogeneous Modalities in a Vector Quantized Framework for Robust Solar Power Forecasting

Ziqing Ma*
maziqing.mzq@alibaba-inc.com
Alibaba Group
China

Wenwei Wang*
duoluo.www@alibaba-inc.com
Alibaba Group
China

Tian Zhou*
tian.zt@alibaba-inc.com
Alibaba Group
China

Chao Chen
cc410784@alibaba-inc.com
Alibaba Group
China

Bingqing Peng
pengbingqing.pbq@alibaba-inc.com
Alibaba Group
China

Liang Sun
liang.sun@alibaba-inc.com
Alibaba Group
USA

Rong Jin†
rongjinemail@gmail.com
Alibaba Group
USA

ABSTRACT

Accurate solar power forecasting is crucial to integrate photovoltaic plants into the electric grid, schedule and secure the power grid safety. This problem becomes more demanding for those newly installed solar plants which lack sufficient data. Current research predominantly relies on historical solar power data or numerical weather prediction in a single-modality format, ignoring the complementary information provided in different modalities. In this paper, we propose a multi-modality fusion framework to integrate historical power data, numerical weather prediction, and satellite images, significantly improving forecast performance. We introduce a vector quantized framework that aligns modalities with varying information densities, striking a balance between integrating sufficient information and averting model overfitting. Our framework demonstrates strong zero-shot forecasting capability, which is especially useful for those newly installed plants. Moreover, we collect and release a multi-modal solar power (MMSP) dataset from real-world plants to further promote the research of multi-modal solar forecasting algorithms. Our extensive experiments show that our model not only operates with robustness but also boosts accuracy in both zero-shot forecasting and scenarios rich with training data, surpassing leading models. We have incorporated it into our eForecaster platform and deployed it for more than 300 solar plants with a capacity of over 15GW. Our code and dataset are accessible at https://anonymous.4open.science/r/FusionSF-770F/.

---

*authors contributed equally to this research.
†The author now works at Meta Platforms, Inc.

CCS CONCEPTS

• **Computing methodologies → Neural networks**.

KEYWORDS

Solar Power Forecasting, Modality Fusion, Vector Quantization, Zero-Shot Learning

1 INTRODUCTION

Solar photovoltaic (PV) plants serve as important contributors to the renewable energy sector, offering significant potential for sustainable energy generation [7, 10, 35]. Accurate solar power forecasting is crucial to balance the electricity supply and demand, and integrate the PV plants into the electricity grid [26].

Solar power forecasting differs from the traditional time series (TS) forecasting problem due to its heavy reliance on weather conditions, especially solar irradiation, cloud cover, temperature, and other meteorological factors [35]. The dynamic processes of these factors follow several physical principles, which are intrinsically complicated and unobserved or only partially observed, thus difficult to be captured by solely historical power series. This indicates that learning only from the historical pattern is usually insufficient in this scenario [5]. On the other hand, a critical problem in solar power forecasting is the noisy historical data and even lack of historical data [26], which is especially true for those newly installed PV plants. In this case, how to build and deploy accurate forecasting models given limited data remains challenging.

To address such challenges, the introduction of additional modalities in addition to historical power data becomes essential. In the realm of solar power forecasting, we deal primarily with three types of information: historical observed inputs, historical observed covariates, and future predicted covariates. Among these, future

predicted covariates, such as numerical weather predictions (NWP), are often considered the most crucial for accurate solar power forecasting [14]. Additionally, historical covariates, including ground-based all-sky camera images, data collected by instruments onboard satellite, and remote-sensing data [35], prove to be extremely valuable. However, the practical application of these technologies is sometimes constrained, as sky cameras and remote sensors are not universally available at solar power plants. On the other hand, satellite images are photos of the Earth taken by imaging satellites, usually in a high resolution and broad geographic area, and several weather phenomena, such as cloud thickness, can be retrieved from this observed data [3]. Although satellite images capture the true contextual information about the Earth in real-time, they cannot provide future predictions for subsequent days. Conversely, NWP data, generated by physics-informed weather models, offer future predictions of meteorological variables. However, they are usually of coarse granularity, and their precision may vary over different variables and weather conditions [35]. Consequently, for practical deployment, we select a combination of NWP, satellite imagery, and historical power data to serve as complementary data sources that effectively address the challenging day-ahead (short-term) solar power forecasting problem.

The next challenge arising in fusing multi-modal data is how to effectively extract and combine valuable information from heterogeneous sources, each with distinct characteristics. For instance, the satellite images are characterized by high volume, yet contain sparse information [29]. Conversely, NWP data are dense in information but often come with systematic biases [35]. Solar power historical data typically suffer from noise contamination [18, 35]. To address this challenge, we propose a Transformer-based architecture that exploits vector quantization (VQ). Our empirical studies in Section 5.4 verified that the VQ layers **help align the distributions of different modalities for better fusion** and **help reduce noise**.

The deployment on numerous newly established solar plants often presents another challenge in terms of limited historical data availability [26, 35]. One appealing feature of our proposed Transformer-based framework is its capability of zero-shot learning by leveraging data from various solar plants. A detailed experiment in Section 5.2 demonstrates **heterogeneous modality might be the key to zero-shot learning**. The complementary nature of diverse data modalities bolsters the model's robustness, as they provide a more comprehensive understanding of the underlying patterns that a single modality alone may not capture.

Moreover, we conduct an in-depth study in Section 5.5 to demonstrate the necessity and efficacy of our trimodality fusion paradigm, even in the context of high-accuracy numerical weather predictions. In Section 5.6, we provide a detailed description of the extensive deployment of FusionSF based on our eForecaster platform [44].

Our contributions are summarized as follows:

(1) We present a multi-modality fusion framework (FusionSF) for short-term solar power forecasting which outperforms contemporary SOTA models and our latest deployed baseline model with an improvement of 30.6% and 9.5%, respectively.

(2) We show our model's strong potential for zero-shot forecasting, thanks to the integration and alignment of multiple modalities. This strategy allows emerging solar plants with insufficient historical data to achieve accurate predictions.

(3) We incorporate a vector quantized design that, through in-depth analysis, demonstrates to facilitate the modality fusion.

(4) We release a Multi-modal Solar Power (MMSP) dataset, which integrates solar power generation records from numerous plants, satellite imagery, and numerical weather predictions. This rich dataset is collected from 88 diverse plants spread over an area of 157,100 square kilometers, covering a duration of 1.5 years.

(5) Our FusionSF is incorporated into our eForecaster platform [44], and provides short-term (day-ahead) solar power forecasting service for more than 300 solar plants with a capacity of over 15GW across three provinces in China.

2 RELATED WORK

There are two main approaches for solar power forecasting, including the physical approach and statistical approach [20]. The physical approach basically uses a deterministic model with mathematical equations to describe the input and output. As statistical models are becoming more popular, in this section we mainly focus on the latest developments in statistical approach, i.e., deep learning based methods.

2.1 Deep networks for time series and spatiotemporal forecasting

Solar power generation forecasting can be approached by pure TS forecasting or spatiotemporal forecasting, with the latter incorporating geographical information. Deep neural networks, especially Transformers, have garnered significant attention in the realm of TS forecasting [32]. Several efficient and high-performance Transformers have been developed for this purpose, such as Informer [40], Autoformer [34], FEDformer [42], FiLM [41], and PatchTST [23]. Concurrently, fully connected models (like Dlinear [37] and LightTS [38]) and convolution-based models (like TimesNet [33]) have also emerged as competitive alternatives in TS forecasting.

In the domain of spatiotemporal forecasting, SimVP model [9] employs convolution for spatial and temporal data processing. Alternatively, ConvLSTM [27] utilizes a recurrent network to capture temporal dependencies. Transformer-based models represent another family of methods that possess the potential to address both temporal and spatial dependencies. To handle a large amount of data, Earthformer [8] divides images into smaller patches, as also demonstrated by models such as Pangu [4], which has developed a large-scale global weather forecasting model and yielded superior performance compared to conventional NWP methods. Despite the remarkable success in TS and spatiotemporal forecasting achieved by deep neural networks, handling solar power forecasting still poses challenges, especially when there is a lack of multi-modal data sources to support the predictions. Furthermore, these models are typically designed to process inputs structured in a grid formation and, consequently, face difficulties when modeling an irregular network of stations, where each individual station may not correspond to a grid point.
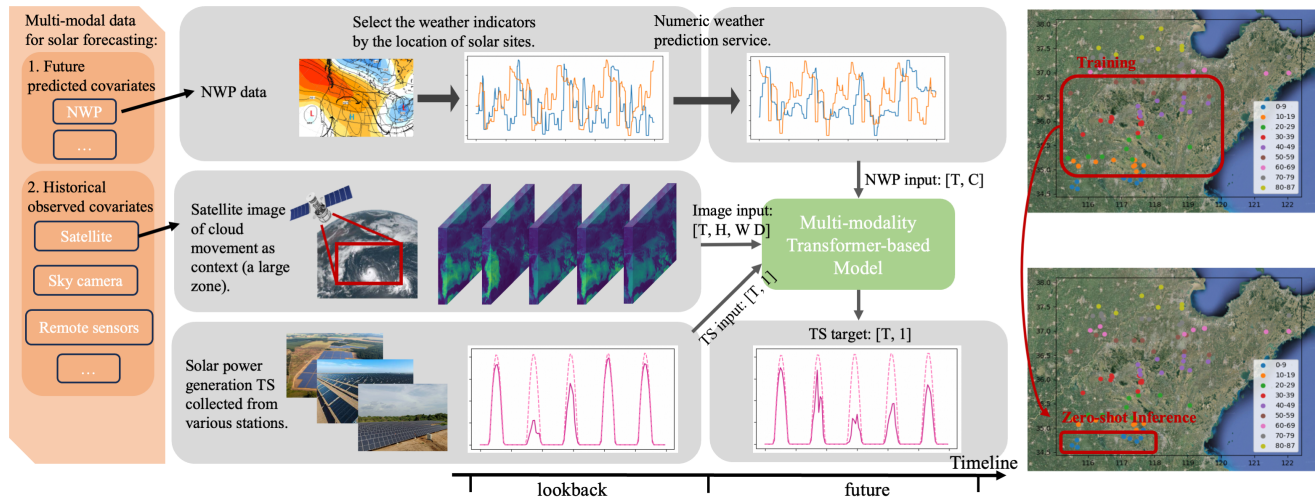
Figure 1: Left: An illustration of our proposed multi-modal framework. The three modalities include solar power historical data, satellite images, and NWP data. Right: Geographical locations of the 88 solar power plants and the zero-shot learning setting. The plants are grouped into sets of 10 and are represented in different colors.

2.2 Multi-modal solar forecasting

NWP, satellite, and sky camera are commonly utilized as key data sources to support solar forecasting [35]. Traditional techniques that rely on NWP data often employ regression-based approaches, as showcased in the Global Energy Forecasting Competition 2014 [14]. However, the effectiveness of these methods heavily depends on the accuracy of weather prediction. CrossViViT [5] integrates the satellite images as surrounding physical context. CorssViVit employs a cross-attention mechanism to effectively combine satellite images and solar power, and incorporates ROPE [28] to handle the coordinates information. Sky images are frequently leveraged to enhance ultra-shot-term solar forecasting [35]. Several studies [1, 19, 21] incorporate sky images as an auxiliary modality to improve solar power prediction. [21] utilizes a Vision Transformer to analyze sky images, coupled with an Informer [40] to process solar power TS. When considering short-term forecasting, the employment of satellite imagery and NWP is both essential and practical as heterogeneous modalities. These modalities offer expansive coverage and are more widely accessible as shown in Figure 1.

2.3 Zero-shot learning for time series

Despite the existence of robust zero-shot learners for natural language, achieving zero-shot learning for time series remains challenging. This difficulty primarily arises from the distribution disparities among TS originating from different domains. However, zero-shot learning within the same domain is possible. N-BEATS [24] acts as a meta-learning adaptation and demonstrates remarkable zero-shot performance in the domain of finance (M3 & M4 dataset). OneFitsAll [43] proves that deep Transformer structure (GPT2) excels as zero-shot learners. CrossViViT [5] proves the zero-shot ability for multi-modality solar forecasting.

3 METHODOLOGY

3.1 FusionSF overall architecture

The overall architecture of our proposed FusionSF is illustrated in Figure 2. We develop a multi-modal framework featuring three encoder branches to handle historical observed solar power inputs $\mathcal{X}_{ts} : [T_{in}, C_{ts}]$, historical observed context $\mathcal{X}_{ctx} : [T_{in}, C_{ctx}, H, W]$, and future predicted covariates $\mathcal{X}_{aux} : [T_{out}, C_{aux}]$, paired with a single decoder branch. Here T_{in} and T_{out} denote the temporal length of inputs and outputs. H and W denote the spatial dimensions of the contextual images; C_{ts}, C_{ctx}, and C_{aux} represent the number of features.

Given that historical observed contexts often include voluminous data sources like satellite imagery and the historical observed inputs are typically characterized by noisy TS data, we implement a vector-quantized (VQ) encoder branch for them. This approach offers two primary benefits: it not only reduces noise in the original data which enhances the robustness of the extracted features, but also facilitates the alignment of modalities with varying information densities. In contrast, future predicted covariates such as the weather prediction generally manifest as smoother signals with less noise. As a result, we can directly input the unprocessed signal into its designated encoder branch without VQ.

We also integrate a Cross Transformer based fusion module before the decoder, employing a key-value/query (KV/Q) cross-attention mechanism to fuse the three modalities. It is important to note that our example only utilizes satellite imagery, numerical weather prediction (NWP), and historical solar power data for illustrative purposes. However, the model's design is versatile and can readily incorporate additional data sources such as sky images or other covariates, using analogous fusion or concatenation modules.

3.2 Feature encoding

Rotary Positional encoding. To accurately model relative distances, we employ the Relative Positional Encoding (RoPE) [28]
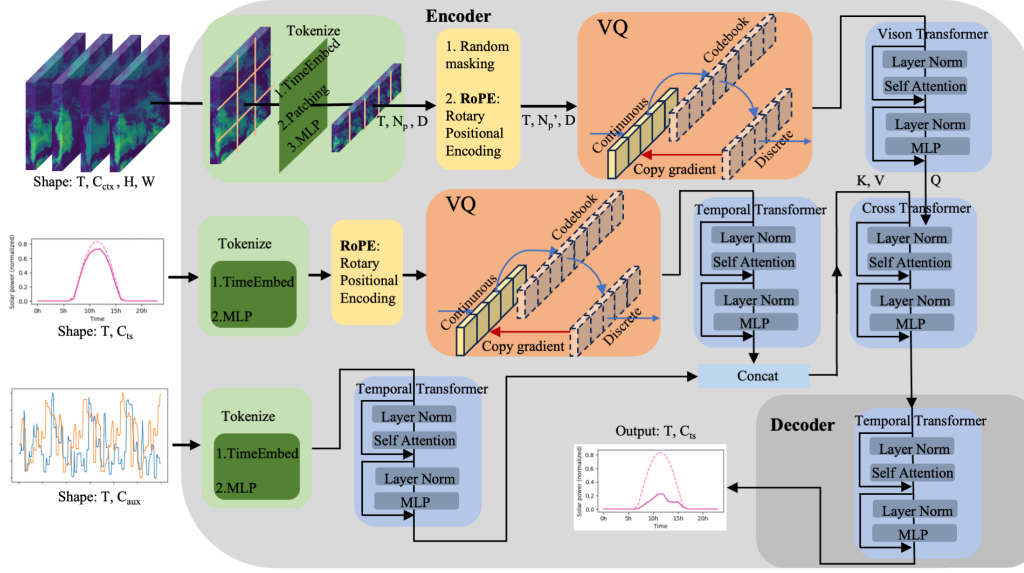
Figure 2: FusionSF architecture. The contextual images are tokenized, randomly masked, vector quantized, and processed with Vision Transformer. The vector quantized solar power TS and NWP covariates are processed with Temporal Transformer. The three modalities are fused with Cross Transformer. In the decoder, the mixed latent representation is processed with Temporal Transformer to make the final output.

which encodes the position with a rotation matrix and meanwhile incorporates the explicit relative position dependency in the self-attention formulation. More details of RoPE can be found in Appendix B.

Patching & Masking. Building upon the approach pioneered in the widely recognized Vision Transformer (ViT) [6], we implement patching to encapsulate small, localized regions of an image. The satellite image modality $\mathcal{X}_{ctx} : [B, T_{in}, C_{ctx}, H, W]$ is divided into N_p non-overlapping patches and subsequently projected into tokens with Multi-Layer Perceptron (MLP):

$$\mathcal{X}_{ctx}^{embed} = MLP(Patching(\mathcal{X}_{ctx})). \tag{1}$$

The shape of $\mathcal{X}_{ctx}^{embed}$ is $[(B*T), N_p, d]$, where B is the batch size, N_p is the number of patches, and d is the hidden dimension. Note that the temporal dimension T is permuted into batch dimension. The time series $\mathcal{X}_{ts}$ and auxiliary series $\mathcal{X}_{aux}$ are embedded without patching: $\mathcal{X}_{ts}^{embed} = MLP(\mathcal{X}_{ts})$, $\mathcal{X}_{aux}^{embed} = MLP(\mathcal{X}_{aux})$, where the shapes of $\mathcal{X}_{aux}^{embed}$ and $\mathcal{X}_{aux}^{embed}$ are $[(B*T), 1, d]$. Following [5], we mask a portion of the tokens in the context during the training phase. A masking ratio is randomly sampled from a uniform distribution, and the corresponding tokens and their positional embedding are masked. During inference, no masking is applied.

Vector quantization (VQ). Vector quantization (VQ) is proposed in VQVAE [30] to represent the image features in discrete space. By viewing quantization as a denoising procedure to improve model robustness, we adopt VQ to limit the patterns of encoding vectors and attain strong generalization. Formally, we initialize a codebook $e \in \mathbb{R}^{K*D}$, where K is the size of the codebook, and D is the dimension of the encoding vector. The encoding vector is denoted as $z_e(x)$, which is then replaced by the closest code in e:

$$z_q(x) = e_k, \text{where } k = \operatorname{argmin}_j ||z_e(x) - e_j||_2. \tag{2}$$

The replacement of vectors interrupts gradient propagation. To address this, VQ employs the straight-through estimator [2] to approximate the gradient by simply copying gradients from the quantized outputs $z_q(x)$ to the encoding vectors $z_e(x)$.

To balance the trade-off between detail preserving and noise removal, we use residual VQ [36] to recursively quantize the encoding vectors. Following VQVAE [30], we introduce a commitment loss to ensure that the encoding vectors are close to the codebook: $L_{cmt} = ||z_e(x) - sg[e]||_2^2$, where sg refers to the stop-gradient operator, which functions as an identity during the forward process but has zero partial derivatives during backpropagation. The codebook is learned through exponential moving averages (EMA) as proposed in [30]. The residual VQ layer (RVQ) is applied on $\mathcal{X}_{ts}^{embed}$ and $\mathcal{X}_{ctx}^{embed}$: $\mathcal{X}_{ctx}^{quantized} = RVQ(\mathcal{X}_{ctx}^{embed})$, $\mathcal{X}_{ts}^{quantized} = RVQ(\mathcal{X}_{ts}^{embed})$.

Transformer-based Encoder. In the encoder stage, the quantized context $\mathcal{X}_{ctx}^{quantized}$ is first processed using the Vision Transformer (VIT) architecture. TVT consists of several components, including layer normalization, multi-head self-attention, MLP, and residual connection:

$$\mathcal{X}_{ctx}^{latent} = VisionTransformer(\mathcal{X}_{ctx}^{quantized}). \tag{3}$$

Additionally, the quantized time series data: $\mathcal{X}_{ts}^{quantized}$ and $\mathcal{X}_{aux}^{embed}$ are processed with Temporal Transformer:

$$\mathcal{X}_{ts}^{latent} = TemporalTransformer(\mathcal{X}_{ts}^{quantized}), \tag{4}$$

$$\mathcal{X}_{aux}^{latent} = TemporalTransformer(\mathcal{X}_{aux}^{embed}). \tag{5}$$

Note that within our proposed framework, any alternative vision-based Transformer or temporal Transformer model can be integrated as a plug-in component to enhance performance.

3.3 Modality mixing

After encoding, it becomes necessary to mix the three modalities. $\mathcal{X}_{\text{ts}}^{\text{latent}}$ and $\mathcal{X}_{\text{aux}}^{\text{latent}}$ are concatenated on hidden dimension, which allows the data aligned according to the hour of the day: $\mathcal{X}_{\text{cat}}^{\text{latent}} = $ concat($\mathcal{X}_{\text{ts}}^{\text{latent}}, \mathcal{X}_{\text{aux}}^{\text{latent}}$).

Furthermore, we employ the Cross Transformer mechanism to integrate the image and (TS) modalities. Within this framework, the image modality is designated as the query (Q), while the TS modality serves as the key (K) and value (V):

$$\mathcal{X}_{\text{mixed}}^{\text{latent}} = \text{CrossAttention}(\mathcal{X}_{\text{ctx}}^{\text{latent}}, \mathcal{X}_{\text{cat}}^{\text{latent}}, \mathcal{X}_{\text{cat}}^{\text{latent}}). \qquad (6)$$

$\mathcal{X}_{\text{mixed}}^{\text{latent}}$ is the final output of the encoder. In the decoder stage, $\mathcal{X}_{\text{mixed}}^{\text{latent}}$ is first processed with an MLP layer and subsequently a Temporal Transformer to output the final prediction $\hat{\mathcal{Y}} : [B, T_{\text{out}}, C_{\text{ts}}]$.

4 BENCHMARK DATASET

This section presents an overview of our proposed Multi-modal Solar Power (MMSP) dataset, which has been made publicly available. For more details, please refer to Appendix A. The statistics of the dataset are summarized in Table 1.

Table 1: Dataset statistics.

Dataset	Data type	Length	Dim	Freq
MMSP(S)	Satellite	25540≈2 years	64×64×1	1h
	NWP	12864 ≈1.5 years	79grid × 15	1h
	solar power ts	12840 ≈1.5 years	10plants × 1	1h
MMSP(L)	Satellite	25540≈2 year	64×64×4	1h
	NWP	12864 ≈1.5 years	79grid × 15	1h
	solar power ts	12840 ≈1.5 years	88plants × 1	1h

Historical time series modality. MMSP dataset encompasses a comprehensive TS dataset of solar power generation, obtained from a network of 88 geographically dispersed solar power plants spanning across a province in China measuring 157,100 square kilometers. The dataset has been downsampled to a resolution of 60 minutes and covers a temporal range from Jan 2021 to June 2022. To facilitate parameter tuning and benchmarking, we select the initial 10 plants to create a smaller dataset MMSP(S).

Historical satellite image modality. The Himawari-8/9 satellites, operated by the Japan Meteorological Agency (JMA), provide invaluable satellite imagery data that has revolutionized weather monitoring and analysis in the Asia-Pacific region.

Future numerical weather prediction modality. The European Centre for Medium-Range Weather Forecasts (ECMWF) offers valuable NWP data that plays a pivotal role in advancing weather forecasting and related research.

5 EXPERIMENT

5.1 Benchmark

Baselines. We perform a thorough evaluation by comparing FusionSF with various SOTA time series baselines, namely Informer [40],

Table 2: Comparative analysis of model performance on MMSP(S) dataset across "All", "Easy", and "Hard" scenarios. We use MAE(↓) and RMSE(↓) as metrics. The best results are highlighted in bold, and the second best results are highlighted with underline.

Models	All (25210)		Easy (18014)		Hard (7196)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Persistence	0.06500	0.13909	<u>0.04763</u>	0.10279	0.10838	0.20319
Mean	0.07632	0.12849	0.07674	0.12614	<u>0.07528</u>	<u>0.13417</u>
Clear sky	0.07347	0.15682	0.05589	0.12196	0.11748	0.22119
Informer [40]	0.07973	0.13086	0.07952	0.12867	0.08025	0.13613
Autoformer [34]	0.07830	0.11702	0.07015	0.09876	0.10285	0.15505
Crossformer [39]	0.06599	0.11259	0.06201	0.10173	0.08440	0.14645
PatchTST [23]	0.06575	0.11755	0.06056	0.10192	0.08320	0.14783
FiLM [41]	0.06995	0.12529	0.05783	0.09468	0.10474	0.18154
Dlinear [37]	0.07609	0.12310	0.06364	0.09762	0.10682	0.17035
LightTS [38]	0.06474	<u>0.11048</u>	0.05724	<u>0.09347</u>	0.08324	0.14413
CrossViVit [5]	<u>0.05789</u>	0.11818	0.04891	0.09924	0.08007	0.15535
FusionSF	**0.04020**	**0.08881**	**0.03891**	**0.08359**	**0.04980**	**0.10690**

Autoformer [34], Crossformer [39], PatchTST [23], FiLM [41], Dlinear [37], and LightTS [38], which are specifically designed for pure TS forecasting tasks. CrossViVit [5] leverages satellite imagery as contextual information to enhance solar forecasting outcomes. Additionally, we introduce some naive statistic methods specifically tailored for solar power forecasting, which turn out to be practically useful and widely applied in industry [26]. *Persistence* [26] uses the past day's true values as the prediction for the current day. *Mean* uses the average power of all historical series in the training set as the prediction. *Clear Sky* computes the theoretical Global Horizontal Irradiance (GHI) at a specific location by its temporal and geographic information, which implies the total irradiance reaches the ground in the absence of clouds [13, 15], and then maps it to solar power. The dataset is divided into training, validation, and test sets with a ratio of [0.6: 0.2: 0.2].

We recognize that exclusively using our benchmark could be perceived as a limitation. Nevertheless, the modality fusion strategy presented here is central to our work and merits further investigation. Lacking a suitable existing benchmark to illustrate our approach, we have released our dataset to the public and concentrated our analysis on it. Testing on alternative two-modality datasets would not sufficiently highlight our principal contribution nor ensure real-world applicability.

Full benchmark. As shown in Table 2, it is observed that the performance of naive baselines is comparable to that of TS baselines, as the weather system is chaotic and the input series from the past 24 hours provides limited guidance. Among the TS forecasting algorithms, LightTS performs the best. However, our proposed trimodality framework demonstrates superior performance to LightTS, with an improvement of 37.9% and 19.6% in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Moreover, our proposed framework outperforms the CrossViVit model, which utilizes two modalities, by 30.6% and 24.9% in terms of MAE and RMSE, respectively.

'Easy' vs 'Hard' scenario. Since GHI exhibits similarity between consecutive days, *Persistence* shows good performance over the
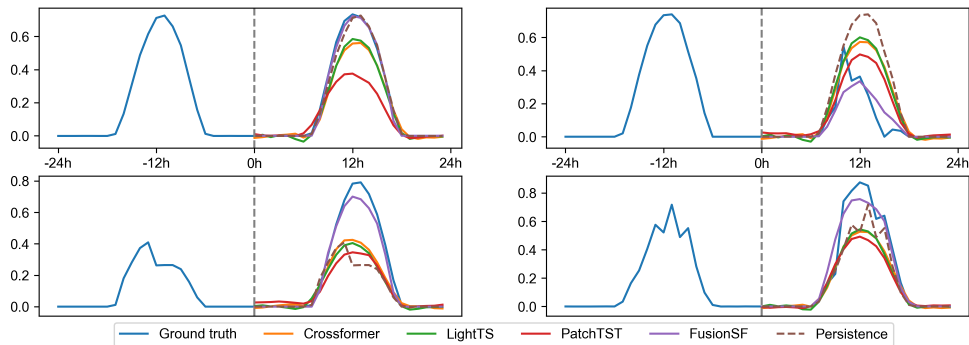
Figure 3: Prediction visualization from FusionSF and other baselines. The first row shows two 'Hard' cases and the second row shows two 'Easy' cases.

full test dataset. We categorize samples into 'Easy' and 'Hard' subsets based on the prediction difficulty metric outlined in [5]. This metric assesses the challenge posed by a sample in terms of its susceptibility to a persistence model. If straightforward "copying and pasting" leads to a high accuracy, the sample is considered as an 'Easy' case. Conversely, if this approach fails to yield accurate predictions, we designate the sample as a 'Hard' one. Specifically, we calculate the ratio of the area under the power curve for the two days, i.e., $r = |\log \frac{y}{y_{prev}}|$, where y represents the power over 24 hours and y_{prev} represents the power over the previous 24 hours. Accordingly, a sample is categorized as 'Easy' if $r < |\log(\frac{2}{3})|$ and 'Hard' otherwise. From Table 2, we can observe that FusionSF exhibits more significant improvement in 'Hard' scenarios than the 'Easy' ones. This outcome underscores the importance of leveraging NWP data (compared with CrossViVit [5]) in handling scenarios with fluctuating weather conditions. To investigate why FusionSF performs well, we plot the prediction of FusionSF, several baselines, and the ground truth in Figure 3, for both 'Hard' and 'Easy' scenarios. It can be observed that FusionSF outperforms other methods significantly in the peak hours when general accurate prediction is most challenging. We present a complete case study with details in Appendix D.

5.2 Zero-shot performance on stations outside the training distribution

Table 3: Comparison of zero-shot performance for different data modality using MAE(↓) as the evaluation metric. (-) indicates improvements on MAE, and (+) indicates degraded performance.

	non-zero-shot	zero-shot		
Plants for training	#0-#9	#10-#19	#10-#29	#10-#39
Satellite+TS	0.05789	+34%	+14%	+32%
Satellite+NWP+TS	0.04020	+4.0%	-0.5%	-1.2%

To demonstrate the zero-shot learning capability of our model, we evaluate its performance on stations that lie outside the training distribution as shown in Figure 4. In Scenario No. 2, we utilized the
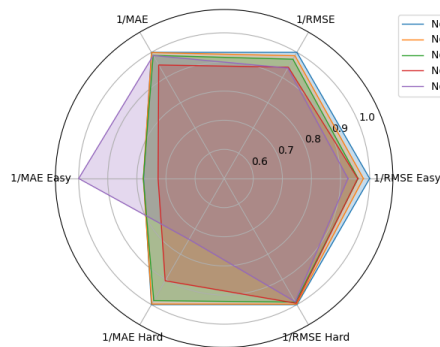

Figure 4: Radar plots for analyzing the model performance for zero-shot learning. The test set includes data from solar plants #0 to #9 and the training set varies. The metrics are rescaled for visual clarity. A larger radar plot indicates better performance.

data from plants #10 to #19 for training purposes, while reserving the data from plants #0 to #9 for testing. Our investigation revealed a performance degradation of 4.0% in the zero-shot learning setting when compared to the scenario where training and testing are executed on the same plants (Scenario No. 1).

To further investigate the impact of training set size, we expanded the training set to include data from 50 plants (plants #10 to #59). Notably, the performance of our model demonstrated an improvement, outperforming the non-zero-shot learning approach by 1.3%. These findings underscore the importance of a larger training set and its positive influence on overall performance.

Note that the non-zero-shot setting (Scenario No. 1) performs better in the 'Easy' scenario while other zero-shot settings (Scenario No. 2 to No. 5) perform better in 'Hard'. This distinction arises from the fact that the former settings can access the patterns present in the target plants and consequently overfit them. While the zero-shot models, especially those trained with more plants, are capable of acquiring a deeper understanding of the relationship between weather conditions and solar power generation.

As shown in Table 3, the introduction of NWP as a third modality demonstrates more robust performance in zero-shot learning compared with the two-modality version. This finding emphasizes the

importance of modality fusion in achieving solidity and reliability in the zero-shot learning context.

5.3 Ablation study

Ablation of data. By incorporating satellite images as a secondary modality (No. 2), we observed approximately a 10.6% improvement in MAE compared to the best TS baseline (No. 1), which aligns with the findings reported by [5]. Additionally, with the inclusion of NWP data in conjunction with the TS data (No. 3), we observed a further 30.6% improvement in MAE compared to the setting No. 2. This indicates that the NWP data offers a more direct and precise prediction of weather conditions within the desired time horizon, while the modeling of the relationship between satellite context and the target series presents challenges.

The trimodality setting (No. 6) exhibits superior performance compared to the NWP+TS (No. 3) setting, with an improvement of 10.7%. This observation suggests that the NWP and context modalities complement each other, leading to enhanced predictive capabilities.

In our research, various resolutions of satellite contexts are examined. We initially employ a resolution of 50km and 64x64 pixels (No. 6). Finer resolutions of 25km (No. 4) and 10km (No. 5) are also evaluated. Notably, the degradation of performance is observed when utilizing the 10km resolution, primarily due to the limited spatial coverage of the context area at this resolution.

Table 4: Ablation study on MMSP(S) dataset for analyzing the impact of data modalities and model structures.

	Methods	MAE	RMSE
Ablation of data	No. 1 TS (LightTS)	0.06474	0.11048
	No. 2 TS+Satellite	0.05789	0.11818
	No. 3 TS+NWP	0.04503	0.09890
	No. 4 TS+NWP+Satellite(25km)	0.04144	0.09046
	No. 5 TS+NWP+Satellite(10km)	0.04267	0.08862
Ours: FusionSF	No.6 TS+NWP+Satellite w/VQ on Satellite&TS	**0.04020**	**0.08881**
Ablation of module	No.7 w/o VQ	0.04124	0.09222
	No.8 w/ VQ only on Satellite	0.04289	0.08875
	No.9 w/ VQ only on TS	0.04266	0.09213
	No.10 w/ VQ on Satellite&TS&NWP	0.04152	0.08985
	No.11 w/o Random Masking	0.04369	0.09139

Ablation of module. We conduct an ablation study on various VQ modules within our proposed framework. Employing VQ on both TS and satellite (not on NWP) leads to the best performance. Additionally, it is noteworthy that the application of random masking to the satellite modality results in a performance enhancement of approximately 8.0%.

5.4 How does VQ adjust distributions?

To elucidate the mechanisms by which vector quantization (VQ) layers contribute to enhanced model performance, we visualize the latent values with and without the VQ layers. In Figure 5 Upper, we observe that the VQ layer functions as a normalizing agent, condensing the distribution of latent values for satellite images. We employ the Kullback-Leibler (KL) divergence as a metric to assess the similarity between the latent distributions of images and TS

data. In the absence of VQ, the KL divergence is 0.264. However, the implementation of VQ results in a significant reduction of the KL divergence to 0.080, thereby indicating a substantial alignment and enhancement of distributional proximity between the latent representations of different modalities.

In the t-SNE visualizations (Figure 5 Lower), the application of VQ delineates each cluster more distinctly and makes different tokens evenly distributed across these clusters.
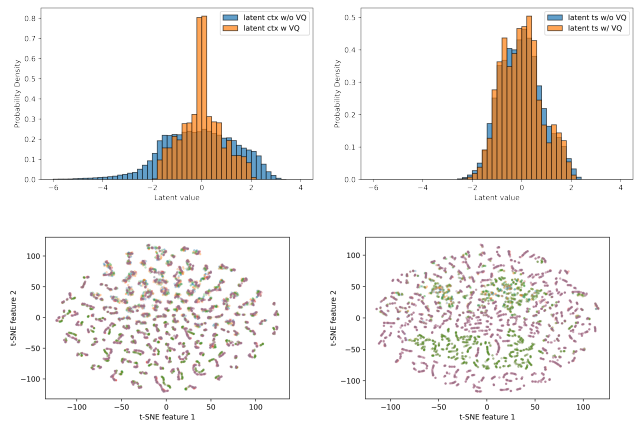

Figure 5: Comparison of latent value distributions employing VQ for satellite images and TS data. The upper panels display the distributions with and without VQ for satellite images (Upper Left) and TS (Upper Right). Lower panels illustrate t-SNE visualizations of latent values corresponding to satellite images with (Lower Left) and without (Lower Right) VQ. Distinct colors denote disparate tokens.

5.5 Where is the limit?

Even with access to the most precise weather forecasts or actual meteorological conditions, our research indicates that multi-modal approaches remain essential.

In the experiments shown in Table 5, we utilized SolarTCN, a lightweight CNN-based backbone model that has been extensively deployed in our real-world solar forecasting projects. This model mainly relies on NWP to build a regression model for solar power forecasting, which is a widely adopted approach [14]. Additionally, we introduce the 5th generation of ECMWF Reanalysis data (ERA5) [12], which combines model data with observation data using data assimilation and is recognized as the best estimation of the state of the atmosphere [11, 17]. It provides a dataset of several weather fields on 0.25° latitude-longitude resolution and 1 hour time step. Recent AI-based weather forecasting models like FourCastNet [25], Pangu-Weather [4], and GraphCast [16] all use ERA5 as the ground truth.

In scenario No. 3, by training and testing SolarTCN with ERA5 as input, the performance is improved (0.0422 on MAE) compared to using NWP (0.0444 on MAE). This MAE value of 0.0422 can be considered as the theoretical upper bound that SolarTCN can reach by continuously enhancing the accuracy of the weather prediction. Note that the coarseness of ERA5 data, which has a resolution of roughly 0.25°, and the lack of actual observed meteorological data for calibration, prevents ERA5 from accurately reflecting real-world weather conditions at solar power stations.

Table 5: Analysis of forecasting error due to NWP inaccuracy. An assessment using ERA5 data as ideal weather reports and Satellite(Real) observations for actual cloud conditions, with SolarTCN serving as the backbone.

	Scenario No.	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6
Method	FusionSF	✓	-	-	-	-	-
	SolarTCN	-	✓	✓	✓	✓	✓
Data	TS: -24h→0h	✓	-	-	-	-	-
	NWP: 0h→24h	✓	✓	-	-	✓	-
	ERA5: 0h→24h	-	-	✓	-	-	✓
	Satellite: -24h→0h	✓	-	-	-	-	-
	Satellite(Real): 0h→24h	-	-	-	✓	✓	✓
Metric	MAE	0.0402	0.0444	0.0422	0.0502	0.0406	0.0399

Similarly, in scenario No. 4, we utilize the satellite images in the future horizon (0h→24h) to represent real cloud conditions (Satellite (Real)). SolarTCN achieves 0.0502 on MAE, indicating sub-par performance. This observation highlights the fact that relying solely on a single modality (No. 2, No. 3, and No. 4), even when using ground truth weather conditions (No. 3 and No. 4), does not lead to satisfactory performance.

While an enhanced weather prediction may improve accuracy, it is not sufficient to achieve perfect predictions. Instead, our experimental results indicate that introducing a new modality might be more promising. In our approach (No. 1), by combining two exogenous modalities including satellite and NWP, our solution can utilize data sources that complement each other and bring more significant improvement.

In future research, by introducing more sources of information pertaining to weather conditions as input, for example, sky images and remote sensing data, we expect to achieve even greater improvements in the accuracy and performance of the solar power prediction model.

5.6 Real-world deployment

As of Jan. 2024, FusionSF has been deployed to predict day-ahead solar power for more than 300 solar plants across three provinces in China. These plants have a total capacity of over 15 GW and generate more than 1.5×10^{10} kWh per year. Our system outperforms the previous forecasting systems (SolarTCN) with a consistent improvement of 1.5% in accuracy. According to [45], minor forecasting errors can lead to a 30-fold increase in imbalance fees in Scandinavian energy markets. While the dynamics of China's electricity market differ, enhancing forecast accuracy is expected to yield significant cost savings, especially considering the diversities among these deployed plants.

In the deployment phase, FusionSF is incorporated into our eForecaster platform [44]. This platform stands as a versatile, modular, and cohesive artificial intelligence framework designed to facilitate diverse applications for electrical forecasting, such as electric load forecasting, wind power forecasting, and solar power forecasting. As illustrated in Figure 6, with eForecaster, developers can implement an end-to-end forecasting pipeline composed of Pre-processing, Feature engineering, Modelling, and Post-processing stages. In terms of data, a database that contains historic solar power, ECMWF high-resolution 10-day forecast (HRES) NWP data, and Himawari satellite data are maintained in the backend, where

all these data are retrieved from their source, and pushed into the database in real-time. When making day-ahead forecasting, the trimodal data, along with other extra data like temporal or season information, constitute the raw input. Specifically, for solar power forecasting, we establish a Pre-processing module where outliers are removed through our robust anomaly detection methods [31], and then imputations are made for missing values. The Feature Engineering module extracts temporal and coordinate features, while the Modeling module allows for the selection and application of specific forecasting algorithms. Finally, users can ensemble and adjust the results in the Post-processing module. Since station capacity changes, power curtailments, and extreme events (e.g., sandstorms, snowstorms) all greatly influence the actual solar power penetration into the power grid, this procedure is crucial in reality. For example, the user can adjust the predicted power directly when equipment maintenance happens.

Notably, FusionSF is trained offline and necessitates only the inference process in online environments. This attribute enables the model to operate with minimal computational resources, obviating the need for GPU support. Benefiting from the zero-shot learning capacity, our algorithm remains competitive in prediction accuracy even though approximately 30% of the solar stations utilizing it have insufficient historical data. Moreover, when new stations are set up, the cold start challenge is eased thanks to FusionSF's advantage in zero-shot learning.
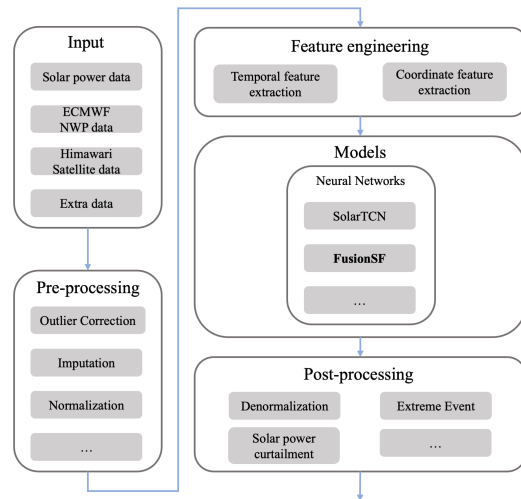

Figure 6: Solar forecasting platform.

6 CONCLUSION

In summary, we aim to propel advancements in solar power forecasting by utilizing a refined VQ (Vector Quantized) multi-modality fusion framework and incorporating multi-modal data sources. This approach is designed to enhance both accuracy and zero-shot learning capabilities for practical and real-world deployment. Through our widespread applications across numerous solar power plants, we demonstrate that this interdisciplinary approach harbors considerable potential for optimizing renewable energy utilization and promoting sustainable energy practices.

REFERENCES

[1] Anas Al-lahham, Obaidah Theeb, Khaled Elalem, Tariq A. Alshawi, and Saleh A. Alshebeili. 2020. Sky Imager-Based Forecast of Solar Irradiance Using Machine Learning. *Electronics* 9, 10 (2020). https://doi.org/10.3390/electronics9101700

[2] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *CoRR* abs/1308.3432 (2013). arXiv:1308.3432 http://arxiv.org/abs/1308.3432

[3] Kotaro BESSHO, Kenji DATE, Masahiro HAYASHI, Akio IKEDA, Takahito IMAI, Hidekazu INOUE, Yukihiro KUMAGAI, Takuya MIYAKAWA, Hidehiko MURATA, Tomoo OHNO, Arata OKUYAMA, Ryo OYAMA, Yukio SASAKI, Yoshio SHIMAZU, Kazuki SHIMOJI, Yasuhiko SUMIDA, Masuo SUZUKI, Hidetaka TANIGUCHI, Hiroaki TSUCHIYAMA, Daisaku UESAWA, Hironobu YOKOTA, and Ryo YOSHIDA. 2016. An Introduction to Himawari-8/9 mdash; Japan rsquo;s New-Generation Geostationary Meteorological Satellites. *Journal of the Meteorological Society of Japan. Ser. II* 94, 2 (2016), 151–183. https://doi.org/10.2151/jmsj.2016-009

[4] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nat.* 619, 7970 (2023), 533–538. https://doi.org/10.1038/S41586-023-06185-3

[5] Oussama Boussif, Ghait Boukachab, Dan Assouline, Stefano Massaroli, Tianle Yuan, Loubna Benabbou, and Yoshua Bengio. 2023. What if We Enrich day-ahead Solar Irradiance Time Series Forecasting with Spatio-Temporal Context? *CoRR* abs/2306.01112 (2023). https://doi.org/10.48550/ARXIV.2306.01112 arXiv:2306.01112

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[7] Michael Emmanuel, Kate Doubleday, Burcin Cakir, Marija Marković, and Bri-Mathias Hodge. 2020. A review of power system planning and operational models for flexibility assessment in high solar energy penetration scenarios. *Solar Energy* 210 (2020), 169–180. https://doi.org/10.1016/j.solener.2020.07.017 Special Issue on Grid Integration.

[8] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. 2022. Earthformer: Exploring Space-Time Transformers for Earth System Forecasting. In *NeurIPS*.

[9] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. 2022. SimVP: Simpler yet Better Video Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 3160–3170. https://doi.org/10.1109/CVPR52688.2022.00317

[10] D González-Zeas, B Erazo, P Lloret, B De Bièvre, S Steinschneider, and Olivier Dangles. 2019. Linking global climate change. *Science of the Total Environment* 650 (2019), 2577–2586.

[11] Yanyi He, Kaicun Wang, and Fei Feng. 2021. Improvement of ERA5 over ERA-Interim in simulating surface incident solar radiation throughout China. *Journal of Climate* 34, 10 (2021), 3853–3867.

[12] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* 146, 730 (2020), 1999–2049.

[13] William F Holmgren, Clifford W Hansen, and Mark A Mikofski. 2018. pvlib python: A python package for modeling solar energy systems. *Journal of Open Source Software* 3, 29 (2018), 884.

[14] Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. 2016. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting* 32, 3 (2016), 896–913. https://doi.org/10.1016/j.ijforecast.2016.02.001

[15] Pierre Ineichen. 2016. Validation of models that estimate the clear sky global and beam solar irradiance. *Solar Energy* 132 (2016), 332–344.

[16] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. 2023. Learning skillful medium-range global weather forecasting. *Science* (2023), eadi2336.

[17] David A Lavers, Adrian Simmons, Freja Vamborg, and Mark J Rodwell. 2022. An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society* 148, 748 (2022), 3152–3165.

[18] Zhuo Li, Kejie Wang, Chenchen Li, Miao Zhao, and Jiannong Cao. 2019. Multimodal Deep Learning for Solar Irradiance Prediction. In *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. 784–792. https://doi.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00144

[19] Zhuo Li, Kejie Wang, Chenchen Li, Miao Zhao, and Jiannong Cao. 2019. Multimodal Deep Learning for Solar Irradiance Prediction. In *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom)*

[19 cont.] *and IEEE Smart Data (SmartData)*. 784–792. https://doi.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00144

[20] Tariq Limouni, Reda Yaagoubi, Khalid Bouziane, Khalid Guissi, and El Houssain Baali. 2023. Accurate one step and multistep forecasting of very short-term PV power using LSTM-TCN model. *Renewable Energy* 205 (2023), 1010–1024. https://doi.org/10.1016/j.renene.2023.01.118

[21] Jingxuan Liu, Haixiang Zang, Lilin Cheng, Tao Ding, Zhinong Wei, and Guoqiang Sun. 2023. A Transformer-based multimodal-learning framework using sky images for ultra-short-term solar irradiance forecasting. *Applied Energy* 342 (2023), 121160. https://doi.org/10.1016/j.apenergy.2023.121160

[22] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[23] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=Jbdc0vTOcol

[24] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2021. Meta-Learning Framework with Applications to Zero-Shot Time-Series Forecasting. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 9242–9250. https://doi.org/10.1609/AAAI.V35I10.17115

[25] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. 2022. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *CoRR* abs/2202.11214 (2022). arXiv:2202.11214 https://arxiv.org/abs/2202.11214

[26] Elissaios Sarmas, Nikos Dimitropoulos, Vangelis Marinakis, Zoi Mylona, and Haris Doukas. 2022. Transfer learning strategies for solar power forecasting under data scarcity. *Nature Scientific Reports* 12 (2022), 14643. https://doi.org/10.1038/s41598-022-18516-x

[27] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf

[28] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. RoFormer: Enhanced Transformer with Rotary Position Embedding. *CoRR* abs/2104.09864 (2021). arXiv:2104.09864 https://arxiv.org/abs/2104.09864

[29] Esra Suel, Samir Bhatt, Michael Brauer, Seth Flaxman, and Majid Ezzati. 2021. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sensing of Environment* 257 (2021), 112339. https://doi.org/10.1016/j.rse.2021.112339

[30] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 6306–6315. https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

[31] Qingsong Wen, Linxiao Yang, Tian Zhou, and Liang Sun. 2022. Robust time series analysis and applications: An industrial perspective. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4836–4837.

[32] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2023. Transformers in Time Series: A Survey. 6778–6786. https://doi.org/10.24963/ijcai.2023/759

[33] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=ju_Uqw384Oq

[34] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. 101–112.

[35] Dazhi Yang, Wenting Wang, Christian A. Gueymard, Tao Hong, Jan Kleissl, Jing Huang, Marc J. Perez, Richard Perez, Jamie M. Bright, Xiang'ao Xia, Dennis van der Meer, and Ian Marius Peters. 2022. A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. *Renewable and Sustainable Energy Reviews* 161 (2022), 112348. https://doi.org/10.1016/j.rser.2022.112348

[36] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. SoundStream: An End-to-End Neural Audio Codec. *IEEE ACM Trans. Audio Speech Lang. Process.* 30 (2022), 495–507. https://doi.org/10.1109/TASLP.2021.3129994

[37] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers Effective for Time Series Forecasting?. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 11121–11128. https://doi.org/10.1609/AAAI.V37I9.26317

[38] Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. 2022. Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures. *CoRR* abs/2207.01186 (2022). https://doi.org/10.48550/ARXIV.2207.01186 arXiv:2207.01186

[39] Yunhao Zhang and Junchi Yan. 2023. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=vSVLM2j9eie

[40] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, Vol. 35. 11106–11115.

[41] Tian Zhou, Ziqing Ma, Xue Wang, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, and Rong Jin. 2022. FiLM: Frequency improved Legendre Memory Model for Long-term Time Series Forecasting. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/524ef58c2bd075775861234266e5e020-Abstract-Conference.html

[42] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *39th International Conference on Machine Learning (ICML).*

[43] Tian Zhou, PeiSong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One Fits All:Power General Time Series Analysis by Pretrained LM. arXiv:2302.11939 [cs.LG]

[44] Zhaoyang Zhu, Weiqi Chen, Rui Xia, Tian Zhou, Peisong Niu, Bingqing Peng, Wenwei Wang, Hengbo Liu, Ziqing Ma, Xinyue Gu, Jin Wang, Qiming Chen, Linxiao Yang, Qingsong Wen, and Liang Sun. 2023. Energy forecasting with robust, flexible, and explainable machine learning algorithms. *AI Magazine* 44, 4 (dec 2023), 377–393. https://doi.org/10.1002/aaai.12130

[45] Øyvind Sommer Klyve, Magnus Moe Nygård, Heine Nygard Riise, Jonathan Fagerström, and Erik Stensrud Marstein. 2023. The value of forecasts for PV power plants operating in the past, present and future Scandinavian energy markets. *Solar Energy* 255 (2023), 208–221. https://doi.org/10.1016/j.solener.2023.03.044

A DETAILS OF THE DATASET

This section provides a comprehensive overview of our proposed Multi-modal Solar Power (MMSP) dataset. Table 1 details critical aspects of the dataset, including the type of data, the temporal span of the data collection, the dimensions of the data, and the resolution of individual modality. We provide a comprehensive version denoted as MMSP(L) and a smaller version as MMSP(S). We have made this dataset publicly accessible to facilitate knowledge sharing and collaborative research. To ensure confidentiality, we employ anonymization techniques on geographical information (latitude and longitude) of the power plants, NWP data, and satellite data. We also normalize solar power measurements based on capacity.
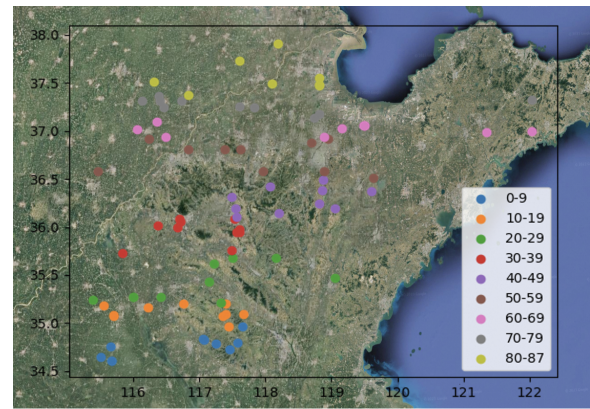
A.1 Solar power time series

Figure 7: Geographical locations of the 88 solar power plants. The plants are grouped by sets of 10 and represented in different colors.
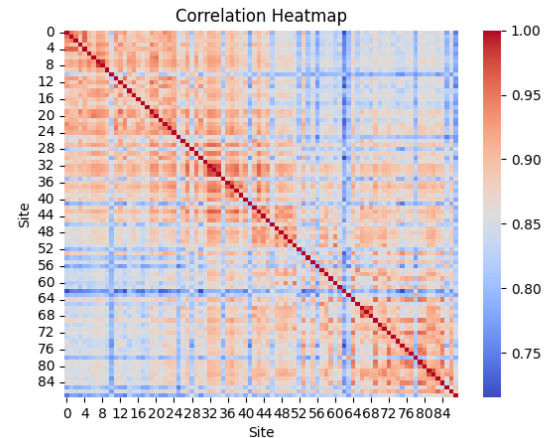

Figure 8: Correlation heatmap of 88 solar power plants.

The MMSP dataset is a comprehensive multi-modal dataset comprising time series of solar power generation. This dataset has been collected from a network of 88 solar power plants located across a

province in China, covering a vast area of 157,100 square kilometers. The geographical locations of these solar plants are illustrated in Figure 7. The correlation among the 88 plants is depicted in Figure 8. It is observed that the plants located in close proximity exhibit a higher correlation, indicating a stronger relationship between solar power generation and the weather conditions specific to a particular location. The dataset covers a temporal range from January 2021 to June 2022, allowing for a comprehensive analysis of solar power trends.

The original resolution of the dataset is 10 minutes, but for convenience, it has been downsampled to a resolution of 60 minutes. The time series data has undergone a removal process to eliminate abnormal samples. However, it still includes instances of power restriction conditions, where the power generation of the plants is limited to a very low level despite favorable solar conditions. It is important to note that such conditions are infrequent within the dataset, and although they may be considered as noise, we have chosen to retain these samples for further analysis. To facilitate parameter tuning, we have chosen the initial 10 plants to construct a smaller dataset called MMSP(S).

Note that the dataset spans a duration of only 1.5 years, which may be considered relatively short for training a large-scale weather system model. However, this is a common scenario for many solar plants, as they are newly deployed and have limited historical data available. To address the challenge of insufficient data, we propose FusionSF as a joint model for all plants instead of creating individual models for each plant. By doing so, we aim to capture and establish the relationship between complex weather patterns and solar power generation.

A.2 Satellite image

The geostationary satellites Himawari-8 and Himawari-9 are equipped to capture high-resolution imagery across East Asia, Oceania, and select regions of the Pacific Ocean. These satellites provide an extensive data resource, proving invaluable for advancements in meteorological and climate research. The Himawari-8/9 data encompasses various satellite products, including visible, infrared, and water vapor imagery, enabling scientists to investigate atmospheric phenomena, monitor severe weather events, and study cloud dynamics. The availability of Himawari-8/9 satellite image data has greatly contributed to advancements in weather forecasting, climate research, and the understanding of regional weather patterns.

The Advanced Himawari Imagers (AHIs) on the Himawari8/9 satellite capture complete views of the Earth's surface in 16 different observation bands. These bands consist of three for visible light, three for near-infrared, and ten for infrared wavelengths. These observations are taken every 10 minutes and provide a spatial resolution that varies between 0.5 to 2 kilometers [3].

Himawari-8/9 are geostationary satellites that jointly offer uninterrupted coverage of the target region. Figure 9 showcases a series of images depicting the transition from morning to night within a single day. From these visuals, it is evident that the imagery is dependent on sunlight reflection, resulting in clouds being undetectable during the night. It is only during the daytime that satellite imagery can capture visible cloud formations.

We conducted a preliminary experiment to determine the optimal number of bands required for solar power forecasting. After careful consideration, we select the three visible bands (blue: Albedo_01, 0.47 μm, green: Albedo_02, 0.51 μm, red: Albedo_03, 0.64 μm) and one near-infrared band (Albedo_04, 0.86 μm) as the context satellite image. Our analysis demonstrated that utilizing these four bands is sufficient for accurate solar power forecasting. A sample of selected 4 channels is demonstrated in Figure 10. In the MMSP(s) dataset, we use the satellite image of 64x64 pixels and only keep the first channel. In the MMSP(L) dataset, we keep all 4 channels.

We employ a spatial and temporal downsampling approach to effectively manage the resolution and frequency of our dataset. Specifically, on the spatial dimension, we initially select an area of 640x640 pixels, corresponding to a 5km resolution. However, to reduce computational complexity, we downsample the input image to 64x64 pixels, equivalent to a 50km resolution. The selected area and the locations of plants of interest are shown in Figure 11. On the temporal dimension, we downsample the data to a frequency of every 60 minutes. This downsampling approach allowed us to maintain essential temporal information while reducing the overall volume of data. The dataset spans January 1, 2021, to December 31, 2022, encompassing the entire timeline of solar power data used in our study.

A.3 Numerical weather prediction

We rely on the Numerical Weather Prediction (NWP) data offered by The European Centre for Medium-Range Weather Forecasts (ECMWF). Renowned for its expertise in global atmospheric modelling, ECMWF grants access to high-resolution NWP datasets. These datasets are generated using the state-of-the-art Integrated Forecasting System (IFS), which empowers researchers with comprehensive and reliable information for medium-range weather predictions.

This data is streamed as an online service and is regularly updated four times a day. It offers global weather predictions for several days in advance, with a temporal resolution of 60 minutes. This real-time and regularly updated nature of the dataset allows for accurate and up-to-date forecasting of weather patterns at a global level.

The NWP dataset contains essential meteorological information that plays a crucial role in forecasting and understanding weather patterns. From the NWP dataset, we select 17 columns that contain crucial weather features. These features encompass a range of variables including wind, temperature, pressure, cloud cover, and solar radiation. The selected feature names are as follows:

(1) Clear-sky direct solar radiation at surface,
(2) Direct solar radiation
(3) Downward UV radiation at the surface
(4) Surface solar radiation downwards
(5) Surface net solar radiation
(6) Surface pressure
(7) Sunshine duration
(8) Low cloud cover
(9) Total cloud cover
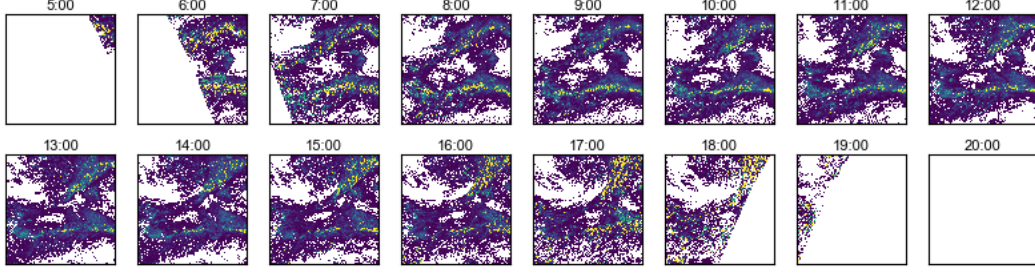(10) 2 metre temperature
(11) 2 metre dewpoint temperature
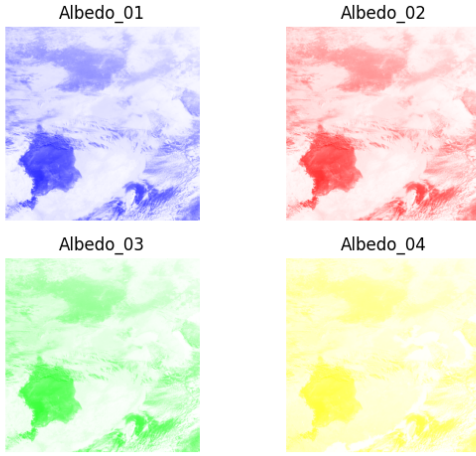
Figure 9: The cloud optical thickness over the daytime.

Figure 10: Visualization of satellite data on Himawari8/9 (first 4 bands: Aldebo 01 to Aldebo 04).

Figure 11: The selected coverage area of satellite imagery data and the locations of the solar power plants.

(12) Skin temperature

(13) Total precipitation

(14) 100 metre U wind component

(15) 100 metre V wind component

The NWP data is provided at a resolution of 10km. To align the NWP data with the solar power plant, we assign the plant to the nearest point on the NWP grid. Before inputting the weather features into the model, we perform a normalization process. This step ensures that all the weather variables are on a consistent scale.

B ROTARY POSITION EMBEDDING (ROPE)

Positional encoding has been proven to be an effective component within Transformer architectures. In our specific scenario, relative positional encoding is more suitable than absolute positional encoding. This is due to the fact that the relevance of a solar station is largely determined by the proximity.

Rotary Positional Encoding (RoPE) [28] is an efficacious method for encoding relative position information. The fundamental objective of RoPE is to d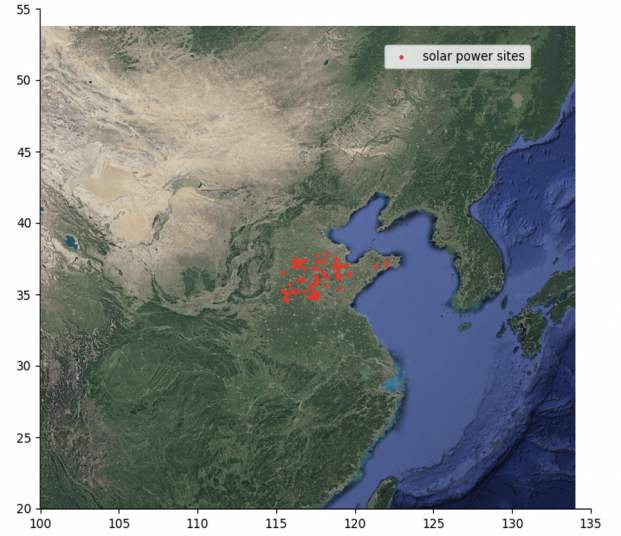evise a mechanism whereby the inner product inherently captures and represents positional data in terms of relative distances and relationships:

$$\langle f_q(\mathbf{x}_m, m), f_k(\mathbf{x}_n, n) \rangle = g(\mathbf{x}_m, \mathbf{x}_n, m - n), \tag{7}$$

where $\mathbf{x}_m$ and $\mathbf{x}_n$ are the embeddings of query (Q) and key (K). Their relative distance is $m - n$. So the goal is to solve the functions $f_q(\mathbf{x}_m, m)$ and $f_k(\mathbf{x}_n, n)$ to conform the aforementioned relation.

Following a thorough mathematical derivation as outlined in [28], we arrive at the formulation of $f_{q,k}$ that adheres to the previously mentioned relation. The expression in a d-dimensional space is given by:

$$f_{q,k}(\mathbf{x}_m, m) = \mathbf{R}^d_{\Theta,m} \mathbf{W}_{q,k} \mathbf{x}_m, \tag{8}$$

where

$$
\mathbf{R}_{\Theta,\mathbf{m}}^{d} =
\begin{bmatrix}
\cos m_1 & -\sin m_1 & 0 & \cdots & 0 & 0 \\
\sin m_1 & \cos m_1 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cos m_2 & -\sin m_2 & \cdots & 0 \\
0 & 0 & \sin m_2 & \cos m_2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \cos m_{\frac{d}{2}} & -\sin m_{\frac{d}{2}} \\
0 & 0 & 0 & 0 & \sin m_{\frac{d}{2}} & \cos m_{\frac{d}{2}}
\end{bmatrix}
$$

is the rotary matrix with pre-defined parameters $\Theta = \{\theta_i = 10000^{-2(i-1)/d}\}, i \in [1, 2, ..., d/2]$.

Leveraging the sparsity of $\mathbf{R}_{\Theta,\mathbf{m}}^{d}$, a more computationally efficient realization could be implemented in the code [28], as follows:

$$
\mathbf{R}_{\Theta,m}^{d} x =
\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_{d-1} \\ x_d \end{pmatrix}
\otimes
\begin{pmatrix} \cos m\theta_1 \\ \cos m\theta_1 \\ \cos m\theta_2 \\ \cos m\theta_2 \\ \vdots \\ \cos m\theta_{\frac{d}{2}} \\ \cos m\theta_{\frac{d}{2}} \end{pmatrix}
+
\begin{pmatrix} -x_2 \\ x_1 \\ -x_4 \\ x_3 \\ \vdots \\ -x_{d-1} \\ x_d \end{pmatrix}
\otimes
\begin{pmatrix} \sin m\theta_1 \\ \sin m\theta_1 \\ \sin m\theta_2 \\ \sin m\theta_2 \\ \vdots \\ \sin m\theta_{\frac{d}{2}} \\ \sin m\theta_{\frac{d}{2}} \end{pmatrix}
\tag{9}
$$

So given input x and its positional embeddings sin and cos, the algorithm for employing RoPE in a self-attention mechanism is shown in Algorithm 1.

Algorithm 1 Self Attention with Rotary Positional Encoding

Input: x of shape [B, N, D], sin of shape [B, N, D], cos of shape [B, N, D]

Output: out of shape [B, N, D]

1: **procedure** FORWARDPASS(x, sin, cos)
2:      $q \leftarrow$ TO_Q(x)   ▷ Project input to query
3:      $k, v \leftarrow$ TO_KV(x).CHUNK(2)   ▷ Project input to key and value
4:      $q \leftarrow q * cos +$ ROTATE_EVERY_TWO(q) $* sin$ ▷ Function of $\mathbf{R}_{\Theta,m}^{d} x$ defined in (9)
5:      $k \leftarrow k * cos +$ ROTATE_EVERY_TWO(k) $* cos$ ▷ Function of $\mathbf{R}_{\Theta,m}^{d} x$ defined in (9)
6:      $dots \leftarrow$ EINSUM("$bid, bjd-> bij$", q, k)   ▷ Compute attention matrix
7:      $attn \leftarrow$ SOFTMAX($dots$)
8:      $attn \leftarrow$ DROPOUT($attn$)
9:      $out \leftarrow$ EINSUM("$bij, bjd-> bid$", $attn, v$)
10:      $out \leftarrow$ TO_OUT(out)   ▷ Project output with linear layer
11:      **return** out
12: **end procedure**

C IMPLEMENTATION DETAILS

The training of FusionSF is conducted on a single Nvidia V100 GPU, utilizing a batch size of 16. During the training phase, the AdamW optimizer [22] was leveraged, accompanied by a weight decay parameter set to 0.05. The following list delineates the hyperparameters configured for FusionSF:

```
patch_size: [8, 8]
image_size: [64, 64]
ctx_channels: 1
ts_channels: 1
pe_type: rope
use_glu: True
freq_type: lucidrains
max_freq: 128
ctx_masking_ratio: 0.99
ts_masking_ratio: 0
dim: 64
depth: 12
heads: 8
mlp_ratio: 4
dim_head: 64
dropout: 0.4
num_mlp_heads: 1
decoder_dim: 128
decoder_depth: 4
decoder_heads: 6
decoder_dim_head: 128
vq_in_ts: True,
vq_in_ctx: True,
vq_in_guide: False,
```

Listing 1: Hyperparameters of FusionSF
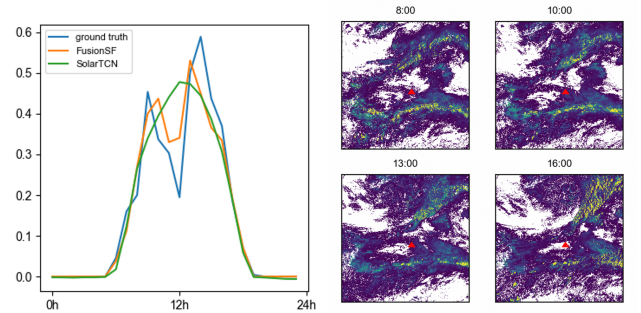
D A CASE STUDY

Figure 12: An example of 24h power predictions. The left figure shows the true power curve over two consecutive days and predictions by FusionSF (TS+NWP+Satellite) and SolarTCN (TS+NWP). The right figure shows the cloud optical thickness over the daytime of the second day. The site position is marked by a red triangle.

In this section, we provide an example to demonstrate how the multi-modality data helps to improve the predictions on 'hard' cases. As shown in Figure 12 (Left), the target day is not a typical sunny day, where power increases with a clear-sky pattern before noon and then decreases sharply in the afternoon. There is a bias in weather prediction, as a result, the predicted power curve (by SolarTCN) with only TS and NWP as input deviates from the true curve. Usually, such a phenomenon is owed to the motion of clouds, where some thick cloud covers the site in the afternoon. Figure 12 (Right) shows the cloud coverage during the daytime, in which the red triangle marks the site position. Notice that a cloud cluster is moving northeast, and the site is on its edge before noon and then obscured. This explains why the scale of power is slightly smaller

Table 6: Comparative analysis of model performance on MMSP(S) dataset across "All", "Easy", and "Hard" scenarios. We use MAE(↓) and RMSE(↓) as metrics, including parameter size and floating point operations (FLOPs) to measure model efficacy. The best results are highlighted in bold, and the second best results are highlighted with <u>underline</u>.

Models	Parameters	FLOPs	All (25210)		Easy (18014)		Hard (7196)	
			MAE	RMSE	MAE	RMSE	MAE	RMSE
Persistence	-	-	0.06500	0.13909	<u>0.04763</u>	0.10279	0.10838	0.20319
Mean	-	-	0.07632	0.12849	0.07674	0.12614	<u>0.07528</u>	<u>0.13417</u>
Clear sky	-	-	0.07347	0.15682	0.05589	0.12196	0.11748	0.22119
Informer [40]	7.5M	175M	0.07973	0.13086	0.07952	0.12867	0.08025	0.13613
Autoformer [34]	7.1M	204M	0.07830	0.11702	0.07015	0.09876	0.10285	0.15505
Crossformer [39]	58M	328M	0.06599	0.11259	0.06201	0.10173	0.08440	0.14645
PatchTST [23]	9.5M	85.3M	0.06575	0.11755	0.06056	0.10192	0.08320	0.14783
FiLM [41]	4.7M	42.5M	0.06995	0.12529	0.05783	0.09468	0.10474	0.18154
Dlinear [37]	1.2K	34.6K	0.07609	0.12310	0.06364	0.09762	0.10682	0.17035
LightTS [38]	0.11M	321K	0.06474	<u>0.11048</u>	0.05724	<u>0.09347</u>	0.08324	0.14413
CrossViVit [5]	3.8M	1.24B	<u>0.05789</u>	0.11818	0.04891	0.09924	0.08007	0.15535
FusionSF	4.3M	1.25B	**0.04020**	**0.08881**	**0.03891**	**0.08359**	**0.04980**	**0.10690**

than the previous day and a significant cutdown occurs in the early afternoon.

E FULL BENCHMARK

In Table 6, we provide the performance along with an analysis of the computational complexity in the benchmark. While FusionSF introduces additional modalities, the resultant increase in computational complexity remains acceptable in comparison to baseline models.