Multimodal LLMs as Image-to-Tabular Translators for Ultra-Short-Term Photovoltaic Forecasting

Diego de Carvalho Neves da Fontoura[1], Marilton Sanchotene de Aguiar[1]

[1] Graduate Program in Computing (PPGC) – Center for Technological Development (CDTec)
Federal University of Pelotas (UFPel)
Pelotas – RS – Brazil

{dcnfontoura,marilton}@inf.ufpel.edu.br

Abstract. *Ultra-short-term photovoltaic forecasting is strongly affected by cloud dynamics, yet integrating visual information typically requires complex, dedicated computer vision architectures. This study proposes an alternative approach in which multimodal LLMs (OpenAI and Gemini) serve as image-to-tabular translators. The method converts directional sky-camera images into numeric variables, which are then fused with conventional predictors, such as inverter and meteorological data. Experimental evaluations across 5-, 10-, and 15-minute horizons demonstrate that LLM-enhanced configurations consistently outperform a strong tabular baseline. The results suggest that LLMs can serve as a modular and auditable intermediate layer between raw imagery and standard forecasting models. Limitations include modest performance gains and reliance on a single pilot installation, requiring future multi-site validation.*

Keywords: Photovoltaic forecasting; Multimodal large language models; Sky images; Image-to-tabular translation; Time-series forecasting.

1. Introduction

The increasing penetration of photovoltaic (PV) generation makes ultra-short-term forecasting especially relevant for operational planning, reserve allocation, storage control, and the stable integration of renewable energy into power systems [Ahmed et al. 2020, Antonanzas et al. 2016, Voyant et al. 2017]. At 5–15 minute horizons, forecasting accuracy is strongly affected by rapid local cloud transitions, which are not always well represented by tabular meteorological measurements alone [Chow et al. 2015, Zhen et al. 2020]. This challenge is also relevant in the broader context of renewable-energy integration, since improving short-horizon PV predictability contributes to more reliable operation under highly variable atmospheric conditions.

Recent work has increasingly explored sky images, public cameras, satellites, optical flow, convolutional networks, transformers, and multimodal fusion for solar nowcasting [Sharma and Elmenreich 2021, Paletta et al. 2022, Paletta et al. 2023, Sarkis et al. 2024, Straub et al. 2024, Xu et al. 2024, Zhang et al. 2023]. These studies consistently show that visual information can improve short-term irradiance and generation forecasting.

However, they also reveal an important methodological constraint: image information is typically incorporated through bespoke computer-vision pipelines, such as segmentation stages, motion-estimation procedures, or specialized deep visual encoders [Paletta et al. 2022, Paletta et al. 2023, Straub et al. 2024, Xu et al. 2024, Zhang et al. 2023]. As a result, transforming raw sky images into structured predictors often remains tightly coupled to domain-specific visual architectures.

This paper investigates an alternative computational route in which multimodal large language models (LLMs) act as *image-to-tabular translators*, leveraging recent advances in visual-language models and multimodal instruction tuning [Alayrac et al. 2022, Li et al. 2023, Liu et al. 2023]. In the proposed formulation, grouped multi-view sky images are processed to produce numeric cloud-related contrast proxies, which are then integrated with power, meteorological, luminosity, solar-geometry, and temporal variables in conventional forecasting models.

Rather than replacing forecasting models with LLMs, the proposed formulation introduces a modular and interpretable intermediate layer between raw imagery and tabular pipelines. Grounded on a pilot dataset, the study evaluates whether LLM-derived image variables add predictive signal beyond a strong tabular baseline. More specifically, we test the hypothesis that such variables provide complementary information under leakage-aware temporal evaluation, using an expanding-window walk-forward protocol consistent with rolling-origin assessment for time-series forecasting [Bergmeir et al. 2018]. To support transparency and reproducibility, the complete experimental artifacts (code, sanitized datasets, prompts, figures, and results) are publicly available.[1]

The main contributions are: (i) a formalization of multimodal LLMs as image-to-tabular translators for ultra-short-term PV forecasting; (ii) a pilot multimodal dataset integrating camera interpretations, inverter power, meteorology, luminosity, and solar geometry; (iii) a leakage-aware experimental protocol based on expanding-window walk-forward validation; and (iv) empirical evidence that LLM-derived variables provide modest but consistent gains over a strong tabular baseline.

The remainder of the paper is organized as follows. Section 2 discusses related background on ultra-short-term PV forecasting, sky-image methods, and multimodal LLMs. Section 3 presents the study design and dataset, detailing the pilot setup, data sources, preprocessing steps, and the LLM prompting protocol. Section 4 outlines the experimental protocol. Section 5 reports the experimental results and discusses their implications. Section 6 outlines the main threats to validity, and Section 7 concludes the paper.

2. Background

Ultra-short-term PV forecasting under fast cloud variability has been studied through persistence baselines, statistical models, machine learning, and deep learning [Voyant et al. 2017, Antonanzas et al. 2016, Ahmed et al. 2020]. When the forecasting horizon is only a few minutes ahead, recent power values are often highly informative, but visual cloud evolution may still add relevant local information, especially during ramps [Chow et al. 2015, Garniwa et al. 2023, Sarkis et al. 2024].

The literature on sky-image forecasting has expanded considerably. Recent studies combine optical flow and recurrent networks [Garniwa et al. 2023],

[1] `https://github.com/diegonevesdafontoura/semish-2026-multimodal-llm-pv-forecasting`

polar-invariant representations [Paletta et al. 2022], attention-based sequence modeling [Ukwuoma et al. 2024], transformer-based visual encoders [Xu et al. 2024], and multimodal fusion with weather and satellite data [Paletta et al. 2023, Straub et al. 2024, Zhang et al. 2023]. These results support the idea that image information improves short-horizon forecasting, but they also show that image processing is typically embedded in specialized computer vision architectures.

In parallel, multimodal LLMs have evolved toward more reliable joint processing of text and images [Alayrac et al. 2022, Li et al. 2023, Liu et al. 2023]. Yet, in PV forecasting, their use remains underexplored. The relevance of investigating them in this context is not to replace the forecasting model itself with an LLM, but to test whether they can operate as structured perception modules that convert sky imagery into numeric predictors. This fact creates a distinct methodological possibility: instead of building a bespoke vision pipeline, one may use a multimodal LLM as an interpretable intermediate layer that preserves compatibility with conventional regression models while enabling explicit comparison between scenarios with and without image-derived information.

From a broader perspective, this work aligns with integration challenges between software and hardware systems in real-world sensing environments, where heterogeneous physical sensors, visual acquisition devices, and computational models must be connected through structured representations suitable for predictive modeling.

3. Study Design and Dataset

3.1. Pilot Setup and Data Sources

The study uses data from a pilot photovoltaic installation comprising a visual acquisition subsystem, a grid-connected inverter, a local luminosity-sensing module, an external meteorological source, and astronomical solar-position computation. Visual data were acquired with an Intelbras iM7 S Full Color outdoor camera, while active power measurements were obtained from a Solis-1P5K-4G inverter. In addition, local luminosity readings were collected by an ESP32-based sensing module and transmitted to a local server, and meteorological variables were retrieved from the Open-Meteo service. Figure 1(a) shows the installation context, and Figure 1(b) presents an example of the grouped camera views used for multimodal interpretation.

The data acquisition workflow is script-based. Inverter power was collected from the device's local web interface by periodically issuing HTTP requests to extract the instantaneous power variable (`webdata_now_p`). The ESP32 device sent luminosity readings to a Flask-based local server through a dedicated endpoint, which stored timestamped records including raw and normalized values. Meteorological variables such as temperature, relative humidity, dew point, pressure, wind speed, wind direction, wind gusts, radiation, and precipitation were periodically queried from Open-Meteo and stored in CSV format. Astronomical variables, including solar altitude and azimuth, were computed separately and later integrated into the forecasting dataset.

The Intelbras iM7 S Full Color camera provides Full HD acquisition (1920×1080) with motorized directional control, while the Solis-1P5K-4G inverter provides the active power measurements used as forecasting targets. These hardware characteristics are relevant because camera optics, positioning, and acquisition stability directly affect the visual content processed by the multimodal LLMs.

(a) Physical setup of the pilot PV installation.

(b) Grouped multi-view sky image used as LLM input.

Figure 1. Pilot setup and grouped sky image used in the proposed pipeline.

The acquisition scripts were designed to operate within controlled daytime windows and to align collection times to full-minute timestamps, reducing temporal inconsistency across sources. After collection, all sources were synchronized to a common 5-minute grid for forecasting. These implementation details are relevant because camera optics and positioning, inverter polling, ESP32 sensing behavior, meteorological update frequency, and timestamp alignment directly affect the quality, reproducibility, and transferability of the multimodal dataset.

3.2. LLM-derived Variables and Preprocessing

For each 5-minute timestamp, the grouped sky image is interpreted independently through multimodal inference, i.e., visual image input combined with a fixed textual prompt that produces structured numeric outputs by OpenAI's GPT-4o and Google's Gemini 2.5 Flash [OpenAI 2025, Google DeepMind 2025]. Each provider returns directional values for south, north, east, west, and top views, plus an aggregated score. Table 1 summarizes the main variable groups.

Table 1. Main numeric variables derived from multimodal image interpretation.

Variable group	Range	Interpretation
`value_{dir}`	0.0 to 1.0	OpenAI-derived directional cloud-related contrast proxy.
`value_{dir}_gemini`	0.0 to 1.0	Gemini-derived directional cloud-related contrast proxy.
`result`	0.0 to 1.0	Aggregated OpenAI cloud-related contrast proxy.
`result_gemini`	0.0 to 1.0	Aggregated Gemini cloud-related contrast proxy.

Only numeric outputs were used in the predictive pipeline. Records with values outside the expected range were discarded, and all sources were synchronized on a 5-minute grid. Only samples collected between 07:00 and 19:00 were retained. This filtering removes periods with low or null solar irradiance, ensuring that the dataset focuses on operational photovoltaic conditions. After filtering valid targets, the final modeling dataset contains 27,686 samples.

This adjustment reduces the artificial variance introduced by nighttime or near-zero generation periods, resulting in a more realistic and challenging forecasting setup.

All preprocessing steps were strictly causal to avoid temporal leakage. Power was resampled to 5-minute means, meteorological variables were forward-filled, and lagged features (1, 2, 3, 6, and 12 steps) and rolling statistics were computed from shifted series. Forecasting targets were defined as future shifts for 5-, 10-, and 15-minute horizons.

Additional features include directional variability, inter-model differences, and horizon–zenith contrasts, resulting in 32 baseline predictors and up to 50 features when both LLM sources are used.

3.3. Multimodal LLM Prompting Protocol

To ensure reproducibility and consistency across providers, a structured prompting strategy was adopted for both OpenAI and Gemini models. The exact prompting template used in the experiments is shown in Figure 2.

```
=== PROMPT ===
prompt_text = """ This image is divided into 5 parts and represents the sky in different directions, based on
captures made by cameras located XXXX:
The center image represents the sky seen directly upwards (TOP).
The top image is NORTH.
The bottom image is SOUTH.
The image on the right is EAST.
The image on the left is WEST.
Each image shows the sky at the same moment but in different directions.
Please analyze the cloud cover in each of the 5 parts of the image and assign a numerical score between 0.0 and
1.0, where:
0.0 = completely clear sky (no clouds)
1.0 = completely covered by clouds (overcast)
Return:
north_value:
south_value:
east_value:
west_value:
top_value:
result: simple or weighted average of the values above.
open_description: A descriptive classification of the overall sky, based on the weighted value, using the categories:
Clear_sky (0.0 – 0.2)
Partly_clear (0.3 – 0.4)
Partly_cloudy (0.5 – 0.7)
Cloudy_sky (0.8 – 1.0)
Avoid explaining the process; just present the values and the final classification based visually on the provided
image. """
```

Figure 2. Prompt template for image-to-tabular translation.

A fixed prompt template (in English) instructs the model to assign normalized cloud-related contrast values to five directional views (north, south, east, west, and top) in the range $[0, 1]$, where lower values indicate visually clearer conditions and higher values indicate stronger cloud presence or occlusion. The model returns directional variables (e.g., `north_value`, `south_value`) and an aggregated value (`result`), optionally accompanied by a categorical sky description. The output format is constrained to numeric values and predefined fields to facilitate automated parsing and integration into the forecasting pipeline.

Each grouped sky image (comprising south, north, east, west, and top views) was processed independently using the same instructions, without contextual carryover across timestamps.

To reduce variability, the prompt template was kept constant across all observations. The provider models were kept fixed throughout the experiments (GPT-4o for OpenAI and Gemini 2.5 Flash for Google Gemini), using the same prompt structure across all timestamps. The outputs were parsed automatically, and records with values outside the expected numeric range were discarded to ensure consistency.

Although different providers may internally interpret visual features differently, the protocol enforces a common output structure, enabling direct comparison and fusion of the resulting variables. This design prioritizes auditability and modularity, allowing the LLM layer to be treated as a structured perception component within the forecasting pipeline.

All implementation details, including prompts, preprocessing steps, evaluation scripts, sanitized datasets, figures, and results, are structured to support full reproducibility and are publicly available in the accompanying repository.

In addition, ChatGPT (OpenAI) was used exclusively for English language polishing and minor editorial refinement during manuscript preparation. It was not used to generate scientific results, fabricate data, create references, or alter the experimental analyses. All scientific design, analyses, interpretations, and final content validation remain the responsibility of the authors.

4. Experimental Protocol

Four feature scenarios are evaluated:

1. **Baseline tabular**: meteorology, luminosity, solar geometry, cyclical time variables, lagged power, and rolling power statistics.
2. **Baseline + OpenAI**: baseline tabular plus OpenAI-derived image variables.
3. **Baseline + Gemini**: baseline tabular plus Gemini-derived image variables.
4. **Baseline + both LLMs**: baseline tabular plus OpenAI and Gemini variables together.

Three forecasting horizons are considered: 5, 10, and 15 minutes ahead. The models evaluated in all scenarios are Linear Regression, Random Forest, and HistGradientBoosting. Model selection is performed per scenario and horizon.

The primary evaluation protocol is expanding-window walk-forward validation. The modeling dataset is chronologically sorted and split into three folds. In each fold, the training block grows over time, followed by a validation block of 4,152 samples and a test block of 2,768 samples. The training sizes are 13,843, 16,611, and 19,379 samples in folds 1, 2, and 3, respectively. The mean validation RMSE across folds selects the best model for each scenario, and performance is reported as the mean and standard deviation on the test blocks. After validation-based model selection within each scenario, feature scenarios are compared using their corresponding walk-forward test performance. This protocol is stronger than a single fixed split because it reduces dependence on one specific temporal window.

For completeness, a chronological single-split experiment was also run as a sensitivity analysis after the leakage corrections. It produced the same qualitative ranking of scenarios as the walk-forward protocol, so the article emphasizes the walk-forward results as the main evidence.

5. Results and Discussion

Table 2 summarizes the main walk-forward results after restricting the analysis to operational daylight periods (07:00–19:00). In all horizons, the best LLM-enhanced configuration outperforms the strongest tabular baseline. At 5 minutes, the best result is achieved with HistGradientBoosting in the OpenAI-augmented scenario, reducing RMSE from 509.19 to 504.31 and increasing R^2 from 0.8438 to 0.8470. At 10 minutes, the best result is achieved with Random Forest using the combined OpenAI+Gemini scenario, reducing RMSE from 577.36 to 574.33 and improving R^2 from 0.8003 to 0.8025. At 15 minutes, the best result is also obtained when both LLM sources are combined with Random Forest, reducing RMSE from 600.66 to 596.38 and increasing R^2 from 0.7849 to 0.7881.

Table 2. Main walk-forward results comparing the strongest tabular baseline with the best LLM-enhanced configuration at each horizon.

Horizon	Best tabular baseline					Best LLM-enhanced setting				
	Scenario	Model	RMSE	Std.	R^2	Scenario	Model	RMSE	Std.	R^2
5 min	Baseline tabular	HistGradientBoosting	509.19	67.25	0.8438	Baseline + OpenAI	HistGradientBoosting	504.31	64.50	0.8470
10 min	Baseline tabular	Random Forest	577.36	65.43	0.8003	Baseline + OpenAI + Gemini	Random Forest	574.33	64.42	0.8025
15 min	Baseline tabular	Random Forest	600.66	61.41	0.7849	Baseline + OpenAI + Gemini	Random Forest	596.38	58.48	0.7881

Although the observed improvements are modest, they are consistent across a more restrictive, physically meaningful evaluation setting. After restricting the analysis to daylight operational periods (07:00–19:00), the best LLM-enhanced configuration outperforms the strongest tabular baseline across all three horizons, reducing RMSE from 509.19 to 504.31 at 5 minutes, from 577.36 to 574.33 at 10 minutes, and from 600.66 to 596.38 at 15 minutes. In ultra-short-term forecasting, where recent power values dominate predictive performance, achieving consistent gains under this setting is inherently difficult. Therefore, even small reductions in RMSE indicate that LLM-derived variables capture complementary information not fully represented by conventional predictors.

The reduction in performance gains compared to preliminary experiments can be explained by the removal of low-variance periods associated with near-zero generation. In such conditions, simple autoregressive models tend to perform well, potentially inflating baseline performance. By focusing only on active generation periods, the evaluation becomes more realistic, and the contribution of multimodal LLM-derived variables is assessed under stricter conditions.

The ranking of scenarios also reveals an important pattern. OpenAI-derived variables provide the best performance at the shortest horizon (5 minutes), while the combined OpenAI+Gemini configuration becomes superior at 10 and 15 minutes. This suggests that the two providers encode partially distinct aspects of the visual scene, and that their combination becomes increasingly informative as the forecasting horizon increases, when autoregressive effects weaken, and cloud dynamics play a larger role.

Feature-importance inspection for the 5-minute horizon provides additional insight. In all scenarios, `power_5min` is the most influential predictor, followed by variables such as `radiation_wm2`, `power_roll_mean_6`, and cyclical time features. At the same time, image-derived variables such as `result` and `result_gemini`, as well as directional contrast features, consistently appear among relevant predictors, supporting their role as complementary sources of information.

Figure 3 illustrates the predictions produced by the best-performing configurations in a representative evaluation window. The models capture the main intra-day dynamics of photovoltaic generation, including ramp-up and ramp-down periods. The predicted curves remain close to the observed trajectory across all horizons, with the largest deviations concentrated around sharper transitions, which is consistent with the quantitative results.

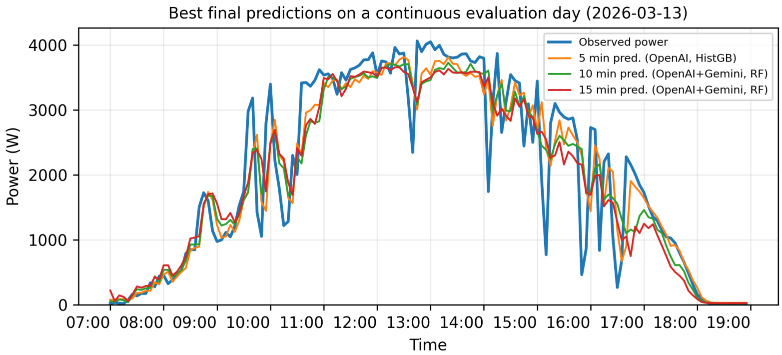

Figure 3. Observed PV power and predictions for the best configurations across all horizons.

5.1. Single-Split Sensitivity Analysis

For completeness, a chronological single-split experiment was also conducted after the leakage corrections. Its results were consistent with the main walk-forward evidence, preserving the same qualitative ordering of scenarios across horizons: the strongest LLM-enhanced settings remained superior to the strongest tabular baseline, with OpenAI features performing best at 5 minutes and the combined OpenAI+Gemini setting performing best at 10 and 15 minutes. Although the article emphasizes walk-forward validation as the main source of evidence, this agreement across validation schemes reinforces the robustness of the central finding that LLM-derived image variables contribute a complementary predictive signal.

5.2. External Holdout Test on Unseen Days

As an additional robustness check, three full days (2026-02-25, 2026-02-26, and 2026-02-27) were removed from the modeling dataset before the main training and validation pipeline. These days were therefore not visible during training, validation, or internal testing. After model selection on the remaining data, the best 5-minute configuration (Baseline + OpenAI with HistGradientBoosting) was retrained on the reduced modeling set and evaluated on this external holdout.

The aggregated results on the external holdout were MAE = 276.81, RMSE = 489.48, and $R^2 = 0.8638$ over 410 samples. Table 3 details the performance by day. The best behavior was observed on 2026-02-27, with RMSE = 133.84 and $R^2 = 0.9909$, consistent with the smoother, more structured intra-day profile observed on that day. By contrast, 2026-02-25 and 2026-02-26 were more challenging due to stronger short-term fluctuations and sharper local variations. Even so, the model preserved the overall trajectory of photovoltaic generation in all three cases, as illustrated in Figure 4.

These results reinforce the paper's main conclusion: the selected LLM-enhanced 5-minute model is not only competitive under walk-forward validation but also generalizes to fully unseen days outside the internal model-selection pipeline, providing additional evidence that the proposed formulation remains robust beyond the temporal partitions used for model selection.

Table 3. External holdout results for the best 5-minute model on three unseen days.

Subset	MAE	RMSE	R^2	n
Overall	276.81	489.48	0.8638	410
2026-02-25	378.55	588.83	0.6483	142
2026-02-26	351.56	569.94	0.7935	144
2026-02-27	73.51	133.84	0.9909	124

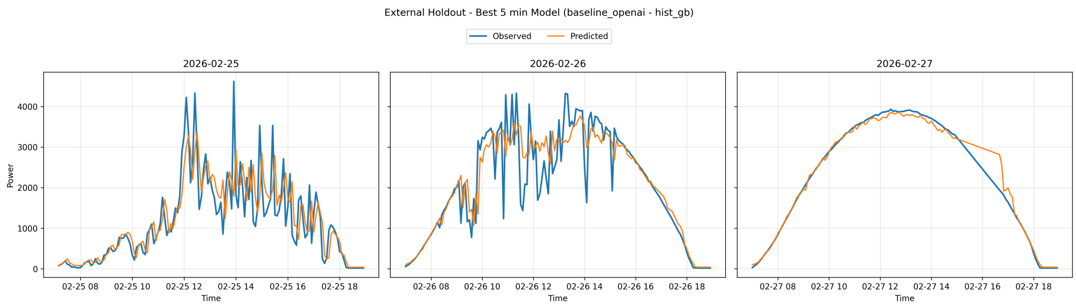

Figure 4. Observed and predicted PV power for the best 5-minute model on three unseen holdout days.

5.3. Additional Remarks on Statistical Validation

The present comparison is based primarily on walk-forward RMSE and R^2, complemented by scenario consistency across horizons and by the agreement between walk-forward and chronological single-split validation. Given the modest absolute gains and the limited scope of the pilot dataset, the results should be interpreted as evidence of methodological feasibility and initial predictive utility rather than as a definitive statistical claim of superiority. A more extensive formal significance analysis, including error-series tests across larger datasets and multiple sites, remains an important direction for future work.

6. Threats to Validity

The study still has important limitations. First, the evidence comes from a single pilot installation, so external validity across sites, climates, seasons, camera placements, and PV configurations remains open. The present results should therefore be interpreted as evidence of methodological feasibility and initial predictive utility, rather than as a definitive demonstration of broad generalization.

Hardware- and acquisition-specific factors, including camera lens and field of view, directional capture stability, inverter polling reliability, ESP32 sensor behavior, and timestamp alignment across heterogeneous sources, may affect the external validity of the reported results.

Second, the meteorological variables are not fully co-located with the PV module, which may introduce spatial mismatches during rapid local cloud transitions. This issue is particularly relevant in ultra-short-term forecasting, where small-scale atmospheric variability can affect the correspondence between local sky conditions and tabular weather observations.

Third, the dataset covers a limited temporal window and reflects the operational characteristics of a specific pilot setup. Although this is sufficient for a first controlled experiment, larger, longer multi-site datasets are still necessary to assess robustness across wider variability regimes and to determine whether the observed gains persist across different seasonal and climatic contexts.

Fourth, the proposed LLM-derived variables are constrained by prompt design, provider behavior, and the consistency of multimodal numeric outputs. While the study treats these outputs as structured predictors and applies the same preprocessing and forecasting protocol across scenarios, some degree of model-specific variability may remain, particularly because aggregated image scores are provider-generated rather than deterministically computed outside the LLM. This means that the reported results should be understood not only as evidence about the data, but also as evidence of the viability of this particular image-to-tabular formulation.

Fifth, the observed performance gains are modest. However, this should be interpreted in light of the baseline's strength, which already includes recent power values, lagged information, rolling statistics, radiation, luminosity, solar geometry, and cyclical temporal features. In this context, even small improvements are informative, since they suggest that image-derived LLM variables contribute a complementary signal beyond a competitive tabular reference. Operational deployment may also be constrained by API latency and inference cost, particularly in high-frequency real-time scenarios.

Finally, the study does not aim to establish that multimodal LLMs outperform specialized end-to-end vision architectures for PV forecasting. Instead, its goal is to evaluate whether LLMs can serve as auditable, modular intermediate layers for converting sky imagery into structured predictors. Direct comparisons with dedicated visual deep-learning pipelines, broader ablation studies, and larger multi-site validation remain important directions for future work.

7. Conclusion

This paper investigated a computational alternative for ultra-short-term photovoltaic forecasting, using multimodal LLMs as *image-to-tabular translators*. Instead of embedding sky imagery in a dedicated end-to-end visual architecture, the proposed approach converts grouped directional camera views into numeric cloud-related variables that can be directly fused with conventional predictors such as power, weather, luminosity, solar geometry, and temporal features.

The empirical results show that LLM-enhanced configurations consistently outperform a strong tabular baseline across 5-, 10-, and 15-minute forecasting horizons, supporting the study's central hypothesis that multimodal LLM-derived visual variables can provide complementary predictive information beyond conventional tabular predictors. The observed gains are modest but stable across evaluation settings and scientifically meaningful because the baseline already includes strong autoregressive, physical, and temporal predictors. A feature-importance inspection further suggests that the image-derived variables provide complementary information rather than replacing the dominant effect of recent power and radiation measurements.

The contribution of this work does not lie in replacing existing forecasting models with LLM-based approaches, but rather in introducing a modular, interpretable mechanism for incorporating visual information into tabular forecasting pipelines. By framing multimodal LLMs as image-to-tabular translators, the study provides a practical alternative to tightly coupled computer-vision architectures, enabling easier integration, inspection, and comparison within conventional modeling workflows.

More broadly, the proposed formulation may be relevant beyond the specific energy domain studied here. The idea of using multimodal LLMs as structured perception layers for converting visual observations into tabular predictors may be applicable to other multimodal time-series problems in which raw images contain contextual information that is difficult to integrate into conventional forecasting models.

Future work includes expanding the dataset in terms of time and geographic diversity, comparing the proposed formulation against specialized computer vision pipelines and pre-trained visual encoders such as CLIP and DINOv2, and evaluating more formal explainability analyses to quantify the contribution of image-derived variables. Additional directions include testing few-shot and visually calibrated prompts, defining deterministic aggregation rules outside the LLM, and measuring API cost and latency under realistic high-frequency deployment conditions, ideally in larger multi-site validation settings.

References

Ahmed, R., Sreeram, V., Mishra, Y., and Arif, M. (2020). A review and evaluation of the state-of-the-art in pv solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124:109792.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., and Barr, I. (2022). Flamingo: a visual language model for few-shot learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Proc. of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pages 23716–23736, New Orleans, USA. Curran Associates, Inc.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F., and Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111.

Bergmeir, C., Hyndman, R., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120(C):70–83.

Chow, C. W., Belongie, S., and Kleissl, J. (2015). Cloud motion and stability estimation for intra-hour solar forecasting. *Solar Energy*, 115:645–655.

Garniwa, P. M., Rajagukguk, R. A., Kamil, R., and Lee, H. (2023). Intraday forecast of global horizontal irradiance using optical flow method and long short-term memory model. *Solar Energy*, 252:234–251.

Google DeepMind (2025). Gemini 2.5 flash model. `https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash`.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proc. of the*

40th International Conference on Machine Learning (ICML), volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742, Honolulu, Hawaii. PMLR.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Proc. of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pages 34892–34916, New Orleans, USA. Curran Associates, Inc.

OpenAI (2025). Gpt models with vision. `https://platform.openai.com/docs/guides/images`.

Paletta, Q., Arbod, G., and Lasenby, J. (2023). Omnivision forecasting: Combining satellite and sky images for improved deterministic and probabilistic intra-hour solar energy predictions. *Applied Energy*, 336:120818.

Paletta, Q., Hu, A., Arbod, G., Blanc, P., and Lasenby, J. (2022). Spin: Simplifying polar invariance for neural networks application to vision-based irradiance forecasting. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5182–5191, New Orleans, USA. IEEE.

Sarkis, R., Oguz, I., Psaltis, D., Paolone, M., Moser, C., and Lambertini, L. (2024). Intraday solar irradiance forecasting using public cameras. *Solar Energy*, 275:112600.

Sharma, E. and Elmenreich, W. (2021). A review on physical and data-driven based nowcasting methods using sky images. In Arai, K., editor, *Advances in Information and Communication*, pages 352–370. Springer International Publishing, Cham.

Straub, N., Herzberg, W., Dittmann, A., and Lorenz, E. (2024). Blending of a novel all sky imager model with persistence and a satellite based model for high-resolution irradiance nowcasting. *Solar Energy*, 269:112319.

Ukwuoma, C. C., Cai, D., Bamisile, O., Yin, H., Nneji, G. U., Monday, H. N., Oluwasanmi, A., and Huang, Q. (2024). An attention fused sequence -to-sequence convolutional neural network for accurate solar irradiance forecasting and prediction using sky images. *Renewable Energy*, 237:121692.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., and Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105:569–582.

Xu, S., Zhang, R., Ma, H., Ekanayake, C., and Cui, Y. (2024). On vision transformer for ultra-short-term forecasting of photovoltaic generation using sky images. *Solar Energy*, 267:112203.

Zhang, L., Wilson, R., Sumner, M., and Wu, Y. (2023). Advanced multimodal fusion method for very short-term solar irradiance forecasting using sky images and meteorological data: A gate and transformer mechanism approach. *Renewable Energy*, 216:118952.

Zhen, Z., Liu, J., Zhang, Z., Wang, F., Chai, H., Yu, Y., Lu, X., Wang, T., and Lin, Y. (2020). Deep learning based surface irradiance mapping model for solar pv power forecasting using sky image. *IEEE Transactions on Industry Applications*, 56(4):3385–3396.