Vision-Language Models for Network Model-Single-Line Diagram Consistency Verification: A Feasibility Study

Charles Alba, *Graduate Student Member, IEEE*, Dean Miller, Rishabh Jain, *Member, IEEE*, Karthik Kumar, Seong Lok Choi, *Member, IEEE*, Fei Ding, *Senior Member, IEEE* and Benjamin Kroposki, *Fellow, IEEE*

***Abstract*—Single-line diagrams (SLDs) are central to power system planning and operations and must remain consistent with the machine-readable network models used by downstream applications – a manual, hard-to-scale reconciliation made increasingly pressing as network models are revised ever more frequently. Vision-Language Models (VLMs) could automate this, but their application to SLDs remains under-explored, in part because power systems is a low-resource domain lacking expert-validated benchmarks. We conduct a feasibility study on VLMs for SLD Consistency Verification: identifying discrepancies between an SLD and its text-based network model representations. We first find that general-purpose VLMs display underwhelming performances. Hypothesizing that this stems from an inability to attend to the relevant component within a dense diagram, we evaluate whether contrastive VLMs (e.g., CLIP) can localize queried components; finding that they too struggle, whereas traditional Optical Character Recognition (OCR) proves surprisingly effective, we propose an OCR-anchored cropping strategy. Integrating this with VLMs substantially improves performance, though still below expectations in safety-critical scenarios. As a preliminary feasibility study, we hope our findings direct attention toward resources spanning two future pathways for GenAI in power systems: (1) curating expert-validated benchmark datasets and (2) developing a power-systems foundation VLM.**

***Index Terms*—Power system single-line diagram, generative AI, vision-language models**

I. Introduction

THE rise of Generative Artificial Intelligence (GenAI) has fueled applications across a range of domains – from medicine to law [1] to energy [2]–[4]. Recent advances in multi-modal generative models – which can process and reason over two or more data modalities such as text, images, audio, or structured data – have opened opportunities in the power systems domain. One such opportunity is the use of Vision Language Models (VLMs) [5]–[8], which integrate image and text to analyze, interpret, and perform tasks – such as identifying errors or discrepancies – on single-line diagram (SLD) images, which are common in power systems [9]–[11].

Despite this potential, the application of VLMs to SLDs remains under-explored. One such task is SLD Consistency Verification: confirming that an SLD and its corresponding text-based network model agree – that the components, connectivity, and electrical parameters depicted in the diagram match those encoded in the machine-readable model used by power-flow, state-estimation, and energy-management software [10], [11]. Because inaccuracies degrade the downstream applications that rely on it – state estimation, contingency analysis, and the operational decisions they inform – its correctness is critical [12], [13]. As network models are revised ever more frequently, and because these updates are typically performed by hand, each is an opportunity for the diagram and model to diverge [9], [14]. Maintaining SLD-model consistency is therefore a quality-assurance prerequisite for the interconnection, planning, and energy-management workflows that rely on these models. Common errors include connectivity discrepancies (e.g., a switch in the wrong state) and parameter errors, in which values differ (e.g., mismatched voltages, power flows) [11]–[13]. Ensuring the model remains consistent with its SLD, however, is a largely manual task, in which engineers cross-check drawings against model records – a process that is labor-intensive and difficult to scale [10].

Unlike several of the aforementioned domains, which benefit from benchmark datasets whose images and texts are carefully curated, labeled, and validated by domain experts before being used to train and evaluate models [1], comparable resources do not exist for SLDs. Real-world diagrams and their network models are typically held privately by utilities or generated by proprietary software – raising ethical barriers to assembling a public benchmark for this task [3], [4]. This – alongside other vision-to-text tasks in the power systems domain – has often led to the hypothesis, albeit unproven, that power systems is a low-resource domain: scarce public data means these models are rarely exposed to it during training, leading to sub-par downstream performance [2]–[4]. Similar challenges have been recognized in fields like medicine and law [1], where data needed to train and evaluate models is difficult to obtain owing to privacy, confidentiality, or regulatory constraints. Unlike SLDs in power systems, however, these fields have progressed from early feasibility studies to expert-curated benchmarks [1] – resources that have since anchored research addressing the low-resource problem.

Even within the power-systems domain, other tasks across other modalities – graph-based learning for AC optimal power flow, time-series forecasting of power demand, and wind-speed and solar-irradiance prediction from geospatial data – have progressed from early feasibility studies to mature, openly available benchmarks [15]. The vision-to-text modality on which SLD understanding depends, however, has yet to make that transition; this work is intended as the feasibility study from which such a progression could begin for SLDs.

SLD Consistency Verification has yet to take this first step.

This work was authored by the National Laboratory of the Rockies (NLR) for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by Laboratory Directed Research and Development program. Seong Choi, Fei Ding and Benjamin Kroposki are with the Power Systems Engineering Center, National Laboratory of the Rockies, Golden, CO, USA (contact: fei.ding@nrel.gov). Charles Alba is with the Washington University in St Louis, Dean Miller is with Princeton University and Karthik Kumar is with the Rochester Institute of Technology. This work was completed during CA, DM and KK's time at NLR.

This work therefore serves as such a feasibility study. First, we task several commonly used VLMs with identifying discrepancies between a text-based network model and its SLD, where one has been intentionally manipulated. Finding that the VLMs produce underwhelming performances, we hypothesize that this stems in part from their inability to attend to the relevant component within a dense, full-diagram input. We therefore experiment with contrastive vision-language models – such as Contrastive Language-Image Pre-training (CLIP) – alongside a traditional Optical Character Recognition (OCR) approach. Finding that the former struggle to localize the queried component whilst OCR performs surprisingly well, we propose a crop-first pipeline that localizes the component via OCR and passes that focused region to the VLM, substantially improving discrepancy-detection performance – though still below what safety-critical deployment would demand. These findings lead us to recommend 2 next steps: (1) curating an extensive, expert-validated benchmark dataset in the power systems domain and (2) developing a power-systems foundation VLM.

II. Data and experimental Setup

As noted above, a central challenge in obtaining SLDs and their corresponding text-based network models is that they are typically held privately by utilities or generated by proprietary software – the latter of which could raise ethical or legal concerns if used to train or evaluate VLMs.

We therefore constructed our evaluation data from simulated and internally generated transmission and sub-transmission cases. For the IEEE 39-bus system, we used the `PyPowSybl` library [16] to generate both the SLDs and their corresponding network-model representations, which came in the form of JSON files. Additionally, we included internal SLDs, for which corresponding representations (in JSON format as well) were generated using an internal Neo4j-powered simulation workflow. These JSON files encode the network model corresponding to each SLD, following the bus-branch and node-breaker representations standard in power system modeling: buses with their voltage magnitudes and phase angles and branches (lines and transformers) with their connectivity and power-flow quantities (e.g., MW and MVAR values), alongside individual components – buses, breakers, transformers, generators, and loads – and their physical connections as relationships.

We used JSON as the text-based network-model representation because it was the native output of our data-generation workflows and could be reliably parsed by both VLMs and our evaluation pipeline. Its structure also enabled programmatic comparison between model outputs and ground truth, whereas converting it to formats like CIM/XML or PSS/E `.raw` would make this process more verbose and less suited to language-model evaluation. We intentionally manipulated the JSON files produced from power-flow simulations to introduce incorrect details spanning switch states, power-flow magnitude and direction (e.g., MW and MVAR values or signs), and component interconnections. In total, we obtained 139 SLD-JSON pairs, each containing between zero and three intentionally introduced discrepancies.

While this sample size falls well short of a benchmark, it is intended to support a feasibility study. Curating a sufficiently diverse, generalizable, and robust benchmark would require annotation or verification by power systems experts, or close collaboration with power utilities. Our aim is to use these findings to assess whether a push toward larger, more diverse community benchmarks is warranted.

III. Methods and results

A. Can VLMs identify discrepancies between SLD and their text-based network model representations?

We perform the experiments of our proposed methods on two advanced and state-of-the art models – `google/gemma-4-26B-A4B-it` [6] and `meta-llama/Llama-3.2-90B-Vision` [5].

Specifically, we prompt the model to "Identify the discrepancies between the SLD and JSON representation." and provide additional contextual information in the system prompt, including detailed descriptions of common component symbols (e.g., switch, generator and transformer symbols).

Our results reveal underwhelming performances among VLMs, with `Llama-3.2-90B-Vision` achieving 14.93%, 66.67% and 24.4% in precision, recall and F-score, respectively. `Gemma-4-26B-A4B-it` fares much better, displaying achieving low-to-moderate performances of 26.53%, 86.67% and 40.63% in precision, recall and F-score, respectively, but still short of the performance required for real-power deployment tasks, particularly in safety-critical scenarios.

To put the results in contexts, the models successfully flag most of the genuine discrepancies present (high recall), particularly for `gemma-4-26B`. However, majority of the discrepancies they report are false alarms (low precision). In other words, the models over-sensitively and indiscriminately report inconsistencies – a behavior consistent with the VLM mistaking its own misreadings of the dense diagram for genuine mismatches. In power systems, such a false-alarm rate is impractical for deployment, as operators cannot act on a verifier wrong in most cases it raises.

B. Can contrastive vision-language pretrained models (e.g., CLIP) accurately identify, segment and crop listed components from the model?

Finding that the VLMs display underwhelming performances when given the full diagram, we speculate this stems from difficulty visually resolving the specific component being verified, as SLDs can be dense with many closely packed components and fine-grained text labels. We therefore hypothesize that directing the model's attention to the relevant component – by localizing and cropping it prior to verification – would improve performance, which we evaluate in Section III-C.

This raises two questions: (1) Can advanced contrastive vision-language pretrained models (VLPM) identify power system components from an SLD, and (2) if so, how do we use them to localize a component within an SLD?

To do so, we compare state-of-the-art VLPMs – Google's CLIP-based `OWL-ViT` [7] and `Grounding DINO` [8]. The former builds on CLIP – a vision-language model trained with

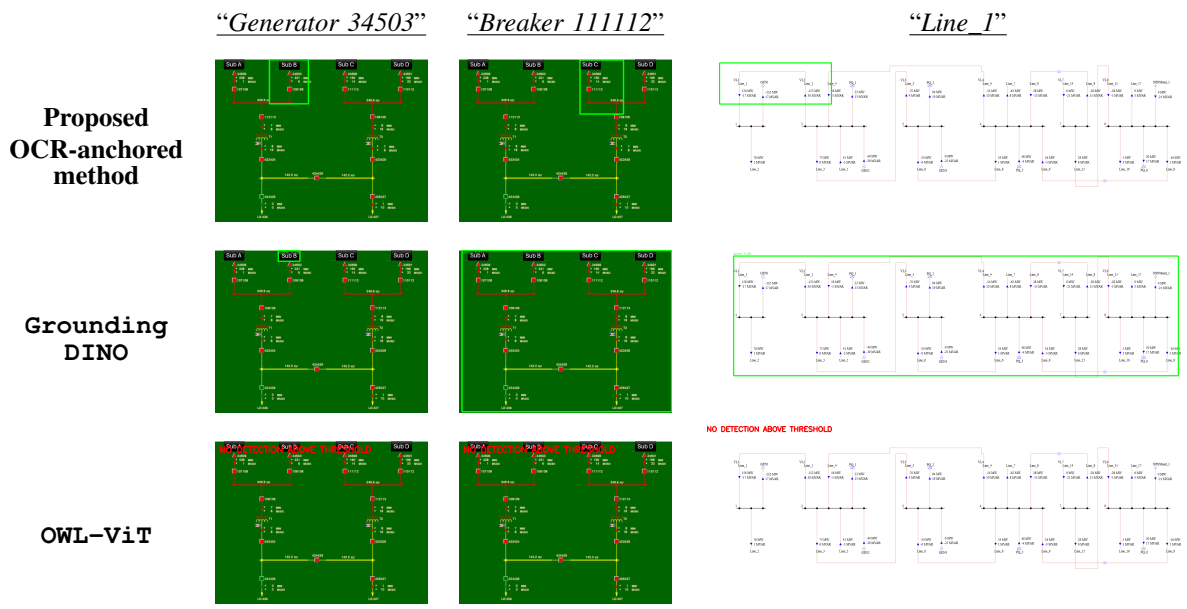

Fig. 1. Examples comparing our proposed OCR-anchored cropping method (top), `Grounding DINO` (middle) and `OWL-VIT` (bottom) when being tasked to identify a specific component in the SLD. As noted in the middle, `Grounding DINO` either highlights the incorrect component (e.g., *Generator 34503*) or selects the entire SLD (e.g., *Breaker 111112* and *Line 1*), whereas `OWL-VIT` consistently fails to detect any given component in the SLD.

a contrastive objective to embed images and text into a shared space – whilst the latter instead uses a series of visual encoders to localize the image region a textual prompt refers to.

Specifically, we parse each network model JSON to enumerate its listed components, and for each component we construct a textual query containing the component type and its ID as found in the SLD (e.g., "Transformer T1"), prompting each detector to localize the corresponding region within the SLD.

Our findings show that both `OWL-ViT` and `Grounding DINO` perform extremely poorly, as illustrated in Figure 1. In most cases, the detector either fails to localize any of the queried components or returns a box spanning nearly the entire SLD rather than the specific component. This is likely attributable to a domain mismatch: these detectors are trained predominantly on generic images, where the objects they learn to localize differ fundamentally from the densely packed symbols and fine-grained text labels present in SLD, as best illustrated in Figure 2.

Given that each component in an SLD is annotated with a text label corresponding to its identifier in the network model, we also take a step back to explore whether classic OCR could be used to locate them. OCR is a technique for detecting and transcribing text within an image – localizing each region of text and recognizing the characters it contains. However, OCR returns only the location of the text label itself, rather than the component together with its neighbouring, inter-connected components. To circumvent this, we propose a strategy built on top of it, where we construct bounding boxes to dynamically crop the SLD around the components specified in the query.

To construct a bounding box from the coordinates of the labels identified by OCR, determining suitable widths and heights for each bounding box is then pertinent: too narrow a crop may exclude neighbouring and inter-connected components, whilst too wide a crop may render the pipeline

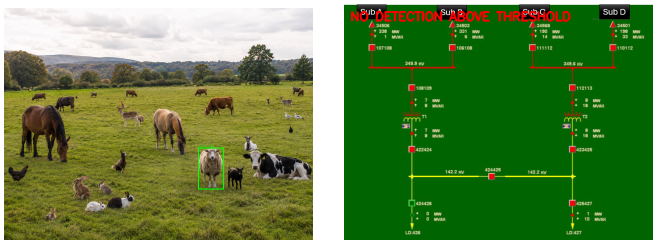

Fig. 2. An example of a CLIP-based `OWL-ViT` when asked to identify a sheep (left) from a generic image versus identifying "Transformer T1" from a SLD (right), illustrating how contrastive VLPM may struggle in SLDs despite excelling in generic images.

no more effective than supplying the full image. Although fixed dimensions relative to the image size are one option, SLD components are often heterogeneously distributed – some regions densely packed, others sparse, the latter necessitating larger crops. We therefore propose an algorithm that dynamically sizes each crop according to the density of components surrounding those specified in the query, as follows:

1) Where c_j is the image coordinates of any label identified via OCR within the SLD (e.g., "Bus 1", "G1"), and c_i is the image coordinates of the component of interest.

$$\Delta_j = \min_{i \neq j} \| c_i - c_j \|_2 . \tag{1}$$

2) For the coordinates of the component of interest in the image c_i and its neighboring set of components C,

$$\Delta_i = \min_{k \in C} \| c_i - c_k \|_2 . \tag{2}$$

3) With parameter α and empirical averages $\bar{\Delta}_i$ and $\bar{\Delta}_j$, choose the working scale

$$\bar{\Delta} = \begin{cases} \bar{\Delta}_i, & \text{if } \dfrac{\bar{\Delta}_i}{\bar{\Delta}_j} > \alpha, \\ \bar{\Delta}_j, & \text{otherwise.} \end{cases} \tag{3}$$

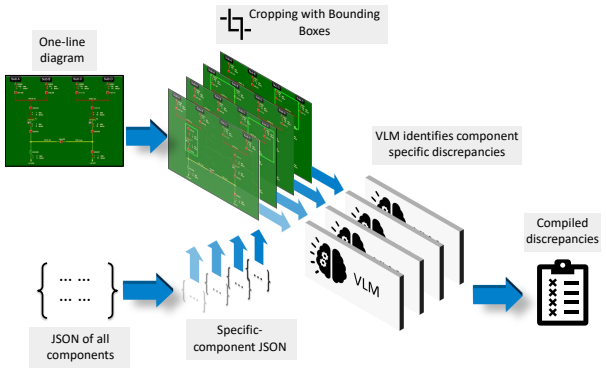

Fig. 3. Overview of the proposed OCR-anchored cropping pipeline: per-component crops and their JSON are passed to the VLM and compiled into a final discrepancy list.

4) Set padding with β as an adjustable hyper-parameter

$$\text{Pad} = \beta\,\bar{\Delta}. \tag{4}$$

Accordingly, we use the `SVTR` model [17] from the `PaddleOCR` package [18] to detect and recognize the text labels in the SLD (e.g., Bus 1", G1"). From the image coordinates of these detected labels we construct the bounding boxes (Equations 1-4) used to crop the SLD around the components specified in the user's query.

As noted in Figure 1, our proposed strategy performs substantially better than the open-vocabulary detectors such as `OWL-ViT` and `Grounding DINO`. Having established that the OCR-grounded framework reliably localizes queried components, we integrate it into the pipeline, supplying the VLM a focused crop rather than the full diagram. This allows us to test whether directing the model's attention to the specific component – as raised earlier in this paragraph – improves its ability to identify discrepancies.

C. Will integrating our proposed OCR-anchored strategy better assist VLMs in identifying discrepancies between SLD and their text-based network model representations?

Thus far, we have established that (1) VLMs produce underwhelming performances when tasked with identifying discrepancies between an SLD and its text-based network model; (2) contrastive VLPMs struggle to localize the queried components; and (3) a traditional OCR-anchored approach shows promise in doing so. This naturally leads us to ask whether combining the OCR-anchored strategy with the VLM as a sequential pipeline yields better discrepancy-detection performance – stemming from our earlier hypothesis that directing the model's attention to the relevant component, by localizing and cropping it prior to verification, would improve performance.

As illustrated in Fig. 3, our overall pipeline works as such: OCR, specifically SVTR [17], is used to detect and locate the exact coordinates of any labels denoted in the SLD (e.g., "Bus 1", "G1"). This is combined with the proposed OCR-centered framework (in Section III-B) to create "bounding-boxes" used to crop the image based on the components specified by the user in the query. This crop is then fed to the VLM for analysis, allowing the model to focus on the intended regions.

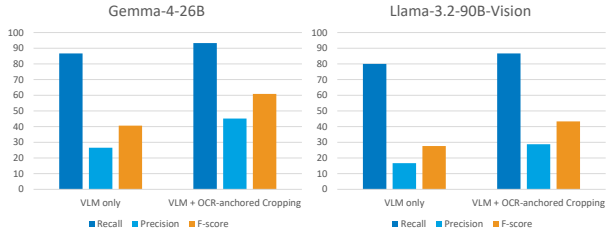

Fig. 4. Results comparing using only VLMs with the pipeline integration of VLMs with our proposed OCR-anchored cropping strategy (in Section III-B)

In doing so, we hope to mitigate the likelihood of VLMs misinterpreting or analyzing incorrect regions of the SLD. In the event where the query contains multiple components, each component could be parsed, analyzed in parallel, before being compiled post-analysis, per Fig. 3. If the OCR-centered framework fails to detect any of the queried components, the pipeline falls back on providing the entire SLD to the VLM. The same prompts and setups from Section III-A were used.

As shown in Figure 4, combining VLM and OCR-anchored cropping achieves substantial gains over the baselines across all evaluation metrics. Specifically, for `Llama-3.2-90B-Vision`, the corresponding improvements are from 14.93% to 30.95% for precision, 66.67% to 86.67% for recall, and 24.4% to 45.62% for F-score. For `gemma-4-26B`, precision improves from 26.53% to 45.16%, recall from 86.67% to 93.33%, and F-score from 40.63% to 60.87%. Overall, `gemma-4-26B` achieves much better performances and the OCR-anchored integration into VLMs allows it to attain moderate performances.

While these gains are notable, and they allowed `gemma-4-26B` in particular to achieve moderate performances, they still fall short of the performance expected for real-world deployment, particularly in tasks surrounding safety-critical power systems. In particular, the low precision and moderate F-scores imply that many flagged discrepancies are false alarms, ultimately risking alert fatigue, with operators dismissing genuine discrepancies alongside the spurious ones.

IV. Discussion

As GenAI accelerates progress across industries, calls for its application to the power systems domain – particularly to the manual, labor-intensive tasks – continue to grow [4]. VLMs are especially promising, as the sector is heavily diagram-driven: SLDs, three-line diagrams, and protection and wiring schematics are central to planning, operation, and maintenance [9]–[11]. Ensuring that an SLD remains consistent with its network model is a critical quality-assurance step, made necessary because models and diagrams are continually revised and often manually entered and updated [9], [14]. This reconciliation has traditionally been performed by hand – labor-intensive and difficult to scale – presenting a natural opportunity for VLMs to assist [12], [13]. As network models are revised ever more frequently, automated SLD-model verification grows

increasingly valuable as a scalable check on the data integrity that downstream applications depend on.

Our study serves as an early-stage, preliminary feasibility study of VLMs for network model-SLD consistency verification, framed specifically as identifying discrepancies between a text-based network model and the system depicted in its SLD. More broadly, this is one instance of a larger problem – querying and interpreting power-system diagrams by posing targeted questions of an SLD (or similar schematic) and reasoning over its contents – of which discrepancy identification is just one example [10].

Taken together, results of Sections III-A and III-B suggest that modern general-purpose VLMs and open-vocabulary detectors may underperform in this setting in part because power systems is a low-resource domain: having seen little such data during training, these models struggle when confronted with power-system schematics [1], [3], [10]. Our proposed OCR-anchored approach – a comparatively classical computer-vision approach – showed promise when integrated as a pipeline with VLMs, lending further support to this interpretation. The improvements observed suggest that the models may have struggled to attend to too many components at once; narrowing focus to the relevant region appears to alleviate this. Additionally, the high recall but low precision suggests the models are overly cautious – flagging any potential discrepancy – at the cost of reporting many that are not present. Finally, the low precision and moderate F-scores from our efforts to integrate an OCR-anchored approach with VLMs suggest that Closing this gap likely requires more than attention direction – namely, the resources this study motivates: expert-validated benchmarks and a power-systems foundation VLM pretrained on schematic imagery.

We note this preliminary study lacks an evaluation set large enough for robust conclusions. We hope that our findings can bring attention to the need for larger, more rigorous benchmark datasets in this domain – resources whose creation will require funding, cooperation from utilities, and substantial manual effort from experts versed in power systems. We hope this study demonstrates enough promise to motivate that investment, offering an early evidence base for the value of pursuing VLM-based tools for power-system diagram understanding, alongside the curated datasets such tools will require.

Future Directions

1) Curating extensive benchmark datasets in the power systems domain: Our findings reinforce that power systems is a low-resource domain in the pre-trained model era which demands vast amounts of domain-specific data. In the power systems domain, such data is scarce and not publicly available – guarded by proprietary software or utilities, with some unavailable for training due to security concerns [10]. Henceforth, building rigorous benchmarks requires sustained funding and cross-sector cooperation between utilities, vendors, and research institutions, as well as considerable manual effort of engineers versed in power systems. Comparable efforts have already borne fruit in other constrained domains, such as medicine or law [1] and other modalities within the power system domain (e.g., graph representations in `GridSFM`'s benchmark [15]). Closing that gap is pertinent for the subsequent progress of realizing VLM's potential in this domain.

2) Developing an open-source power-systems foundation VLM: General-purpose VLMs struggle with this domain's fundamentals – identifying components, reading labels, and reasoning over their relationships from images, as demonstrated from our study [3], [4]. This could be attributed the small volume of power-system data these models encounter during training. A foundation VLM imbued with domain knowledge of power systems is therefore needed to internalize the domain's symbols, conventions, components, and topological relationships. This has already borne fruit in other modalities in the power-grid domain, such as the `GridSFM` Foundation Model for AC Optimal Power Flow [15] and its graph representations, thanks to the progression from feasibility studies to benchmarks – something which could likewise catch on in the vision-language domain.

REFERENCES

[1] B. K. Rachmat, T. Gerald, Z. Z. Slb, and C. Grouin, "QA analysis in medical and legal domains: A survey of data augmentation in low-resource settings," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (SRW)*, 2025, pp. 1132–1144.

[2] S. L. Choi, R. Jain, P. Emami, K. Wadsack, F. Ding *et al.*, "egridgpt: Trustworthy AI in the control room," National Renewable Energy Laboratory (NREL), Golden, CO (United States), Tech. Rep., 2024.

[3] C. Alba, "Benchmarking small language models in the renewable energy domain," in *Proceedings of the 18th IEEE/ACM International Conference on Utility and Cloud Computing*, 2025, pp. 1–2.

[4] H. F. Hamann, B. Gjorgiev, T. Brunschwiler, L. S. Martins, A. Puech, A. Varbella, J. Weiss, J. Bernabe-Moreno, A. B. Massé, S. L. Choi *et al.*, "Foundation models for the electric power grid," *Joule*, vol. 8, no. 12, pp. 3245–3258, 2024.

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle *et al.*, "The llama 3 herd of models," Meta, Tech. Rep., 2024.

[6] Gemma Team, S. El Abd, V. Aggarwal, R. Algayres *et al.*, "Gemma 4 Technical Report," Google, Tech. Rep. arXiv:2607.02770, 2026.

[7] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, *et al.*, "Simple open-vocabulary object detection," in *European conference on computer vision*. Springer, 2022, pp. 728–755.

[8] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*. Springer, 2024, pp. 38–55.

[9] A. Monticelli, *State Estimation in Electric Power Systems: A Generalized Approach*. Boston, MA: Springer, 1999.

[10] C. Wu, H. Yu, Y. Liu, and C. Gong, "Towards reliable power grid modeling from drawings: A review of intelligent understanding, topology inference, and model generation," *Machines*, vol. 14, no. 4, p. 371, 2026.

[11] A. Abur and A. Gómez-Expósito, *Power System State Estimation: Theory and Implementation*. New York: Marcel Dekker, 2004.

[12] F. Wu and E. Liu, "Detection of topology errors by state estimation," *IEEE Transactions on Power Systems*, vol. 4, no. 1, pp. 176–183, 1989.

[13] W.-H. E. Liu, F. F. Wu, and S.-M. Lun, "Estimation of parameter errors from measurement residuals in state estimation," *IEEE Transactions on Power Systems*, vol. 7, no. 1, pp. 81–89, 1992.

[14] E. M. Lourenço, E. P. R. Coelho, and B. C. Pal, "Topology error and bad data processing in generalized state estimation," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3190–3200, 2014.

[15] W. Yang, A. Britto, T. Spina *et al.*, "GridSFM: A foundation model for AC optimal power flow," Microsoft Research, Tech. Rep., 2026.

[16] LF Energy, "Powsybl (power system blocks)," 2025.

[17] Y. Du, Z. Chen, C. Jia, X. Yin, T. Zheng, C. Li, Y. Du, and Y.-G. Jiang, "SVTR: Scene text recognition with a single visual model," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI-22*, 7 2022, pp. 884–890.

[18] C. Cui, T. Sun, M. Lin, T. Gao, Y. Zhang, J. Liu, X. Wang *et al.*, "Paddleocr 3.0 technical report," *arXiv preprint arXiv:2507.05595*, 2025.