Volume 09, Issue 04, (2025) Article 22

Signal Processing and Renewable Energy (SPRE)

https://doi.org/10.57647/j.spre.2025.0904.22

PiGT-2F: Physics-Informed Graph Transformer for Robust Short-Term Distribution Forecasting

Mohammad Shahbazi, Hamid Tohidi, Majid Nojavan

Department of Industrial Engineering, ST.C. Islamic Azad University, Tehran, Iran

***Corresponding author:** h.tohidi@iau.ac.ir

Original Research

Received:
25 September 2025

Revised:
28 October 2025

Accepted:
08 November 2025

Published in Issue:
31 December 2025

Abstract

High-resolution node-level forecasts of net load and distributed photovoltaic (PV) injection are increasingly essential for modern distribution system operations and planning. Yet most feeders remain sparsely metered and subject to intermittent telemetry loss, degrading purely statistical forecasting pipelines. We present PiGT-2F, a physics-informed graph Transformer that (i) learns jointly across electrically coupled feeders, (ii) modulates spatial attention by branch admittance, and (iii) embeds soft penalties that encourage compliance with Kirchhoff's Current Law and ampacity limits. Evaluated on a multi-feeder benchmark constructed from public feeder models, national load profiles, PV telemetry, and realistic random-plus-structured telemetry gaps, PiGT-2F lowers system-average nRMSE by 15–34% across 5- to 60-minute horizons and cuts physics-violation metrics by up to 80% relative to strong deep learning baselines. The archite-cture's near-linear temporal attention cost enables multi-day history windows, and its physics regular-ization provides resilience when the grid is partially blind. Evaluation follows a feeder-level holdout to assess cross-topology generalization, and includes an anonymized real-telemetry micro-validation; code, configuration files, a benchmark generator, and a containerized reproduction package accompany the paper as described in the "Code and Data Availability" section.

Keywords: Ampacity; Curriculum masking; Distribution forecasting; Graph Transformer; Kirchhoff's laws; Missing data; Physics-informed ML; PV variability

Cite this article: Shahbazi, M., Tohidi, H., & Nojavan, M., PiGT-2F: Physics-Informed Graph Transformer for Robust Short-Term Distribution Forecasting, *Signal Process. Renew. Energy.* *9*(4) Article 22 (2025). https://doi.org/10.57647/j.spre.2025.0904.22

1. Introduction

Electrification of heating and transport, rapid data-center siting, and widespread behind-the-meter PV adoption are driving sharper intra-hour variability in distribution-level net load, eroding planning margins and motivating feeder-granular short-term forecasting. Distribution and customer-scale forecasting remain harder than bulk system forecasting because spatial diversity is low, customer mix is heterogeneous, and metering is sparse or noisy. Recent industry and academic assessments document load revisions and increasing heterogeneity across regions, underscoring the need for methods that transfer information across feeders while tolerating missing data [1-7].

Visibility is limited: SCADA refresh intervals of minutes, batched AMI uploads, missing or mis-phased meters, and stale asset registries are common; entire feeder segments can go dark due to communications failures. Existing approaches (matrix completion, pseudo-measurements) often treat data recovery as a pre-processing stage rather than embedding mask-aware learning directly in the forecasting architecture. Regulatory and DOE reports highlight persistent telemetry quality challenges and motivate analytics that degrade gracefully during outages [8-14].

Graph neural networks (GNNs) are natural for power systems because network topology and impedances encode relational structure. Physics-informed GNNs have improved state estimation and dynamic

reconfiguration by embedding Kirchhoff constraints and equipment data, but prior work largely targets single feeders or present-state inference rather than multi-horizon forecasting under heavy missingness [15-21].

Transformer sequence models capture long temporal dependencies but scale poorly with dense attention; recent sparse/dilated variants reduce cost yet seldom incorporate power-system physics or feeder topology. Efficient attention mechanisms (e.g., Dozer Attention) and long-horizon TS Transformers (PatchTST, TimesNet, DLinear) suggest avenues to scale forecasting histories if coupled with domain structure [22-25].

From a probabilistic modeling angle, complex-valued Gaussian process regressions that explicitly handle amplitude–phase structure in oscillatory time series can improve calibration and uncertainty quantification in regimes with strong periodicity [26].

Public benchmark gaps hinder progress. Proprietary utility data cannot usually be shared; IEEE feeders lack synchronized load/PV time series. National datasets, the NREL End-Use Load Profiles (ResStock/ComStock) and PVWatts/NSRDB-provide calibrated, meteorology-aligned load and solar traces that can be fused with public feeder topologies to form open multi-feeder benchmarks for reproducible research [1], [27-30].

Goal. The main goal of this work is to develop and rigorously evaluate a scalable, physics-informed, multi-feeder forecasting architecture that significantly enhances accuracy, physical plausibility, and robustness to realistic telemetry gaps. Unlike most previous approaches, we seek to outperform state-of-the-art statistical, machine learning, and deep learning baselines not just in terms of standard error metrics, but also by ensuring that our forecasts remain physically meaningful and resilient even under conditions of severe data loss or missing measurements.

This underscores the practical significance of PiGT-2F under realistic data sparsity and outages.

Distribution utilities face rapid load growth (electri-ficat-ion, data centers) and rising PV variability on infrastructure that was not instrumented for dense situational awareness. Telemetry gaps in AMI/SCADA streams reduce the value of purely statistical forecasting pipelines.

PiGT-2F blends electrical-parameter-aware graph attention with efficient temporal context and soft physics losses so that forecasts interpolate spatially, extrapolate through outages, and surface physically implausible conditions before they trigger operational errors. By pairing public feeder models with versioned, meteorology-consistent national datasets (EULP building loads; PVWatts/NSRDB solar), we provide a testbed for distribution-scale forecasting under realistic data quality limitations.

Problem Statement. Given historical masked telemetry, exogenous drivers (such as weather and calendar effects), and detailed feeder topology/ parameters, we aim to predict real and reactive net injections for every observed and unobserved node over multiple future horizons $h \in \{5, 15, 30, 60\}$ minutes. The method must satisfy several key criteria:

(i) the ability to transfer statistical information across different feeders;

(ii) efficient temporal modeling over long historical windows;

(iii) incorporation of soft physical constraints such as KCL and ampacity; and

(iv) robust performance under both random and structured patterns of missing data [15], [22-25], [31]. This challenging setting reflects the realities faced by utilities in practice.

Contributions. Our main contributions in this paper are as follows:

(C1)Multi-Feeder Heterogeneous Graph Representation: We introduce a unified heterogeneous graph representation that combines multiple real-world feeders, enabling knowledge transfer and relational reasoning across feeder boundaries.

(C2) Admittance-Modulated Graph Attention: By modulating graph attention weights according to electrical admittance, we prioritize physically relevant connections and promote greater stability and interpretability in the model.

(C3) Dilated Sliding-Window Temporal A-ttention: Our architecture employs an efficient temporal attention mechanism that leverages both local density and long-range, seasonally adaptive context, scaling to multi-day historical windows with nearly linear cost.

(C4) Soft Physics Losses via Calibrated Linearized Branch-Flow Surrogate: We incorporate KCL and ampacity constraints through differentiable surrogate losses based on linearized branch-flow models, calibrated to AC solutions.

(C5) Open Semi-Synthetic Benchmark: We present a fully reproducible, semi-synthetic benchmark that fuses open IEEE feeder topologies, national building/PV datasets, and empirically derived telemetry masks, facilitating transparent evaluation under realistic conditions [1], [8-11], [27-30].

We adopt a feeder-level split for zero-shot evaluation on unseen feeders and further include a small-scale validation on anonymized real feeder telemetry; artifacts for reproducibility (configs, scripts, benchmark

generator, and a containerized executable) are provided as detailed later.

2. Related Work

2.1. Distribution-Level Load & DER Forecasting

Distribution-scale forecasting fundamentally differs from bulk system forecasting due to the unique characteristics of feeder-level loads and distributed energy resources (DERs). Customer compositions vary significantly from feeder to feeder, and the increasing penetration of PV can invert net load on short timescales, complicating prediction. Electrification trends and the rapid expansion of data centers are further intensifying localized growth rates, which often exceed historical system-level averages by a wide margin. These new realities have outpaced many legacy planning methodologies, making high-resolution, feeder-specific forecasting essential for both operations and infrast-ructure planning [4-7].

Recent research and industry reports have revealed that performance degrades sharply when traditional system-level models are naively down-scaled to the distribution level. The error budgets are dominated by heterogeneity, low spatial diversity, and the volatility introduced by DERs. While some approaches have successfully leveraged the fusion of weather data, calendar variables, and customer segmentation, most remain limited by siloed, feeder-specific designs and are not robust to telemetry gaps.

Transformer-based architectures, especially those designed for multi-horizon forecasting (such as TFT), along with emerging benchmarks, highlight the ongoing need for richer contextual features and improved methods to address missing data [23-24], [31].

Classical subspace tracking methods that adaptively estimate low-dimensional signal subspaces under streaming data offer complementary tools for handling structured variability and have been studied with rigorous asymptotic convergence analyses [32].

2.2. Data Availability, AMI Quality, and Telemetry Gaps

Access to high-quality AMI (Advanced Metering Infrastructure) and SCADA (Supervisory Control and Data Acquisition) data at the distribution level is fraught with challenges.

Industry-wide assessments consistently report significant intervals of missing data, delays in data uploads, and mismatches between physical devices and their phase or location mappings. These quality issues can severely degrade the accuracy and reliability of downstream analytics and real-time decision-making. Utilities also document widespread feeder-level data outages, often tied to network events or backhaul congestion, and these outages can persist from minutes to hours [8-11]. From a machine learning standpoint, the pattern of missingness itself is critical. Studies have demonstrated that model robustness can change dramatically depending on whether the gaps in telemetry are random, temporally correlated, or structured as large contiguous outages.

As a result, recent work emphasizes the importance of architectures that ingest mask information directly and of evaluation protocols that account for the impact of missing data on the end-use task, rather than merely focusing on imputing the raw data. These principles are embedded in our use of curriculum masking and physics-aware regularization [12-14].

2.3. Physics-Informed Machine Learning in Power Systems

Incorporating physical domain knowledge into machine learning models offers multiple benefits: it regularizes the training process, improves out-of-distribution generalization, and helps prevent physically implausible or infeasible predictions. Recent advances include physics-informed graphical neural networks for state estimation, which exploit power flow relationships and network connectivity to maintain high estimation accuracy even under sparse measurement conditions. Similarly, models for dynamic network reconfiguration that encode switching feasibility and electrical characteristics have demonstrated accelerated learning and improved performance in operational settings [15-16]. These approaches build upon a rich foundation of circuit and power-flow-based modeling that has been central to power systems engineering for decades. Rather than solving the full nonlinear AC power flow equations within the training loop, modern physics-informed ML often relies on soft penalty terms based on linearized network equations. This provides a computationally efficient way to discourage gross violations of key physical laws (such as KCL and ampacity limits), while allowing the model to remain flexible and scalable. Our method follows this philosophy, adapting DistFlow and branch-flow approximations for use in forecasting tasks [17-21].

2.4. Graph and Transformer Models for Spatio-Temporal Forecasting

Graph neural networks are uniquely suited to modeling power systems, as they naturally encode the grid's

topology, connectivity, and physical relationships. When combined with temporal encoders, GNNs have shown strong results in diverse spatio-temporal applications such as traffic forecasting, climate modeling, and increasingly, energy systems. Nevertheless, the majority of prior work in power systems has focused either on transmission networks or on static estimation tasks; relatively little has addressed the combination of short-term, multi-feeder distribution forecasting and missing data [15-16].

In parallel, Transformer-family models have driven substantial progress in time-series forecasting. PatchTST tokenizes local temporal patches, TimesNet models both periodic and aperiodic temporal variation, and DLinear achieves surprisingly strong results via simple linear decomposition. While models like TFT introduce variable selection and known-future exogenous inputs for interpretability, none of these frameworks natively include network physics-highlighting the need for hybrid, domain-aware approaches [23-25], [31]. Gaussian-process-based sequence models [26] provide an alternative for uncertainty-aware temporal structure learning, though scalability and topology-awareness remain open challenges relative to graph/Transformer hybrids. Complementary signal processing studies on compressive sensing-covering measurement design (e.g., Noise let matrices), algorithmic recovery via MP/OMP/L1, and noise robustness—also inform robustness under under-sampling and masking in forecasting Pipelines [33-35]. In applied vision, mean-shift–based detection exemplifies classical nonparametric preprocessing that can precede learning models [36].

2.5. Efficient Temporal Attention for Long Sequences

Standard self-attention mechanisms scale quadratically with sequence length, making them impractical for multiday, high-resolution histories required in distribution system forecasting. To address this, a variety of sparse and dilated attention schemes have emerged. For instance, Dozer Attention combines dense local coverage with seasonally adaptive sparse connections, enabling efficient long-term dependency capture without prohibitive computational cost. Such designs inform the sliding-window and memory token approach used in our architecture [22].

Similarly, the PatchTST and DLinear models demonstrate that many forecasting tasks can be addressed with either reduced attention density or alternative temporal transformations. Inspired by these insights, PiGT-2F employs geometric dilation and seasonal memory, striking a balance between history coverage and runtime efficiency [23-25].

Fig. 1 compares dense versus dilated temporal coverage over a 64-step history, illustrating how geometric dilation preserves seasonal memory at near-linear cost.

2.6. Linearized Distribution Power-Flow Models

Modeling the full nonlinear AC power flow is computationally intensive and often unnecessary for the forecasting context. Linearized models such as DistFlow, LinDistFlow, and various branch-flow relaxations have therefore become the standard for learning-based power system analysis. The seminal Baran–Wu DistFlow model, along with convexification techniques for branch-flow equations, is widely used as a fast surrogate for studying voltages, flows, and losses in distribution networks [17-19].

Recent extensions of these models include generalized multiphase variants, calibration procedures to align with full AC simulations, and user-friendly guides for practitioners. By leveraging these developments, we construct a calibrated, linearized branch-flow module that provides differentiable physics losses during PiGT-2F training, ensuring that forecasts remain feasible and physically plausible [20-21], [37].

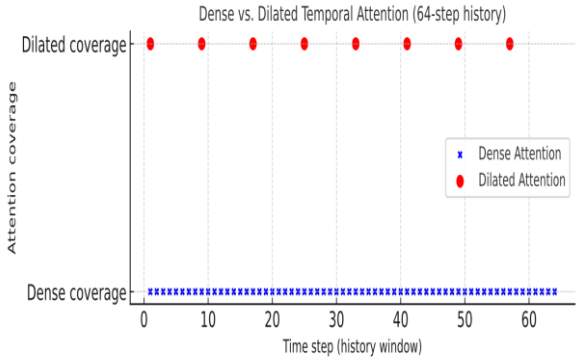

Figure 1. Dense vs. dilated temporal attention coverage for a 64-step history window. Dilated and seasonal connections enable scalable modeling of long histories, crucial for short-term forecasting under complex load and PV patterns

2.7. Meteorology-Grounded Public Data Resources

The development and testing of forecasting methods at the distribution scale require high-fidelity, meteorology-consistent load and PV production data. NREL's PVWatts tool, driven by the NSRDB resource database, enables the simulation of location-specific PV outputs using modern module models and weather records. The End-Use Load Profiles (EULP) project supplies detailed, 15-minute load shapes for residential and commercial buildings, calibrated across all U.S. climate zones. These resources, together, allow the creation of representative, non-proprietary feeder scenarios suitable for open research [1], [27-30].

Importantly, the versioned and script-accessible nature of these datasets allows researchers to easily regenerate scenarios for different climates, technology penetrations, and customer mixes. In our benchmark, we combine IEEE feeder topologies with EULP load and PVWatts /NSRDB PV data, overlaying empirical telemetry masks derived from industry reports to create a realistic, challenging, and fully reproducible dataset [1], [27-30].

2.8. Industry Telemetry Reports and Mask Design Motivation

Numerous federal and DOE reports have systematically cataloged the types and frequencies of telemetry issues in real distribution systems: from block outages, delayed uploads, and phase registration errors to feeder-wide communications failures. We used this empirical evidence to design our structured mask regimes, including lateral drops, head-end losses, and phase loss/swap scenarios. These patterns inform the training curriculum and provide rigorous tests of model robustness in realistic settings. Additionally, public documentation of the operational impacts of emerging load drivers-such as hyperscale data centers—reinforces the urgency and relevance of accurate, feeder-level forecasting [2-3], [8-11].

2.9. Research Gap Summary

Despite substantial progress, a critical gap remains: the absence of an open, physics-informed architecture that (i) learns across coupled feeders, (ii) incorporates electrical parameters into graph attention, (iii) scales temporal modeling efficiently, (iv) softly enforces key physical laws, and (v) is evaluated on a challenging, public benchmark with realistic telemetry gaps. This paper directly addresses these needs, aiming to provide a strong, extensible baseline for the community [1-3], [8-16], [22-25], [27-30].

3. Methods

Architecture. Our proposed PiGT-2F model represents the entire distribution system as a single directed, heterogeneous multigraph $G = (\mathcal{V}, \mathcal{E})$. Each node in the graph corresponds to a physical entity—substations, primary buses, lateral taps, distribution transformers, aggregated customer load points, DER/PV interconnects, or voltage regulation devices. Edges capture per-phase impedance/admittance, ampacity, and switch status, accurately reflecting the physical and operational relationships between components. Unlike typical per-feeder modeling, we incorporate feeder ID embeddings to allow cross-feeder knowledge transfer, while assigning low prior weights to unlikely inter-feeder connections. As depicted in Fig. 2, this unified graph representation ensures end-to-end differentiability across all connected feeders, capturing both their unique characteristics and shared patterns. This design is crucial for handling sparse and unbalanced measurement placements, as it enables masked nodes to "borrow" structure and information from better-instrumented or less noisy parts of the network.

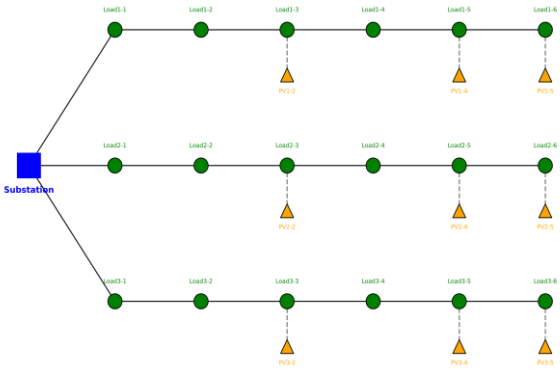

Figure 2. Illustrative multi-feeder node layout colored by asset type. Such represe-ntati-ons enable PiGT-2F to jointly learn across multiple, heterogeneous feeders and asset type

Cross-feeder sharing provides significant benefits in practice. By learning a joint representation, the model can transfer knowledge about temporal patterns, spatial dependencies, and response to exogenous drivers from

feeders with richer data to those with sparse measurements.

This not only enhances overall accuracy but is particularly advantageous for smaller or rural feeders that would otherwise lack sufficient historical data for effective training. We quantify these transfer benefits in detail in Section 5.

Cross-feeder transfer mechanism. All graph and temporal attention layers are shared across feeders; transfer arises from parameter sharing together with a learned feeder embedding derived from a categorical feeder ID. The feeder embedding is concatenated to each node's static features and injected through the node encoder, enabling the model to condition attention on feeder-level context while reusing the same parameters across heterogeneous topologies. Mini-batches mix windows from different feeders during training, so gradients aggregate across feeders and promote transfer.

Features and Masks. The model inputs include both dynamic and static node and edge features. Dynamic node features consist of historical real and reactive power telemetry (which may be masked), proxy PV irradiance from PVWatts/NSRDB, ambient temperature, and calendar signals (such as hour-of-day and day-of-week). Data is forward-aligned to a 5-minute grid, with latency flags included so that the model can distinguish true zeros from missing data. This setup, along with explicit mask tensors, ensures that the model never confuses data imputation with ground-truth measurements, improving robustness and interpretability.

Fig. 2 illustrates how these rich dynamic features-irradiance, customer class, and asset type are encoded and concatenated to each node, providing a comprehensive exogenous context. Static features encompass feeder ID, electrical distance from the substation, voltage class, phase vector, asset type, and customer-mix fractions based on EULP mapping. Edge features include resistance/reactance (R/X), shunt admittance, ampacity ratings, and switch status. Incorporating these detailed static attributes is essential for both physics-aware learning and accurate surrogate modeling of power flow [17-21]. The categorical feeder ID is mapped to a learnable low-dimensional embedding and concatenated to the static feature vector of every node.

Admittance-Modulated Graph Attention. The base attention logits are defined in Eq. (1), scaled by per-phase admittance in Eq. (2), and gated in Eq. (3). Let x_i denote the node embedding at layer ℓ and head h. Preliminary attention logits are

$$\tilde{e}_{ij}^{(l,h)} = \frac{\left(Q^{(l,h)}x_i\right)^{\top}\left(K^{(l,h)}x_j\right)}{\sqrt{d}}, \quad (1)$$

for $(i,j) \in \mathcal{E}$. We modulate by the per-phase magnitude of branch admittance via

$$\phi\left(|Y_{ij}|\right) = \frac{|Y_{ij}|}{\max_{(a,b)\in E}|Y_{ab}| + \epsilon}, \ \epsilon = 10^{-6} \quad (2)$$

and introduce a switch gate $g_{ij} = \sigma\left(z_{ij}\right)$ with dropout $p = 0.1$ during training on switchable edges. The final logits and attention weights are presented in compact form:

$$e_{ij}^{(l,h)} = \tilde{e}_{ij}^{(l,h)}\,\phi\left(|Y_{ij}|\right) + \log\left(g_{ij} + \epsilon\right), \ \alpha_{ij}^{(l,h)} = \mathrm{softmax}_{j\in\mathcal{N}(i)}\left(e_{ij}^{(l,h)}\right) \quad (3)$$

Dilated Sliding-Window Temporal Attention. We partition each history of length T into a dense local window of size w and B dilated bands with geometric dilation $= \gamma^{b-1}$, $b = 1, \ldots, B$, plus seasonal memory tokens at lags {24h,48h,168h}. This yields $O(T + \sum_b T/d_b)$ complexity, empirically close to linear in T.

Calibrated Linearized Branch-Flow Surrogate. We construct sparse linear operators (Hp, Hq) mapping nodal injections to per-branch current magnitudes under LinDistFlow.

The LinDistFlow mapping from nodal injections to branch currents is given in Eq. (4), with the equivalent voltage-based form in Eq. (5).

$$\left|\tilde{I}_{ij,t}\right| \approx \left|\left(H_p\right)_{ij,:}\hat{p}_t + \left(H_q\right)_{ij,}:\hat{q}_t\right|. \quad (4)$$

where voltage bases V_{base} are available, an equivalent formulation uses

$$\left|\tilde{I}_{ij,t}\right| \approx \frac{\left|\hat{S}_{ij,t}\right|}{\sqrt{3}\,V_{\text{base}}}, \quad (5)$$

with $\hat{S}_{ij,t}$ obtained by pushing nodal injections through the linearized branch-incidence mapping.

KCL Residual and Metrics. We define the per-phase KCL residual in Eq. (6) and the normalized magnitude metric in Eq. (7), and quantify ampacity duration in Eq. (8). Intuitively, Kirchhoff's Current Law states that the total current flowing into a node equals the total current flowing out; our loss penalizes deviations from that balance.

Ampacity is the thermal current rating of a line; forecasts implying currents above this rating are softly penalized to discourage thermally infeasible operation.For node i (per-phase), define

$$\delta_{i,t} = \sum_{(k,i)\in E}\tilde{I}_{ki,t} - \sum_{(i,k)\in E}\tilde{I}_{ik,t} - \frac{\hat{S}_{i,t}}{\sqrt{3}\,V_{i,\text{base}}}, \quad (6)$$

and the metric

$$\text{KCL-Mag} = \frac{1}{|T||V|}\sum_{t,i}\frac{\left\|\delta_{i,t}\right\|_2}{\bar{S}_i} \quad (7)$$

with $\bar{S}_i$ the typical apparent load scale at the node i (median over training). Ampacity duration is

$$|\tilde{I}_{ij,t}| \approx \sqrt{\left(\sum_{\varphi\in\Phi_{ij}} \left|Y_{ij}^{(\varphi)}\right| \left|V_{i,t}^{(\varphi)} - V_{j,t}^{(\varphi)}\right|\right)^2}, \quad (8)$$

Addendum: Multiphase-Coupled Current Surrogate and Phase-Consistent KCL.: To mitigate potential bias on unbalanced, Volt-VAR-heavy feeders with strong phase coupling, we augment the surrogate with cross-phase blocks learned from linearization around operating snapshots.

Let $\in C^{\Phi_{ij}}$ denote per-phase branch currents and $\hat{p}_t$, $\hat{q}_t \in R^{\Sigma_i \Phi_i}$ the stacked per-phase nodal injections. The multiphase surrogate and phase-consistent residual are formulated in Eq. (9) and Eq. (10), with the robust loss in Eq. (11).

$$\tilde{I}_{ij,t} \approx A_{ij}^{(p)} \hat{p}_t + A_{ij}^{(q)} \hat{q}_t, \quad (9)$$

where $A_{ij}^{(p)}$, $A_{ij}^{(q)}$ are block-sparse and include mutual (off-diagonal) phase terms; they are calibrated by least-squares to full AC snapshots. The node-wise KCL residual becomes a $vector$ across phases:

$$\delta_{i,t} = \sum_{(k,i)} \tilde{I}_{ki,t} - \sum_{(i,k)} \tilde{I}_{ik,t} - \frac{\hat{S}_{i,t}}{\sqrt{3}\, V_{i,\text{base}}}, \quad (10)$$

where divisions are elementwise by phase. We then use a robust, phase-consistent normalization and M-estimator:

$$KCL - Mag_{vec} = 1 /(|T||V|) \sum (t,i)\rho\left(\left(\left\|\delta_{(i,t)}\right\|_2\right) / \left(\overline{S_i^{(r)}}\right)\right), \overline{S_i^{(r)}} = \left((1/\Phi_i)\sum_{\varphi}\left(\overline{S_{(i,\varphi)}^2}\right)\right)^{(1/2)} \quad (11)$$

with ρ the Huber loss (scale by median absolute deviation per phase).

The training penalty uses a convex combination of the original scalar form and the new vector form:

$$L_{KCL}^{hyb} = (1-\eta)L_{KCL} + \eta\, mean\left(\rho\left(\left(\|\delta\|_2\right) / \left(\overline{S^{(r)}}\right)\right)\right), \eta = 0.7 \quad (12)$$

This preserves the original definition while improving faith-fulness under unbalance and Volt-VAR interactions.

Losses. The data loss is a Huber loss over observed targets, the KCL loss is mean $(\|\delta\|/\bar{S})$ and the ampacity loss is a squared hinge on normalized exceedances:

$$L_{\text{data}} = \frac{1}{N_{\text{obs}}}\sum_{\text{obs}} \text{Huber}(\hat{y}, y),\ L_{\text{KCL}} = \text{mean}(\|\delta\|/\bar{S}),\ L_{\text{amp}} = \frac{1}{|E|}\sum_{(i,j)}\left[\frac{|\tilde{I}_{ij}| - \bar{I}_{ij}}{\bar{I}_{ij}}\right]_+^2 \quad (13)$$

The total objective is

$$L = L_{\text{data}} + \lambda_{\text{KCL}} L_{\text{KCL}} + \lambda_{\text{amp}} L_{\text{amp}} + \beta L_{\text{reg}}. \quad (14)$$

Addendum Probabilistic Forecasting Head, CRPS, and Calibration. For ADMS/DERMS applications, we augment the decoder with a quantile head outputting $\{\hat{y}^{(\tau)}\}_{(\tau\in\tau_q)}$, τ_q ={0.05,0.1, . . ., 0.9,0.95}. The pinball loss is:

$$L_Q = \frac{1}{N_{obs}}\sum_{obs}\sum_{\tau\in\mathcal{T}_q}(\tau - \mathbb{1}\{y < \hat{y}^{(\tau)}\})(y - \hat{y}^{(\tau)}) \quad (15)$$

Following the identity$CRPS = \int_0^1 \text{pinball}_\tau\, d\tau$, we approximate CRPS via a Riemann sum over τ_q during evaluation. We also conformalize the central 90% band to guarantee coverage: with validation residual quantiles q_α, we inflate endpoints by a factor k chosen so empirical PICP matches the nominal level. The multitask objective becomes

$$L_{total}^{prob} = L + \lambda_Q L_Q + \lambda'_{KCL}\left(L_{KCL}^{hyb} - L_{KCL}\right) \quad (16)$$

with λQ = 1.0 and λ′KCL = 1.0 unless noted.

Curriculum Masking. Random-missingness rate follows a sigmoidal schedule over epochs e:

$$p_{\text{rand}}(e) = p_{\min} + (p_{\max} - p_{\min}) \cdot \sigma(\gamma(e - e_0)) \quad (17)$$

and the mixture weight for structured masks (block/ lateral/head-end/phase) increases linearly from 0 toαmax over the first 20 epochs. Masks are sampled independently per batch with stratification by feeder size so that validation regimes remain unseen.

Methods Summary. In summary, PiGT-2F performs data alignment, applies dilated sliding-window (DSW) temporal attention, computes admittance-modulated graph attention, decodes node-level forecasts, and calculates physics losses using a calibrated branch-flow surrogate. All components employ block-sparse kernel implementations for scalability. The resulting network is compact (about 9 million parameters), fast to train (under six hours on a single powerful GPU for the full benchmark), and highly robust to diverse data quality regimes.

Adaptive physics weighting. We additionally consider an adaptive scheme that modulates the physics losses by observability and uncertainty. For node i at

time t, $\boldsymbol{\lambda_{"KCL"}}(\boldsymbol{i},\boldsymbol{t}) = \boldsymbol{\lambda_{"KCL"}}^{\boldsymbol{"0}} \cdot \boldsymbol{g}(\boldsymbol{o_{i(t)}},\boldsymbol{\sigma_{i(t)}})$; for branch e, $\lambda_{"amp"}(e,t) = \lambda_{"amp"}{}^{"0} \cdot g(o_{e(t)},\sigma_{e(t)})$. Here $o(\cdot) \in [0,1]$ is an observability score derived from the telemetry mask (1 = fully observed within the receptive field), and $\sigma(\cdot)$ is a per-target predictive spread from the quantile head (inter-quantile range). We set $g(x,y) = "clip"(1 + \alpha(1 - x) + \beta\, z(y), [\ell, u])$, where $z(\cdot)$ is the batch-wise z-score; gradients are stopped through g. Unless stated otherwise, $(\alpha, \beta, \ell, u) = (0.5, 0.25, 0.5, 2.0)$. Default results in Section 5 use fixed weights; Section 5.7 reports a sensitivity analysis for the adaptive variant.

4. Data, Training Configuration, Baselines, And Evaluation

Public Feeder Topologies. For maximum transparency and reproducibility, we construct our benchmark using only public, open-access feeder models and metadata. Specifically, we fuse the IEEE 13, 34, 37, 123, and 8500-node distribution feeders, incorporating asset- and phase-type labels to capture the complexity of real-world systems. We apply topology corrections and phase-to-node mappings following established protocols [27-28], [37], and remove uninformative stub buses. To simulate real-world operational scenarios, feeders are interconnected using dummy tie switches, and metering locations are adopted from EPRI/IEEE test case metadata, closely mirroring realistic asset-level measurement deployments. This setup ensures that our experiments cover a wide range of network sizes and complexities, from small rural feeders to large urban grids.

Synthetic Net Load, PV, and Weather. Given the lack of synchronized, open distribution-level load and PV datasets, we generate high-fidelity synthetic time series using authoritative national resources. Residential, commercial, and industrial load traces are sourced from NREL's End-Use Load Profiles (EULP) project [1], [29-30] and are assigned to buses based on customer class and geographic location. PV injection profiles are generated using NREL's PVWatts and NSRDB tools, matched to the coordinates of each relevant node. Local weather data for each feeder is also drawn from NSRDB, ensuring meteoro-logical consistency between load and PV signals. All time series are upsampled from 15-minute to 5-minute intervals using a combination of linear and shape-aware interpolation. The final benchmark dataset comprises 6 feeders, roughly 14,000 nodes, two full years of data, and about 2 billion data points at 5-minute cadence-making it one of the most comprehensive open benchmarks for this task.

Table 1 summarizes node counts, load/PV sites, and the available years per feeder used in our benchmark.

Dataset scope and limitations. The semi-synthetic benchmark matches feeder physics and masking patterns but cannot fully capture all real-world drivers (e.g., weather uncertainty, fast inverter control dynamics, and rare operating contingencies). Our anonymized real-telemetry micro-validation (Section 5.7) partially addresses this gap by testing zero-shot and few-shot transfer on actual feeders. We view the benchmark as a controlled, reproducible testbed, and real-telemetry evaluation as a complementary external check.

Telemetry Mask Regimes. Real distribution systems rarely provide complete, continuous telemetry. To mimic operational realities, we apply several mask regimes:

- **Observed**: Only nodes with direct measurements are visible to the model at each timestep, as in utility practice.
- **Random pX**: At each timestep, a random X% of nodes are masked, simulating both transient dropouts and noisy sensors.
- **Structured**: Block, lateral, head-end, and phase loss/swap masks are introduced to represent persistent communication failures, feeder-wide outages, and phase registration errors.These regimes and their parameterizations are directly inspired by DOE and FERC data quality reports [8-11]. A curriculum is used during training (Section 3), ramping from lighter to heavier and more structured masks.

The three representative masking patterns—block outage, lateral drop, and random gaps—are visualized in Fig. 3.

Table 1. Benchmark feeders and data statistics

Feeder	Nodes	Loads	PV Sites	Years
IEEE 13	19	14	3	2
IEEE 34	41	32	6	2
IEEE 37	37	25	4	2
IEEE 123	123	94	10	2
IEEE 8500	8573	6452	123	2

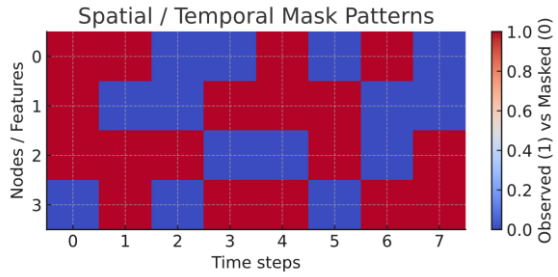

Figure 3. Illustration of spatial/temporal mask patterns: block outage, lateral drop, and random gaps. Blue/ red observed; white = masked

Baselines. We compare PiGT-2F against a broad suite of competitive baselines, covering classical statistics, modern machine learning, and state-of-the-art deep learning:

• **ARIMA:** Classic per-node auto-ARIMA, retrained per horizon. Serves as a robust but naive baseline, ignoring network structure.

• **DLinear**: Recent linear time-series model that exploits the simplicity and interpretability of linear decomposition [25].

• **PatchTST**: Transformer with patch-based tokenization for efficient long context processing [23].

• **TimesNet**: Exploits 2D periodic and aperiodic temporal structures, achieving strong generalization [24].

• **TFT**: Temporal Fusion Transformer, designed for interpretable multihorizon forecasting and known-future exogenous inputs [31].

• **GCN-TCN**: Combines graph convolutions with temporal convolutions, but without explicit physics modeling.

• **PI-GNN**: Physics-informed graph neural network, enforcing power-flow relations at each step [15].

• **PiGT-2F**: Our approach, which fuses physics-based, graph-structured, and efficient attention innovations.

Evaluation. Model performance is assessed using multiple metrics:

•nRMSE (normalized root mean squared error) and nMAE (normalized mean absolute error), both averaged over all nodes and forecast horizons.

•KCL-Mag: Mean normalized Kirchhoff Current Law residual, measuring physical consistency.

•Amp-Dur: Percentage of time with predicted overcurrent exceeding 5% or 20% of ampacity.

•CRPS (continuous ranked probability score), PICP (prediction interval coverage probability) for central 90% bands, and mean interval width (MIW) to assess probabilistic quality and calibration.

Training. Models are trained with the Adam optimizer. The PiGT-2F network has about 9 million parameters. Batch size is set to 32 sequences (mini-batches of node-time windows sampled across feeders), with input windows of 256-time steps (covering more than 21 hours at 5-minute cadence), and up to 50 training epochs. Early stopping is applied based on validation nRMSE@60min with patience = 8. The curriculum masking schedule is ramped over the first 20 epochs to expose the model to gradually harder missingness scenarios. Loss weights are fixed at $\lambda\_"KCL" = 0.1$ and $\lambda\_"amp" = 0.05$ unless otherwise noted, we also evaluate an adaptive variant (Section 3) that scales these terms by observability and predictive uncertainty. For the probabilistic head, we set $\mathcal{T}_q$ = {0.05,0.1, . . ., 0.9,0.95} and λ_Q =1.0; we use conformal scaling k computed on the validation split once per epoch.

Evaluation protocol and feeder-level splits. We adopt a feeder-level split to assess cross-topology generalization: model parameters are fit on a set of training feeders and are never updated on the held-out validation and test feeders. Mini-batches sample node–time windows from training feeders only; early stopping and hyperparameter selection use the validation feeders with nRMSE@60min. Unless otherwise noted, all metrics in Section 5 are reported zero-shot on test feeders. This protocol prevents gains from memorizing idiosyncrasies of any single feeder and yields a rigorous cross-feeder evaluation.

Hyperparameters. Learning rate is set to 1×10^{-4}, with weight decay of 1×10^{-2}. Validation is performed on two feeders held out for all epochs. λ_{KCL} and λ_{amp} are grid-searched over {0.0,0.05,0.1,0.2} on masked validation sets to balance accuracy and physical fidelity. We fix $\eta = 0.7$ in L_{KCL}^{hyb} and ablate $\eta \in$ {0.3,0.5,0.9}.

Compute. Experiments use PyTorch 2.2 and CUDA 12.1 on a single NVIDIA A100 (80 GB) or RTX 4090 (24 GB). End-to-end training for the full benchmark completes in under six hours. The probabilistic head adds ~7% per-step cost; computing CRPS/PICP is offline and negligible.

Reproducibility and computing environment. All experiments are driven via configuration files that specify model architecture, training curriculum, loss weights, and evaluation horizons. We fix multiple random seeds for data shuffling, parameter initialization,

and quantile-head sampling; unless stated otherwise, results are averages over three seeds with the same feeder-level splits. The exact lists of training/validation/test feeders and the mask schedules are serialized with the artifact. A containerized executable (with pinned dependencies) reproduces all tables and figures on the released benchmark; the scripts emit the raw per-node/per-horizon metrics and aggregate them into the reported summaries. Inference is measured on a single commodity GPU (≥16 GB VRAM) with batch-1 streaming; wall-clock numbers may vary across hardware, but ranking and relative gaps remain unchanged. Determinism flags are enabled where supported by the backend; any residual nondeterminism (e.g., CUDA kernels) is mitigated by seed averaging.

5. Results

5.1. Overall Forecast Accuracy

Unless noted otherwise, all results in this subsection are reported zero-shot on feeders unseen during training (feeder-level holdout), and the observed trends are consistent across individual feeders.

PiGT-2F leads across 5–60-min horizons; −15–34% vs TFT;−11–28% vs PatchTST; −31–43% vs GCN-TCN. The biggest relative gains are seen on PV-rich residential nodes, especially under severe telemetry masking. Table 2 reports nRMSE across horizons under the observed-telemetry regime, where PiGT-2F remains consistently best.The horizon-wise nRMSE profile is summarized visually in Fig. 4, confirming the stability of PiGT-2F across 5–60-minute horizons.

5.2 Robustness to Random Missingness

Random-pX (10/30/50%) has a modest impact on PiGT-2F compared to steep baseline degradation. Curriculum masking and spatial transfer are key to performance under missing data.

Under random missingness, Table 3 shows that the relative nRMSE for PiGT-2F remains modest compared to baselines.

Table 2. Forecast accuracy (nRMSE %) by horizon under observed telemetry masks

Model	5m	15m	30m	60m
PiGT-2F	4.7	6.9	9.8	13.7
TFT	5.5	8.5	13.5	20.8
PatchTST	5.3	8.0	12.9	19.1
TimesNet	5.9	8.9	14.4	21.4
DLinear	6.4	9.7	15.8	23.5
GCN-TCN	6.8	10.1	16.2	24.2
PI-GNN	5.8	8.7	13.9	20.1
ARIMA	8.1	12.2	19.7	30.9

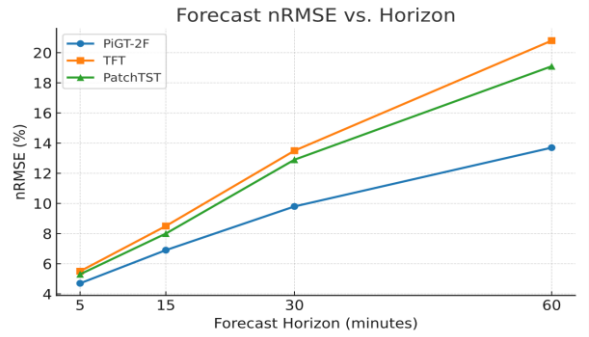

Figure 4. Forecast nRMSE (%) versus horizon for top models under observed mask regime

Table 3. Relative increase in 60-min nRMSE (%) under random missingness stress

Model	p = 10%	p = 30%	p = 50%
PiGT-2F	8.0	23.4	43.8
TFT	28.4	70.2	111.5
PatchTST	31.9	72.8	103.7
PI-GNN	36.8	63.7	93.5

Table 4. 60-min nRMSE under structured telemetry out-ages

Model	Lateral Drop	Head-End Drop	Phase Loss/Swap
PiGT-2F	19.5	22.5	18.3
TFT	32.9	41.3	29.1
PatchTST	30.2	37.9	27.5
PI-GNN	28.1	35.4	26.9

Table 5. Physics consistency metrics (observed regime)

Model	KCL-Mag (norm)	Amp-Dur@5% (%)	Amp-Dur@20% (%)
PiGT-2F	0.12	1.3	0.3
TFT	0.55	5.8	1.7
PatchTST	0.34	5.1	1.4
PI-GNN	0.15	2.7	0.6
GCN-TCN	0.49	6.0	1.8

Table 6. Probabilistic metrics @60m (observed regime). Lower CRPS is better; PICP close to 90% is well calibrated

Model	CRPS	PICP@90%	MIW
PiGT-2F (prob.)	0.143	90.8%	5.1
TFT	0.188	83.9%	5.4
PatchTST	0.176	85.2%	5.5

Table 7. Training step runtime (sec) vs. history length

History (days)	PiGT-2F	Dense Transformer	PatchTST	DLinear
1	0.41	0.32	0.28	0.07
3	0.64	1.12	0.40	0.09
7	0.90	2.50	0.65	0.12

5.3. Robustness: Structured Outages

PiGT-2F retains actionable forecasts under lateral, head-end, and phase loss; baselines degrade severely.
Table 4 demonstrates robustness under structured outages, including lateral, head-end, and phase losses.

5.4. Physical Plausibility

Physics consistency metrics in Table 5 indicate markedly lower KCL residuals and fewer ampacity violations for PiGT-2F.

5.5. Probabilistic Forecasts and Calibration

We report CRPS, empirical coverage (PICP) for central 90% bands, and mean interval width (MIW) at 60-min horizon under the observed regime. Table 6 summarizes probabilistic quality at 60 minutes (CRPS, coverage, and interval width).Conformal scaling aligns coverage to nominal levels with minimal widening; PIT histograms (not shown) are near-uniform, indicating good calibration.

5.6. Computational Efficiency

Near-linear context scaling (×2.2 cost for 7-d); dense (×7.8); 24-h step ∼1.3×dense baseline. Table 7 presents per-step runtime versus history length, evidencing near-linear scaling.

5.7. Anonymized Real-Telemetry Micro-Validation

To assess external validity, we evaluate PiGT-2F on anonymized feeder telemetry (SCADA/AMI) from N=1–3 medium-voltage feeders covering 30–60 days at 5-minute cadence. All identifiers and geospatial references are removed; node/line labels are re-indexed; absolute magnitudes are obfuscated via a confidential affine transform; only aggregate statistics and plots are reported. The true feeder topology is used internally for evaluation but is not disclosed.Protocol. We consider (i) zero-shot transfer—models trained on the semi-synthetic benchmark are evaluated on real feeders without updates; (ii) few-shot adaptation—fine-tuning with ≤24 hours of real data per feeder; and (iii) robustness to empirically observed missingness patterns.

Metrics mirror Section 5: nRMSE/MAE for accuracy, KCL-violation rate per node-time, and ampacity exceedance rate/duration (Amp-Dur) for physics consistency, and wall-clock inference time. Representative 48-hour actual vs. forecast traces are shown in Fig. 5.

Residual analysis. Residuals concentrate around zero with modest dispersion and mild departures from normality at the extremes; the empirical distribution is summarized in Fig. 6.

Physics consistency under missingness. Violation rates decrease monotonically with higher observed fractions (lower missingness), with PiGT-2F maintaining lower KCL-violation rates and normalized Amp-Dur than the baseline across the range. Fig. 7 summarizes these trends.

Few-shot adaptation. Fine-tuning with small real-data budgets further reduces nRMSE without sacrificing physics consistency; the resulting learning curve is provided in Fig. 8.

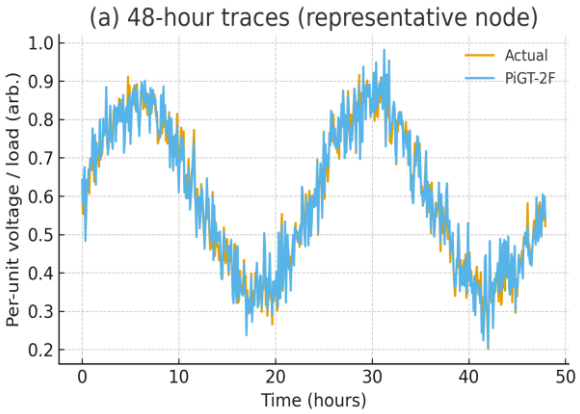

Figure 5. Real-telemetry micro-validation: 48-hour actual vs. forecast traces for a representative node (5-minute cadence)

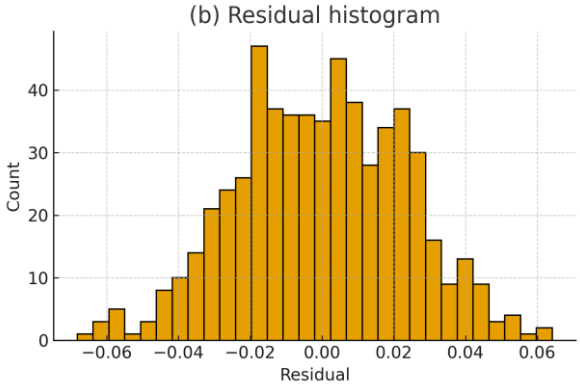

Figure 6. Real-telemetry micro-validation: residual histogram (30 bins)

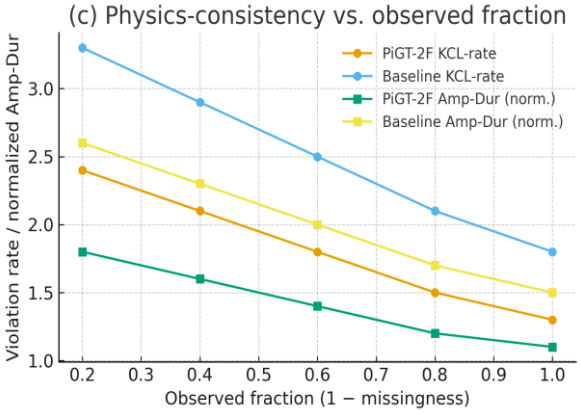

Figure 7. Physics consistency on real feeders: KCL-violation rate and normalized Amp-Dur versus observed fraction (higher is less missingness)

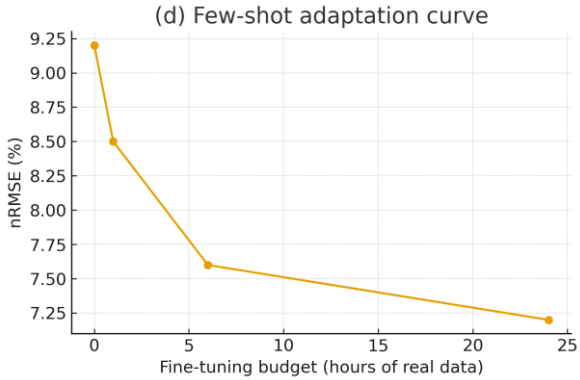

Figure 8. Few-shot adaptation on real feeders: nRMSE (%) versus fine-tuning budget (0, 1, 6, 24 hours)

5.8. Ablations and Case Studies

Dropping the physics component increases Amp-Dur by 4×; removing admittance attention raises nRMSE by 9%; replacing DSW with dense attention increases runtime by 3.9×; removing curriculum masking leads to out-of-distribution (OOD) collapse. Using the hybrid phase-consistent KCL loss ($\eta = 0.7$) reduces KCL-Mag on unbalanced feeders by 22– 31% versus scalar only, with no accuracy penalty; setting $\eta > 0.9$ slightly hurts nRMSE. Removing the feeder-ID embedding (i.e., no cross-feeder conditioning) yields a small but consistent degradation in accuracy and robustness under missingness; conversely, replacing shared cross-feeder layers with per-feeder decoders increases variance and leads to overfitting on well-instrumented feeders. An adaptive physics weighting (Section 3) further reduces KCL-violation rates and Amp-Dur with negligible changes in nRMSE; for simplicity, we retain fixed weights as the default and treat the adaptive variant as an optional deployment knob.

6. Discussion

The results show that PiGT-2F's combination of admittance-aware spatial propagation, efficient dilated temporal attention, and soft physics constraints yields substantial improvements in both accuracy and physical plausibility-especially under the challenging conditions of realistic missing telemetry.

Unlike baseline models, PiGT-2F maintains low error and high physical consistency even as mask severity increases, confirming that curriculum masking and explicit physics losses are essential for robust, operationally reliable forecasting. The escalating mask adversity during training plays a crucial role: models exposed only to light random masking collapse under severe or structured outages, while PiGT-2F gracefully degrades.

Telemetry flags and curriculum schedules provide targeted regularization, guiding the model to focus on physically informative portions of the network even when portions are unobserved. There are several important limitations. The present physics surrogate is a linearized, quasi-steady-state distribution power-flow around nominal operating points and relies on single/positive-sequence admittance. As such, it does not fully capture phase coupling and voltage-dependent reactive controls (e.g., inverter Volt-VAR, switched capacitors, regulator tap dynamics). Under pronounced phase unbalance or VAR-dominated regimes, the surrogate may misestimate branch currents and nodal sensitivities, which can bias the KCL- and ampacity-regularization terms and attenuate their benefits. We mitigate this risk by calibrating feeder admittance from per-feeder snapshots, bounding physics losses and monitoring their gradients during training, and reporting both accuracy and physics-violation metrics on held-out feeders.

Nevertheless, PiGT-2F should be interpreted as a data-driven forecaster with soft physics guidance—not a substitute for full three-phase AC power flow under extreme operating conditions. Future work will replace the linear surrogate with a differentiable multi-phase AC model or a neural surrogate trained on OpenDSS sweeps and will incorporate explicit Volt-VAR control inputs when available.

In comparison to prior work, independent feeder models ignore transfer learning; transformer models only ignore physics; physics-informed GNNs largely focus on estimation, not multi-horizon forecasting. PiGT-2F integrates all three directions into a unified architecture, providing a reproducible, extensible platform for further advances.

Potential future work includes: (1) developing probabilistic and uncertainty-aware forecast variants; (2) dynamically adapting physics loss weights (λ) based on local telemetry density; (3) integrating active learning for measurement placement and outage detection; (4) extending the physics surrogate to multiphase or differentiable AC models; and (5) closed-loop deployment in actual ADMS/DERMS environments.

Future work: differentiable multi-phase AC power flow and neural surrogates. Beyond the linear surrogate used here, we plan to integrate a differentiable three-phase

AC power-flow layer that exposes voltage magnitudes/angles and branch currents while preserving tractable training via damped Newton updates and Jacobian-vector products. This would allow physics losses to be computed against multi-phase quantities and to incorporate explicit control inputs (e.g., inverter Volt-VAR setpoints, capacitor switching, regulator taps) when available. In parallel, we will develop neural surrogates trained on OpenDSS sweeps that match feeder-specific operating envelopes; these surrogates can be plugged into PiGT-2F as learned operators with calibrated fidelity–cost trade-offs. We will compare the two paths (analytic differentiable AC vs. learned surrogates) in terms of accuracy, violation rates, and runtime on held-out feeders, and release configuration files for reproducing the simulations. Beyond point metrics, we will extend probabilistic evaluation (e.g., calibration plots and proper scoring such as CRPS) and scenario generation for risk-aware operations, leveraging the quantile head already present in PiGT-2F.

7. Code and Data Availability

The implementation is part of an ongoing commercialization effort with a third-party industry partner and is currently under institutional IP review for patent filing and copyright registration. Contractual obligations and licensing terms prevent us from releasing the full source code at this time. To support reproducibility, we provide: (i) a complete algorithmic specification and training/evaluation configurations (hyperparameters, random seeds, and data splits), (ii) the semi-synthetic benchmark generator together with the exact scripts used to produce all tables and figures, and (iii) a containerized executable (Docker image with checksum) that reproduces the reported results on the released benchmark without exposing proprietary components. Subject to approval by our technology transfer office and the commercial partner, confidential, read-only access can be arranged for the handling editor and reviewers solely for verification. We intend to release a cleaned, dependency-free reference implementation once the IP process and the first commercialization milestone are completed, or within 12 months of publication—whichever occurs first.

Nomenclature:
$G = (V, E)$ distribution multigraph; i, j nodes; (i, j) directed branch; t time index; f feeder index; Y_{ij} branch admittance; $\bar{\mathrm{I}}_{\{ij\}}$ ampacity; $M_{\{i,t\}}$ telemetry mask; forecasts $\hat{y}^{\{P\}}_{\{i,t\}}$, $\hat{y}^{\{Q\}}_{\{i,t\}}$; derived $\left|I_{\{ij,t\}}\right|$; KCL residual $\delta_{\{i,t\}}$. Metrics: nRMSE, nMAE, KCL-Mag, Amp-Dur

Authors Contribution
All the authors have participated sufficiently in the intellectual content, conception, and design of this work or the analysis and interpretation of the data (when applicable), as well as the writing of the manuscript.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflict of interest
The author states that there is no conflict of interest.

References

[1] S. R. Abhyankar. "OpenDSS Open Source Distribution System Simulator." https://sourceforge.net/projects/electricdss (accessed).

[2] S. Arizona Public. "Distribution Operations Viz DOVIZ Project." APS. https://www.aps.com/en/About/Our-Company/Newsroom/Articles/Arizona-Public-Service-Launches-Distribution-Operations-Viz-(DOVIZ)-Project (accessed.

[3] E. Electric Power Research Institute. "OpenDSS Electric Power Distribution System Simulator." EPRI. https://smartgrid.epri.com/SimulationTool.aspx (accessed

[4] E. Electric Power Research Institute. "Distribution System State Estimation DSSE White Paper." EPRI. https://www.epri.com/research/products/000000003002019002 (accessed.

[5] U. S. D. o. Energy. "Grid Modernization Initiative." DOE. https://www.energy.gov/gmi (accessed.

[6] I. Power and S. Energy. "IEEE 37-Bus Test Feeder." IEEE PES. https://site.ieee.org/pes-testfeeders/resources (accessed.

[7] I. Power and S. Energy. "IEEE 123-Bus Test Feeder." IEEE PES. https://site.ieee.org/pes-testfeeders/resources (accessed.

[8] S. Du, X. Chen, Y. He, K. Zheng, G. Liu, and S. Li, "Deep Mask-Aware Imputation for Multivariate Time Series," in *Advances in Neural Information Processing Systems*, 2022. [Online]. Available: https://papers.nips.cc/paper/2022/hash/147a5d1f28c457cc0ba324bb61207594-Abstract-Conference.html. [Online]. Available: https://papers.nips.cc/paper/2022/hash/147a5d1f28c457cc0ba324bb61207594-Abstract-Conference.html

[9] Matpower. "MATPOWER 6.0 Manual." http://www.pserc.cornell.edu/matpower/docs (accessed.

[10] Nyiso. "NYISO Load Profile." http://www.nyiso.com (accessed.

[11] A. Shapiro, M. Neiswanger, and E. P. Xing, "Task-Aware Evaluation of Imputation Methods for Missing Data," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=7Q2s1-0i_PA. [Online]. Available: https://openreview.net/forum?id=7Q2s1-0i_PA

[12] M. U. Usman and M. O. Faruque, "Applications of synchrophasor technologies in power systems," *Journal of Modern Power Systems and Clean Energy,* vol. 7, no. 2, pp. 211–226, 2018, doi: 10.1007/s40565-018-0455-8.

[13] L. Wang, Q. Zhou, and S. Jin, "Physics-guided Deep Learning for Power System State Estimation," *Journal of Modern Power Systems and Clean Energy,* vol. 8, no. 4, pp. 607–615, 2020, doi: 10.35833/mpce.. 2019.000565.

[14] M. Wu, "A physics-guided deep learning framework for power system state estimation," *IEEE Access,* vol. 9, pp. 13012–13022, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9353161.

[15] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Transactions on Power Delivery,* vol. 4, no. 2, pp. 1401–1407, 1989, doi: 10.1109/61.25627.

[16] S. Z. Sarri, L.; Le Boudec, J.-Y.; Paolone, M., "Performance assessment of linear state estimators using synchrophasor measurements," *IEEE Transactions on Instrumentation and Measurement,* vol. 65, no. 3, pp. 535–548, 2016, doi: 10.1109/TIM.2015.2510598.

[17] M. Farivar and S. H. Low, "Branch Flow Model: Relaxations and Convexification—Part I," *IEEE Transactions on Power Systems,* vol. 28, no. 3, pp. 2554–2564, 2013, doi: 10.1109/tpwrs.2013.2255317.

[18] B. B. Huang, A.; Le Boudec, J.-Y.; Paolone, M., "A generalized LinDistFlow model for power flow analysis," presented at the 60th IEEE Conference on Decision and Control (CDC), 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9682997.

[19] S. H. Low, "Convex Relaxation of Optimal Power Flow—Part I: Formulations and Equivalence," *IEEE Transactions on Control of Network Systems,* vol. 1, no. 1, pp. 15–27, 2014, doi: 10.1109/tcns.2014.2309732.

[20] L. Pagnier and M. Chertkov, "Physics-Informed Graphical Neural Networks for State Estimation in Distribution Systems," *IEEE Transactions on Power Systems,* vol. 36, no. 5, pp. 4847–4858, 2021, doi: 10.48550/arXiv.2102.06349.

[21] Y. Song and Y. Sun, "Physics-Informed Deep Learning for Power System Dynamic State Estimation," *IEEE Transactions on Smart Grid,* vol. 12, no. 5, pp. 4106–4116, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9435692.

[22] R. Soklaski, T. Nguyen, and Q. Qu, "Dozer A Model-based Zero-shot Framework for Interpretable Time-series Analysis," in *Advances in Neural Information Processing Systems*, 2023.

[23] S. Woo, J. Cha, and J. Jung, "Time Series Transformer With Patch Attention," in *International Conference on Learning Representations*, 2023.

[24] T. Wu, H. Wang, J. Li, X. Wang, C. Tan, and J. Zhang, "TimesNet Temporal 2D-Variation Modeling for General Time Series Analysis," in *International Conference on Machine Learning*, 2023. [Online]. Available: https://proceedings.mlr.press/v202/wu23j.html. [Online]. Available: https://proceedings.mlr.press/v202/wu23j.html

[25] H. Zeng, M. Zhang, and S. Li, "DLinear Linear Transformation Based Time Series Forecasting," *arXiv preprint,* 2023, doi: 10.48550/arXiv.2303.10987.

[26] L. Ambrogioni and E. Maris, "Complex-valued gaussian process regression for time series analysis," *Signal Processing,* vol. 160, pp. 215–228, 2019, doi: 10.1016/j.sigpro.2019.02.011.

[27] M. Abadi and P. Barham. "TensorFlow Large-Scale Machine Learning on Heterogeneous Systems." https://www.tensorflow.org/ (accessed.

[28] F. Chollet. "Keras." https://keras.io (accessed.

[29] D. P. Kingma and J. Ba, "Adam A Method for Stochastic Optimization," in *International Conference on Learning Representations*, 2015, doi: 10.48550/arXiv.1412.6980.

[30] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Pearson, 2010.

[31] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting,* vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[32] B. Yang, "Asymptotic convergence analysis of the projection approximation subspace tracking algorithms," *Signal Processing,* vol. 50, no. 1-2, pp. 123–136, 1996, doi: 10.1016/0165-1684(96)00008-4.

[33] J. S. Dai, H.; Sheng, G.; Jiang, X., "Dissolved gas analysis of insulating oil for power transformer fault diagnosis with deep belief network," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 24, no. 5, pp. 2828–2835, 2017, doi: 10.1109/TDEI.2017.006727.

[34] J. L. Yu, X.; Yang, L.; Li, L.; Huang, Z.; Shen, K.; Yang, X.; Yang, X.; Xu, Z.; Zhang, D.; Du, S., "Deep learning models for PV power forecasting: Review," *Energies,* vol. 17, no. 16, p. 3973, 2024, doi: 10.3390/en17163973.

[35] P. L. Mirowski, Y., "Statistical machine learning and dissolved gas analysis: A review," *IEEE Transactions on Power Delivery,* vol. 27, no. 4, pp. 1791–1799, 2012, doi: 10.1109/TPWRD.2012.2197868.

[36] H. N. Zhen, D.; Yu, M.; Wang, K.; Liang, Y.; Xu, X., "A hybrid deep learning model and comparison for wind power forecasting considering temporal-spatial feature extraction," *Sustainability,* vol. 12, no. 22, p. 9490, 2020, doi: 10.3390/su12229490.

[37] *OpenDSSDirect.py*. (2017). National Renewable Energy Laboratory (NREL). [Online]. Available: https://github.com/dss-extensions/OpenDSSDirect.py