

When Does Context Help Forecasting? An Empirical Study of Energy Systems

Xu Zheng¹, Jingchao Ni², Mo Sha¹, Dongsheng Luo³

¹Florida International University, Miami, FL, USA

²University of Houston, Houston, TX, USA

³Singapore Management University, Singapore

{xzhen019,msha}@fiu.edu, jni7@uh.edu, luodongsheng01@gmail.com

Abstract—Devices with similar power histories can have different remaining demand. Operational records may reveal this difference when linked to the right devices. Across four public collections covering water heaters, EV charging, and air conditioning, linear, tree, and neural models compare power history with richer context on matched forecasts. Storage records improve tree and neural water-heater forecasts, and charging requests improve group forecasts beyond measured telemetry. Indoor measurements yield small gains for the linear model. We then reassign records among devices in the same group and measure the resulting error before and after retraining. Retraining recovers much of the initial loss on water heaters and one charging site, with little recovery at the other site. These results distinguish the benefit of adding context from dependence on device associations, connecting input selection to record maintenance. Source adapters, evaluation code, and the reported aggregate results are provided. The code and data are available at <https://github.com/flyingdoog/energy-context-benchmark>.

Index Terms—time-series forecasting, energy systems, operational data, electric vehicles, smart buildings, model evaluation

I. INTRODUCTION

Two electric vehicles can draw the same power while having different remaining energy requirements. Two idle water heaters can resume operation at different times because their stored energy differs. Power history captures what a device has consumed; operational records can describe what it still needs. Which records provide useful forecasting information?

Using context involves choosing both fields and their device associations. A driver’s energy request belongs to a charging session; a temperature sensor may serve several air conditioners. We study these choices through device forecasts and the combined demand of devices present at the query, such as the vehicles currently connected at a charging site.

Existing resources support several views of energy demand. UrbanEV couples charging demand with spatial and external information [1]. The Smarter Mobility Data Challenge evaluates occupancy at station, area, and overall levels [2], while DeDemand supports device-level forecasting with auxiliary information [3]. Our comparison asks how much a field improves forecasts for the same devices and how that benefit depends on keeping its device association.

Across four public collections, we first add states, controls, and user declarations to power-history inputs for linear, tree, and neural models. We then reassign operational records among devices in a query group, preserving their power histories and targets. Both the original predictor and a predictor retrained under that reassignment are evaluated. Retraining separates the immediate error increase from the gap remaining after adaptation, connecting linkage quality to the model update available to an operator.

Stored state improves nonlinear heater forecasts, and charging declarations improve group predictions beyond telemetry. Indoor measurements have smaller, model-dependent benefits. After record reassignment, water heaters and Caltech charging recover much of their initial loss through retraining; Office charging shows little recovery. A deployed model’s sensitivity and its performance after an update can therefore give different assessments of the same linkage.

We provide matched context comparisons and a record-association experiment through a common task interface retaining device identity, group membership, and source-specific forecast intervals. Code, results, and manuscript sources are available online.

II. RELATED WORK

A. Energy data and forecasting tasks

ACN-Data links charging sessions to power, charging controls, and session information [4]. EPFL’s fast-charging release records two charging plugs and heterogeneous vehicle sessions [5], while CU-BEMS couples circuit-level power to indoor environmental measurements [6]. BPA’s CTA-2045 demonstration provides seasonal water-heater reports and control records [7]. These sources distinguish stored state, charging intent, and shared environmental conditions. Our input views retain their query-time roles: requested energy describes intended demand, energy delivered describes current state, and realized departure is a future outcome.

B. Models and record reliance

We use a regularized linear reference, LightGBM [8], and N-HITS [9], with fixed training settings within each family, to compare additive, tree-based, and neural responses to context. Lightweight linear forecasters can remain competitive on multivariate series [10], which motivates keeping a linear reference.

Permutation-based analysis evaluates how a fitted predictor relies on its inputs. Fisher *et al.* distinguish reliance of one model from reliance across well-performing models [11]; conditional approaches account for dependencies between retained and perturbed features [12]. Related time-series work learns explanations for forecasters [13] and augmentations that preserve informative temporal structure [14]. Our intervention moves whole auxiliary record blocks, including missingness masks, within the current device group. Evaluation-only reassignment and retraining then measure fitted-model dependence and adaptation under the same intervention.

III. FORECASTING FROM OPERATIONAL RECORDS

A. Task and information available at a query

At time q , let G_q be the observed device population. For device i , let x_{iq} contain its power history and calendar features, and z_{iq} contain its available operational records and observation masks. A model returns future mean powers

$$\hat{y}_{iq} = f(x_{iq}, z_{iq}), \quad \hat{Y}_{qj} = \sum_{i \in G_q} \hat{y}_{iqj}. \quad (1)$$

The second expression is the group-total forecast for bin j . Its target uses the same membership G_q . Group-mean power divides both sides by $|G_q|$. Device and

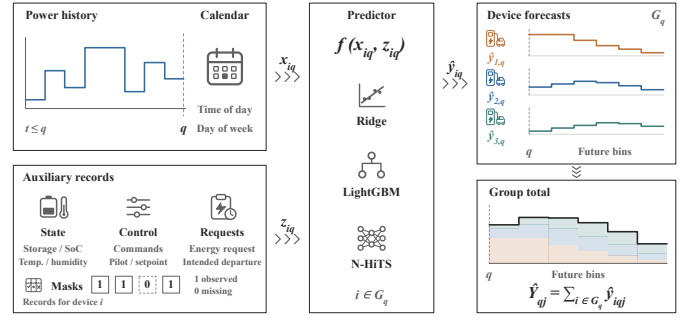


Fig. 1. Context-aware forecasting for the devices observed at a query. Information conditions change the auxiliary inputs, while the power history, device population, and future scoring intervals remain fixed.

group errors answer different questions because errors can cancel or reinforce when summed.

We compare three primary information conditions: *history*, *history plus state*, and *all telemetry*, the last additionally including recorded controls. History includes the calendar. At Caltech, a separate comparison adds query-available charging declarations to all telemetry. For CU-BEMS, the state fields are observed temperature and humidity; state and all telemetry coincide because control commands are not supplied. Fig. 1 shows the common task.

Inputs contain information available by the query, with masks for missing records. After a confirmed charging-session end, subsequent session power is zero; unobserved tails remain missing. Charging arrivals after the query are outside G_q . Group scores require supported targets for every member of G_q throughout the scored interval.

B. Four source collections

Table I summarizes the prepared evaluation populations. BPA queries occur thirty minutes after a changed-command report, retaining recent power and aligned previous-day/week profiles. Its synchronous groups share the report time and command and contain at least ten label-eligible devices. The target is the provider’s reported power, which includes estimated and calibrated values. Each group comprises the eligible field-study devices at that report time.

ACN uses hourly queries for sessions already present at Caltech or Office. Eligible groups contain at least two distinct charging stations and require complete labels for their query-known cohort. The sites have different information coverage: query-time declarations are observed at Caltech but not in the selected Office records. EPFL contributes 1,771 eligible fast-charging sessions on two

TABLE I

SOURCES AND EVALUATION PERIODS. CASES COUNT DEVICE-QUERY PAIRS; EPFL USES ONE QUERY PER SESSION. ACN INCLUDES CALTECH AND OFFICE. CU-BEMS COUNTS COVER ALL SUPPORTED FLOORS; TABLE II USES THE FIVE-FLOOR THERMAL COHORT DESCRIBED BELOW.

Source / setting	Context beyond power	History	Later period	Cases	Dates
BPA heaters	Storage, flags, commands	24 h + profiles	Spring 2018	64,451	63
			Summer 2018	47,237	69
ACN Caltech	State, pilot, declarations	60 min	October 2019	4,858	31
			January 2020	3,983	30
ACN Office	State, pilot	60 min	October 2019	749	23
			January 2020	796	22
EPFL charging	SoC, energy, capacity, setpoint	First 10 min	March–July 2023	739	100
CU-BEMS ACs	Zone temperature, humidity	60 min	Sep.–Oct. 2019	154,316	61
			Nov.–Dec. 2019	135,839	61

plugs. Each query uses the first ten minutes to predict actual delivered power; EPFL is evaluated at device level only. Capacity estimates inferred from complete sessions and realized departures are not inputs.

CU-BEMS contains 55 air-conditioning circuits in one Bangkok building, 51 with linked thermal sensor fields. The matched two-period comparison uses 44 circuits on Floors 2–5 and 7 and intersects dates across the compared models. Floor 1 has no thermal fields; Floor 6’s published file ends in October, leaving five floors with thermal coverage in both periods. Several circuits can share a zone sensor, so the thermal fields describe their common environment rather than an individual device’s stored energy.

C. Splits, intervals, and error measures

Splits follow time. BPA trains on 46,807 winter cases and validates on the last fourteen observed winter dates, with targets crossing the boundary excluded. ACN trains in April–June 2019 and validates in July. EPFL uses 2022 for training, January–February 2023 for validation, and March–July for evaluation, yielding 946, 86, and 739 sessions. CU-BEMS trains in April–June 2019 and validates in July–August. Table I lists the held-out evaluation periods. Model settings are specified in Section IV-A and kept fixed across information conditions.

BPA, ACN, and CU-BEMS score cumulative spans of 30, 90, and 150 minutes using fifteen-minute target bins. EPFL scores 5, 15, and 30 minutes using five-minute bins. Source timestamps and the distinction between point reports and interval measurements are retained by the adapters. For a span of J bins, device MSE is the

average of $(\hat{y}_{iqj} - y_{iqj})^2$ over its bins and eligible cases; MAE uses absolute errors. Cases are averaged within each date, and dates then receive equal weight. Span averages weight the three cumulative spans equally.

Power MSE and MAE are reported in kW^2 and kW , respectively. Group scores use the same definitions after aggregation in (1). CU-BEMS additionally weights its common floors equally. Errors are reported separately for each source in its physical units. The quoted 95% intervals use 1,500 paired calendar-week bootstrap draws, conditional on the fitted predictors.

IV. EXPERIMENTS

A. Models and experimental setup

Ridge uses an intercept, training-derived standardization, and a unit L2 penalty. LightGBM uses 100 trees, 31 leaves, and squared-error fitting for each target bin. N-HITS uses historical and current exogenous inputs, known calendar features, and 1,000 squared-error training updates. Neural results average the errors of three independently fitted seeds. Each charging site or building floor is fitted separately.

We use the stated settings without hyperparameter search. Information conditions share the same targets and eligible cases. The neural Caltech declaration comparison keeps its input width fixed by replacing the added fields with zeros in the telemetry-only control; CU-BEMS likewise zeroes thermal slots in the history-only arm. Thus both neural comparisons vary the supplied information at a fixed input width.

TABLE II

DEVICE FORECASTING WITH DIFFERENT INFORMATION CONDITIONS. MSE (kW^2) AND MAE (kW) AVERAGE THE THREE NATIVE SPANS WITH EQUAL DATE WEIGHTS. CU-BEMS ADDITIONALLY WEIGHTS ITS FIVE COMMON FLOORS EQUALLY; ITS STATE AND ALL-TELEMETRY COLUMNS ARE THE SAME THERMAL ARM. NEURAL ERRORS AVERAGE THREE SEEDS. BOLD MARKS THE LOWEST ERROR ACROSS INFORMATION CONDITIONS WITHIN EACH MODEL AND ROW, INCLUDING TIES AT DISPLAYED PRECISION.

Source / period	Model	History		History + state		All telemetry	
		MSE	MAE	MSE	MAE	MSE	MAE
BPA Spring	Ridge	0.290	0.218	0.314	0.246	0.335	0.273
	LightGBM	0.282	0.192	0.254	0.184	0.253	0.183
	N-HiTS	0.322	0.208	0.297	0.191	0.298	0.189
BPA Summer	Ridge	0.239	0.198	0.256	0.223	0.334	0.290
	LightGBM	0.228	0.176	0.197	0.166	0.196	0.167
	N-HiTS	0.274	0.194	0.237	0.174	0.251	0.185
Caltech Autumn	Ridge	1.183	0.484	1.200	0.498	1.200	0.502
	LightGBM	1.116	0.425	1.103	0.419	1.093	0.415
	N-HiTS	1.202	0.455	1.207	0.440	1.208	0.437
Caltech Winter	Ridge	1.001	0.431	1.043	0.463	1.064	0.473
	LightGBM	1.044	0.385	1.011	0.378	1.000	0.372
	N-HiTS	1.075	0.406	1.080	0.396	1.081	0.392
Office Autumn	Ridge	2.221	0.856	2.588	0.946	2.788	0.997
	LightGBM	2.819	0.893	2.689	0.886	2.665	0.889
	N-HiTS	3.192	0.853	3.169	0.859	3.367	0.921
Office Winter	Ridge	2.623	0.937	2.916	0.992	2.965	1.023
	LightGBM	3.231	1.042	2.923	0.995	2.942	1.004
	N-HiTS	3.372	0.933	3.316	0.941	3.525	1.002
EPFL Later	Ridge	583.010	15.910	529.232	14.616	561.019	15.131
	LightGBM	675.160	17.164	620.133	16.382	614.475	15.959
	N-HiTS	740.517	17.457	798.949	18.326	874.857	18.772
CU-BEMS Sep.–Oct.	Ridge	13.784	1.491	13.630	1.460	13.630	1.460
	LightGBM	3.015	0.427	3.317	0.473	3.317	0.473
	N-HiTS	5.833	0.657	6.420	0.690	6.420	0.690
CU-BEMS Nov.–Dec.	Ridge	12.341	1.429	12.235	1.402	12.235	1.402
	LightGBM	4.111	0.488	4.658	0.558	4.658	0.558
	N-HiTS	6.871	0.698	7.662	0.763	7.662	0.763

B. Forecasting gains from operational context

Table II compares history, state, and control inputs; Table III adds user requests to measured telemetry.

Stored state helps nonlinear heater forecasts. On BPA, adding state lowers LightGBM MSE from 0.282 to 0.254 in spring and from 0.228 to 0.197 in summer, with lower MAE in both periods. N-HiTS follows the same pattern: 0.322 to 0.297 and 0.274 to 0.237. Adding controls beyond state changes the tree results only slightly. Ridge has its lowest error with history alone.

Charging telemetry and user intent play different roles. State and control records improve the tree refer-

ence at both ACN sites. At Caltech, measured telemetry leaves neural MSE close to the history-only result. Requested energy and intended departure describe demand beyond what is observed in the power trace. Table III adds these declarations to the telemetry inputs and compares forecasts on the same vehicle groups. Adding declarations reduces span-average group-mean MSE by approximately 19% and 16% for LightGBM in autumn and winter, and by 25% and 31% for N-HiTS. MAE also decreases in all four comparisons. Both comparisons include sessions with missing declarations through explicit observation masks.

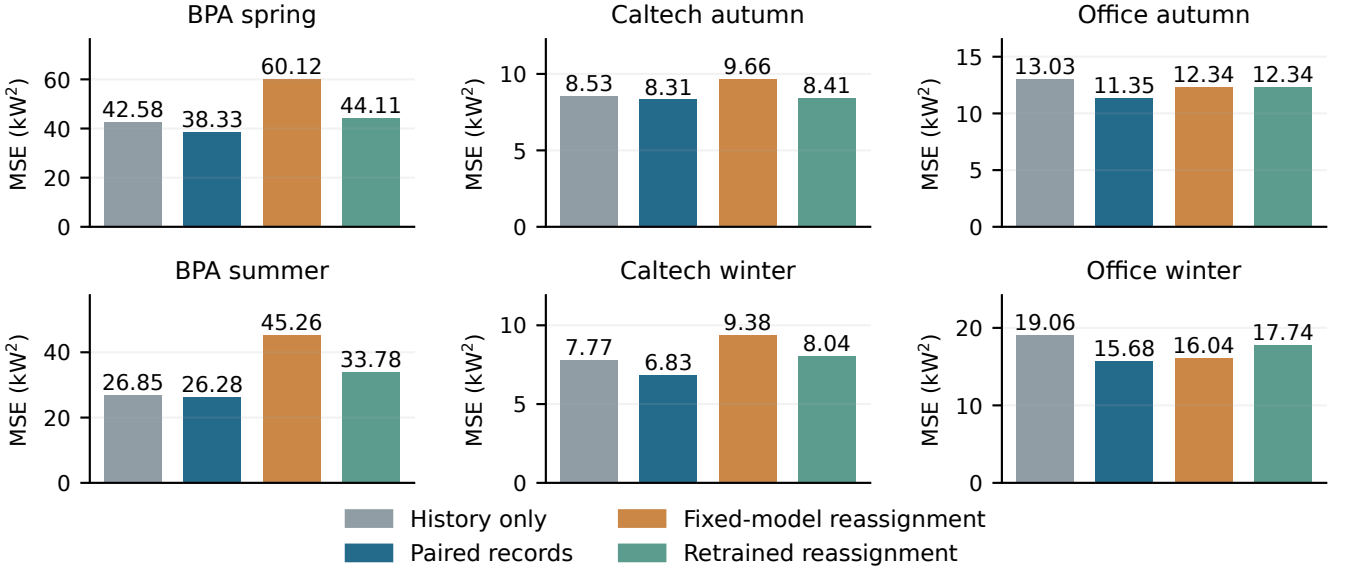


Fig. 2. Group-total MSE for four record-association arms, averaged over native spans. Reassigned arms share three within-query mappings. Vertical scales vary by panel. Retraining recovers much of the fixed-model loss in BPA and Caltech, but not Office.

TABLE III
CALTECH GROUP-MEAN FORECASTING WITH QUERY-TIME
DEMAND DECLARATIONS. MSE / MAE IN kW^2 / kW ;
EQUAL-SPAN AVERAGES. THE NEURAL CONTROL HAS THE SAME
INPUT WIDTH AS THE DECLARATION ARM.

Model	Period	Telemetry MSE / MAE	+ declarations MSE / MAE
LightGBM	Autumn	0.258 / 0.278	0.209 / 0.251
	Winter	0.256 / 0.290	0.215 / 0.248
N-HITS	Autumn	0.307 / 0.301	0.229 / 0.261
	Winter	0.325 / 0.318	0.225 / 0.258

State and environmental measurements favor different models. In EPFL, adding state lowers Ridge MSE from 583.010 to 529.232 and LightGBM MSE from 675.160 to 620.133, with corresponding MAE improvements. N-HITS has its lowest error with history alone. In CU-BEMS, adding thermal measurements lowers Ridge MSE from 13.784 to 13.630 and from 12.341 to 12.235 across the two periods. The history-only tree remains the lowest-error model in both. Unlike individual heater or charging states, these building records describe shared zone conditions. Their different effects favor selecting a model and its input fields together.

C. The value of device-record associations

On BPA and the two ACN sites, we reassign records and measure error before and after retraining. Each

device receives an auxiliary record from a within-query permutation π_q . Power history, calendar, previous-day/week power profiles where available, and targets stay with the receiving device. Auxiliary values and their masks move together. Thus the group retains the same collection of operational records, while individual associations change.

Four arms use the same LightGBM settings: history only, paired records, a fitted paired predictor evaluated with reassigned records, and a predictor retrained with reassigned records. Let their MSEs be R_H , R_P , R_F , and R_R . We report

$$D_F = R_F - R_P, \quad D_R = R_R - R_P. \quad (2)$$

The first measures the fitted predictor’s sensitivity; the second measures the remaining gap after retraining. Their difference $R_F - R_R$ is the recovered error. Three mapping seeds are used in both reassigned arms, averaging errors across mappings. Mappings are generated from query groups and random seeds. The experiment uses the ordinary telemetry views, without the extra Caltech declarations.

Fig. 2 compares the four arms in each site and period. For BPA summer, group-total MSE rises from 26.277 with paired records to 45.256 when only evaluation records are reassigned. Retraining reduces it to 33.780. Relative to the paired error, the fixed-model penalty is 72.23% (95% interval 53.74–94.08%), whereas the remaining retrained penalty is 28.55% (23.86–35.08%).

Spring exhibits the same ordering. The paired model remains the best of these arms, while retraining recovers a large part of the initial loss.

Caltech autumn illustrates an even stronger separation. Paired, fixed, and retrained MSE are 8.313, 9.663, and 8.408. The fixed penalty is 16.24% (1.34–32.26%), while the retrained penalty is 1.15% (–5.86–9.69%). In winter, retraining again improves on the fixed arm but leaves a larger gap: 6.827, 9.376, and 8.040, respectively.

Office supplies a different result. In autumn, fixed and retrained MSE are almost identical, 12.343 and 12.341; in winter, retraining increases the error from 16.044 to 17.738. Both later periods still favor the paired reference over history only. The original associations retain their forecasting benefit after retraining.

V. DISCUSSION

A. Choosing inputs and maintaining records

Input selection should reflect what a record describes. Stored energy, delivered energy, controls, and requests represent reserves, past consumption, operating limits, and intended demand, respectively. Comparing these roles identifies which fields merit integration for a given predictor.

Fixed-model and retrained errors separate immediate linkage sensitivity from the loss persisting after a model update. Paired records define the performance to recover.

B. Study scope and resources

The collections cover device families under fixed model settings. Charging forecasts concern devices present at the query; total future site demand additionally requires an arrival model. Synthetic reassignment changes records' compatibility with individual histories. Its effects and retraining recovery concern that intervention; natural linkage failures can differ in structure and frequency.

The repository supplies source adapters, forecasting/scoring interfaces, reference code, and aggregate results. Source guides describe local data acquisition and preparation under provider terms.

VI. CONCLUSION

Operational records contribute through both their contents and their device associations. Stored state improves nonlinear heater forecasts, and charging requests improve group predictions beyond telemetry. Record reassignment shows that immediate error can greatly exceed the gap after retraining. Input selection and linkage assessment should therefore be evaluated together for the devices and groups being forecast.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under grants ECCS-2242700 and IIS-2529283.

REFERENCES

- [1] H. Li, H. Qu, X. Tan, L. You, R. Zhu, and W. Fan, "UrbanEV: An open benchmark dataset for urban electric vehicle charging demand prediction," *Scientific Data*, vol. 12, no. 1, p. 523, 2025.
- [2] Y. Amara-Ouali, Y. Goude, N. Doumèche, P. Veyret, A. Thomas, D. Hebenstreit, T. Wedenig, A. Satouf, A. Jan, Y. Deleuze, P. Berhaut, and S. Treguer, "Forecasting electric vehicle charging station occupancy: Smarter mobility data challenge," *Journal of Data-centric Machine Learning Research*, vol. 1, no. 16, pp. 1–27, 2024.
- [3] B. Neupane, L. Šikšnys, and T. B. Pedersen, "DeMand: A tool for evaluating and comparing device-level demand and supply forecast models," in *the Workshop Proceedings of the EDBT/ICDT 2016 Joint Conference*, 2016.
- [4] Z. J. Lee, T. Li, and S. H. Low, "ACN-Data: Analysis and applications of an open EV charging dataset," in *ACM International Conference on Future Energy Systems*, 2019.
- [5] Distributed Electrical Systems Laboratory, EPFL, "Level 3 electric vehicle charging dataset," 2025.
- [6] M. Pipattanasomporn, G. Chitalia, J. Songsiri, C. Aswakul, W. Pora, S. Suwankawin, K. Audomvongseree, and N. Hoonchareon, "CU-BEMS, smart building electricity consumption and indoor environmental sensor datasets," *Scientific Data*, vol. 7, p. 241, 2020.
- [7] Bonneville Power Administration, "CTA-2045 water heater demonstration project," Project page and seasonal data downloads, n.d.
- [8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017.
- [9] C. Challu, K. G. Olivares, B. N. Oreshkin, F. Garza Ramirez, M. Mergenthaler Canseco, and A. Dubrawski, "NHITS: Neural hierarchical interpolation for time series forecasting," in *AAAI Conference on Artificial Intelligence*, 2023.
- [10] A. Ma, D. Luo, and M. Sha, "MixLinear: Extreme low resource multivariate time series forecasting with 0.1k parameters," in *International Conference on Learning Representations*, 2026.
- [11] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019.
- [12] C. Molnar, G. König, B. Bischl, and G. Casalicchio, "Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach," *Data Mining and Knowledge Discovery*, vol. 38, pp. 2903–2941, 2024.
- [13] X. Zheng, W. Cheng, Z. Chen, M. Sha, J. Ni, and D. Luo, "Information bottleneck learning for faithful time series forecasting explanations," arXiv preprint arXiv:2607.28124, 2026.
- [14] X. Zheng, T. Wang, W. Cheng, A. Ma, H. Chen, M. Sha, and D. Luo, "Parametric augmentation for time series contrastive learning," in *International Conference on Learning Representations*, 2024.