RETHINKING LONG-TERM TIME SERIES FORECASTING AS PATCH-TO-HORIZON CONTRIBUTION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term time series forecasting (LTSF) is commonly formulated as learning global representations from historical observations, where the contributions of individual historical patches to different forecasting horizons remain implicit. This representation-centric paradigm not only limits the ability to exploit horizon-dependent historical evidence but also obscures how past observations influence future predictions. To address this issue, we reformulate long-term forecasting as Patch-to-Horizon Contribution Learning (PHCL), a contribution-oriented formulation that explicitly learns the contribution mapping from historical patches to future horizons. Based on this paradigm, we propose Patch-Horizon Lifting (PHL), a contribution-aware forecasting framework that constructs a multi-resolution historical evidence space and adaptively routes resolution-patch candidates to horizon-specific contribution estimators. Extensive experiments on standard long-term forecasting benchmarks demonstrate consistent improvements over representative forecasting methods, validating the effectiveness of horizon-specific contribution learning.

1 INTRODUCTION

Long-term time series forecasting (LTSF) has witnessed remarkable progress in recent years, driven by increasingly powerful deep forecasting models. Despite their architectural diversity, including Transformer-based Zhou et al. (2020); Wu et al. (2021); Zhou et al. (2022b); Liu et al. (2022a), patch-based Nie et al. (2023); Stitsyuk & Choi (2025); Tang & Zhang (2025); Wu et al. (2026), MLP and convolutional forecasters Zeng et al. (2023); Wang et al. (2024); Luo & Wang (2024), and frequency-aware models Zhou et al. (2022a); Fei et al. (2025), these approaches largely share a common *representation-centric* paradigm. Specifically, they learn representations of historical observations from which predictions at future horizons are subsequently decoded. However, this shifts the modeling focus toward representing historical observations rather than learning how they contribute to future forecasts.

This perspective raises a question: *What does "contribution" actually mean in long-term forecasting?* In this work, we define the contribution of a historical patch as its relative predictive effect on a target forecasting horizon. Unlike attention weights, which characterize information aggregation during representation learning, contribution measures how historical evidence affects future prediction, making it a property of the forecasting objective rather than an intermediate representation Chen et al. (2025); Kudrat et al. (2025).

This leads to a further question: *Which historical patches contribute to each forecasting horizon, and do these contributions remain invariant across different horizons?* Different forecasting horizons correspond to different prediction objectives and therefore require different historical evidence Tong & Yuan (2025; 2026); Kudrat et al. (2025). A short-term forecast is typically dominated by recent local dynamics, whereas a long-term forecast often depends more on slowly varying trends or periodic structures Liu et al. (2025); Lu et al. (2025); Masserano et al. (2025). Consequently, the predictive value of a historical patch is horizon-dependent Hu et al. (2025); Tong & Yuan (2026). As illustrated in Figure 1, the predictive contribution of historical patches should therefore vary across forecasting horizons rather than be uniformly shared. By compressing the historical window into

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

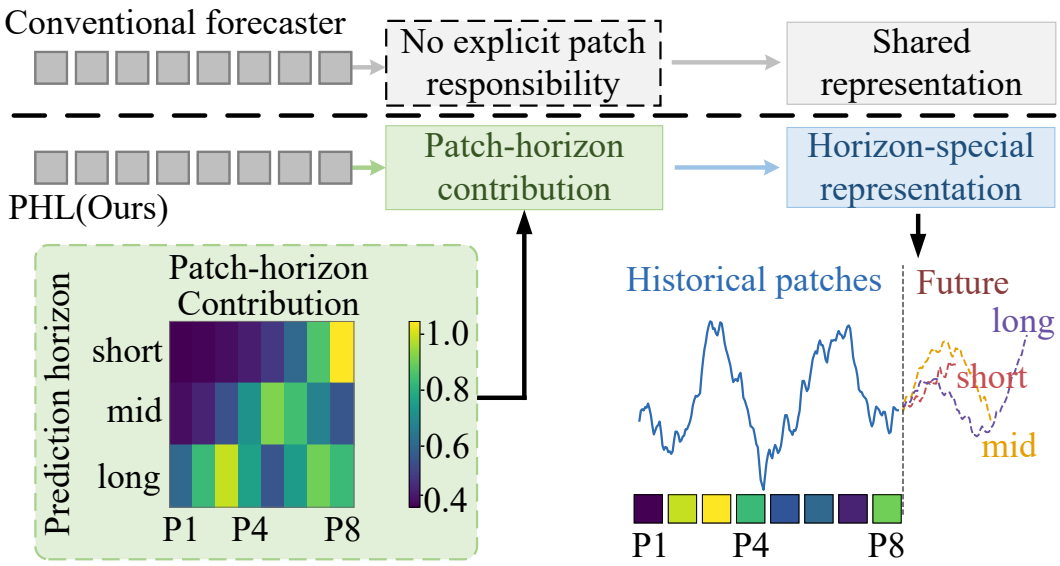


Figure 1: Motivation of Patch-to-Horizon Contribution Learning (PHCL). Different forecasting horizons require different historical evidence, implying that forecasting should explicitly learn horizon-specific historical contributions rather than relying on a shared historical representation.

a shared latent representation, existing representation-centric models implicitly entangle patches with different predictive roles, making horizon-specific evidence difficult to distinguish Wang et al. (2025c); Hu et al. (2025). Thus, representation learning and contribution learning emphasize different objectives: the former seeks expressive historical representations, whereas the latter seeks explicit predictive responsibility for future forecasts.

Motivated by this observation, we introduce Patch-to-Horizon Contribution Learning (PHCL), a contribution-oriented formulation of LTSF. Unlike conventional representation learning, PHCL explicitly models the predictive contribution of each historical patch to each forecasting horizon. This reformulation shifts the central question from *"How should historical observations be represented?"* to *"How does each historical patch contribute to future predictions?"* Compared with recent patch selection or patch enhancement methods Wu et al. (2026); Qiu et al. (2026), which primarily improve intermediate representations, PHCL models horizon-specific predictive contributions, enabling horizon-aware forecasting and patch-to-horizon contribution analysis.

To instantiate PHCL, we propose **Patch-Horizon Lifting (PHL)**, a contribution-aware forecasting framework. PHL constructs complementary multi-resolution historical evidence, identifies horizon-specific informative evidence, estimates patch-level predictive contributions, and reconstructs future forecasts through contribution aggregation. By modeling each stage of this process, PHL assembles predictions from horizon-specific historical contributions rather than decoding them solely from a shared historical representation.

Our main contributions are summarized as follows:

- We introduce PHCL, a contribution-oriented paradigm that reformulates long-term time series forecasting from historical representation learning to horizon-specific contribution learning.
- We propose PHL, an effective instantiation of PHCL that estimates and aggregates horizon-specific historical patch contributions through adaptive routing over complementary multi-resolution temporal evidence.
- We provide an intrinsic interpretation of forecasting by revealing the horizon-specific predictive contributions of historical patches.
- Extensive experiments on standard long-term forecasting benchmarks demonstrate that PHL achieves state-of-the-art or highly competitive performance, supporting the effectiveness of Patch-to-Horizon Contribution Learning.

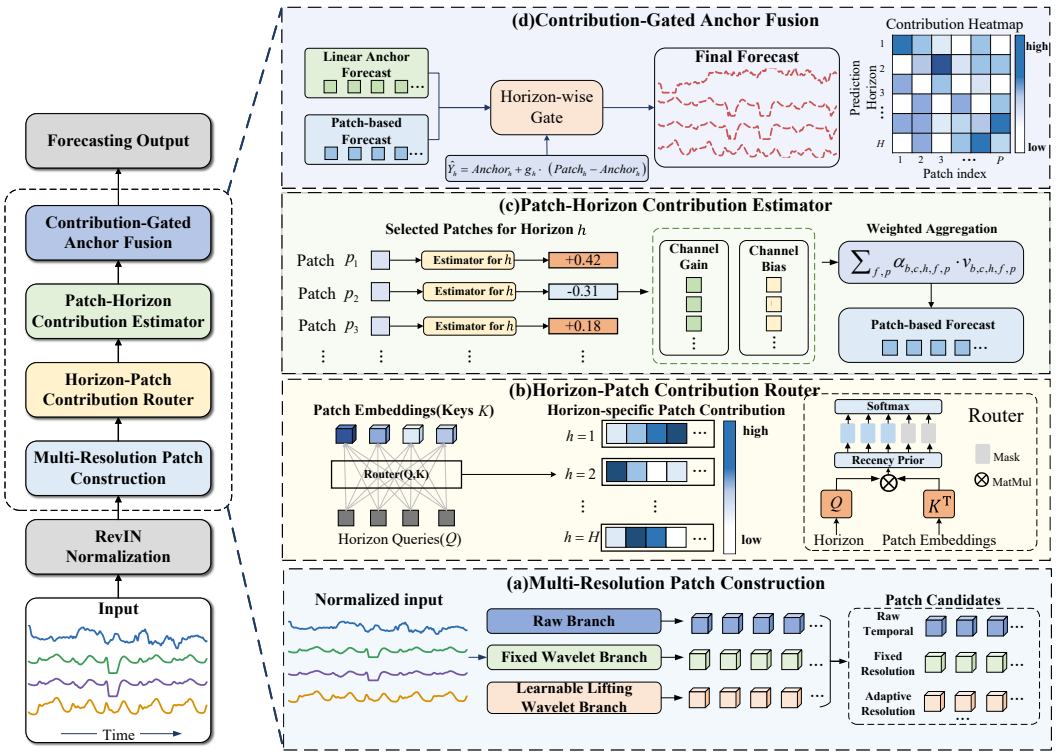


Figure 2: Overview of the proposed Patch-Horizon Lifting (PHL) framework for Patch-to-Horizon Contribution Learning. PHL first constructs raw, fixed-wavelet, and learnable lifting patch candidates, then routes resolution-patch evidence through horizon-specific weights. The selected signed patch contributions are calibrated per channel and fused with an anchor forecast to produce the final prediction.

2 Related Work

Representation learning for LTSF. LTSF models have adopted diverse architectures, including recurrent and convolutional models Lai et al. (2018), Transformers Zhou et al. (2020); Wu et al. (2021); Zhou et al. (2022b); Liu et al. (2022a;b; 2024), and linear, MLP, or convolutional forecasters Zeng et al. (2023); Oreshkin et al. (2020); Challu et al. (2023); Wang et al. (2024); Luo & Wang (2024). Despite their different inductive biases, these methods primarily learn representations of historical observations from which future values are predicted. The role of individual historical segments at different forecast horizons is generally left implicit.

Patch-based forecasting. PatchTST Nie et al. (2023) established patches as effective units for modeling long sequences. Subsequent methods refine patch representations Stitsyuk & Choi (2025); Tang & Zhang (2025) or identify less informative patches Wu et al. (2026); Qiu et al. (2026). These studies motivate selective use of historical patches, but do not explicitly model the signed predictive contribution of each patch to individual forecast horizons.

Multi-resolution modeling and interpretability. Frequency-based forecasters Zhou et al. (2022b;a); Fei et al. (2025) and learnable lifting schemes Sweldens (1998) capture temporal patterns at different resolutions. Attention-based interpretation Lim et al. (2021) and distribution-aware forecasting Wang et al. (2025a); Pan et al. (2026); Wang et al. (2025b) address other aspects of model behavior and robustness. PHCL combines multi-resolution patch candidates with horizon-specific routing and signed contribution estimates, making the modeled relationship between selected historical evidence and each forecast horizon explicit.

3 METHOD

3.1 PROBLEM DEFINITION

Given a multivariate historical sequence $X \in \mathbb{R}^{B\times L\times C}$, where B, L, and C denote the batch size, input length, and number of variables, respectively, long-term forecasting aims to predict the future sequence $\hat{Y} \in \mathbb{R}^{B\times H\times C}$, where H denotes the forecasting horizon.

Unlike conventional approaches that learn a unified representation of historical observations, we formulate long-term forecasting as a **PHCL** problem. The goal is to learn how different historical patches contribute to different forecasting horizons. Specifically, each historical patch is treated as a candidate evidence unit, and PHL learns a horizon-specific contribution distribution over these candidates to estimate their predictive effects. Based on this formulation, the final forecast is obtained by aggregating horizon-relevant historical evidence rather than decoding from a single global representation.

3.2 MODEL OVERVIEW

The overall architecture of PHL is illustrated in Figure 2. Given a historical sequence, PHL first applies reversible instance normalization (RevIN) Kim et al. (2022) to reduce distribution variations across samples and variables. The normalized sequence is then transformed into a multi-resolution evidence pool consisting of diverse resolution-patch candidates. PHL subsequently learns horizon-specific contributions through a contribution router and patch-horizon experts, which identify informative evidence and estimate their predictive effects. Finally, a gated anchor fusion module combines the contribution-based prediction with a stable anchor forecast to obtain the final output. The following sections describe the detailed designs of these components.

3.3 MULTI-RESOLUTION PATCH CONSTRUCTION

To enable horizon-specific contribution learning, PHL first constructs a diverse evidence space from multiple temporal resolutions. Instead of learning a single latent representation of the historical sequence, PHL decomposes the input into complementary resolution-patch candidates, where each candidate serves as a potential source of predictive evidence.

Specifically, PHL adopts three complementary branches: raw temporal representation, fixed multi-resolution decomposition, and learnable adaptive decomposition. The raw branch preserves the original temporal dynamics and provides direct access to time-domain evidence. The fixed wavelet branch applies an undecimated wavelet transform, which maintains temporal alignment across resolution components and preserves the correspondence between decomposed components and historical positions.

To obtain adaptive multi-resolution candidates, PHL introduces a learnable lifting branch based on the split-predict-update scheme. Given an input sequence, we first split it into even and odd temporal samples, denoted as x_{even} and x_{odd}. The branch then performs:

$$d = x_{\text{odd}} - P_\theta(x_{\text{even}}), \quad a = x_{\text{even}} + U_\theta(d), \tag{1}$$

where P_θ and U_θ are learnable depth-wise one-dimensional convolutional filters. Here, d and a denote the detail and approximation components produced by the lifting transform, respectively. Specifically, P_θ predicts odd-position samples from the even-position scaffold, and the prediction discrepancy forms the detail component d that captures local variations. The update operation then incorporates this detail information into x_{even} to obtain the refined approximation component a. Unlike fixed wavelet decomposition, the lifting filters are initialized with Haar-like coefficients and jointly optimized with the forecasting objective, enabling the model to adaptively discover dataset-specific resolution patterns.

For a J-level decomposition, the fixed wavelet and lifting branches each generate $J + 1$ resolution components. Together with the raw branch, PHL obtains

$$F = 1 + 2(J + 1) \tag{2}$$

resolution-specific evidence components. Before patch extraction, we use transposed convolutions to upsample the lifting components to length L, aligning them with the raw and fixed-wavelet components on the original time grid.

Each evidence component is further segmented into overlapping patches with patch length P_l and stride S. Given the look-back length L of the historical input sequence, the number of patches is calculated as

$$P = \left\lfloor \frac{L + S - P_l}{S} \right\rfloor + 1. \tag{3}$$

The resulting resolution-patch candidates are denoted as $x_{b,c,f,p}$, where f indexes the temporal resolution component, and p denotes the historical patch position. These candidates form the evidence pool for subsequent horizon-specific contribution learning, where the predictive role of each historical patch is explicitly estimated with respect to different forecasting horizons.

3.4 Horizon-Patch Contribution Router

Given the constructed patch candidates, PHL learns a horizon-specific contribution distribution to identify which historical evidence is responsible for each future prediction. Unlike conventional attention mechanisms that primarily enhance feature aggregation, the router explicitly models the predictive contribution of each patch for different forecasting horizons.

For each patch candidate, PHL extracts lightweight descriptive statistics, including patch mean, log standard deviation, endpoint value, normalized slope, endpoint variation, evidence component index, and patch position. A key encoder maps these statistics into candidate representations $K_{b,c,f,p}$. Meanwhile, each forecasting horizon is associated with a learnable query vector Q_h, which captures the horizon-dependent preference for historical evidence. The matching score is defined as:

$$s_{b,c,h,f,p} = \frac{Q_h^\top K_{b,c,f,p}}{\sqrt{D_s}} + \lambda \pi_{h,f,p}, \tag{4}$$

where D_s denotes the score dimension and $\pi_{h,f,p}$ is a weak temporal-resolution prior.

The prior provides a simple inductive bias that favors recent patches while allowing longer horizons to access broader historical information:

$$\pi_{h,f,p} = -\left(\frac{1 - r_p}{0.16 + 0.84h/H}\right)^2 - 0.15 u_f h/H, \tag{5}$$

where r_p denotes the normalized patch recency and u_f represents the normalized evidence-component index. The prior does not determine patch selection but only guides the router toward temporally plausible evidence.

PHL then selects the top-k candidates and applies softmax normalization over the selected candidates:

$$\alpha_{b,c,h,f,p} = \mathrm{softmax}(s_{b,c,h,f,p}/\tau). \tag{6}$$

The routing weights satisfy $\sum_{f,p} \alpha_{b,c,h,f,p} = 1$ for each variable and forecasting horizon. This sparse contribution bottleneck forces the model to identify a compact set of historical evidence rather than uniformly aggregating all patches. The resulting routing weights naturally form a patch-to-horizon contribution map, providing an interpretable view of how historical evidence influences different forecasting horizons, as illustrated in Figure 3.

3.5 Patch-Horizon Contribution Estimator

After obtaining the horizon-specific routing weights, PHL estimates the predictive effect of each selected historical patch through patch-horizon contribution estimator. Different from conventional forecasting heads that directly generate future values from latent representations, the estimator module explicitly estimates how much each historical patch contributes to a specific forecasting horizon.

For each resolution-patch candidate, the estimator produces a signed estimated effect:

$$v_{b,c,h,f,p} = \sum_l x_{b,c,f,p,l} W_{h,f,p,l} + b_{h,f,p}. \tag{7}$$

Here, $v_{b,c,h,f,p}$ represents the effect of a specific historical patch to horizon h, rather than a direct prediction of future values. $x_{b,c,f,p,l}$ denotes the l-th feature dimension of the resolution-patch

Models	PHL (Ours)		SEER 2026		DistDF 2026		SRSNet 2025		DUET 2025		xPatch 2025		Amplifier 2025		iTransformer 2024	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.400**	**0.421**	0.423	0.424	0.430	0.429	0.442	0.433	0.443	0.436	0.444	0.429	0.449	0.437	0.454	0.448
ETTh2	**0.311**	**0.369**	0.356	0.384	0.367	0.393	0.376	0.402	0.372	0.398	0.369	0.392	0.389	0.411	0.383	0.407
ETTm1	**0.346**	**0.376**	0.377	0.381	0.378	0.394	0.385	0.395	0.390	0.393	0.387	0.383	0.383	0.396	0.407	0.410
ETTm2	**0.254**	**0.313**	0.276	0.317	0.277	0.321	0.284	0.329	0.280	0.324	0.280	0.319	0.280	0.326	0.288	0.332
Weather	**0.226**	**0.263**	0.246	0.265	0.248	0.275	0.250	0.277	0.251	0.273	0.247	0.266	0.247	0.275	0.258	0.278
Solar	**0.208**	0.249	0.221	**0.224**	-	-	0.250	0.281	0.237	0.233	0.277	0.268	0.222	0.256	0.233	0.262
Electricity	**0.161**	0.254	0.169	**0.251**	0.172	0.267	0.189	0.276	0.172	0.259	0.184	0.266	0.174	0.268	0.178	0.270
Traffic	**0.413**	0.282	0.422	**0.245**	0.417	0.279	0.494	0.307	0.451	0.269	0.500	0.283	0.485	0.315	0.428	0.282

Table 1: Average forecasting results on standard long-term forecasting benchmarks. Results are reported as MSE/MAE averaged over prediction lengths $H \in \{96, 192, 336, 720\}$.

candidate. $W_{h,f,p,l}$ and $b_{h,f,p}$ are learnable parameters initialized before training and optimized jointly with the forecasting objective.

The contribution-based forecast is obtained by aggregating horizon-specific patch contributions:

$$\hat{P}_{b,c,h} = \sum_{f=1}^{F} \sum_{p=1}^{P} \alpha_{b,c,h,f,p} v_{b,c,h,f,p}. \tag{8}$$

Therefore, each prediction can be explicitly decomposed into weighted contributions from historical patches, providing a direct correspondence between historical evidence and future horizons.

3.6 Contribution-Gated Anchor Fusion

Although the contribution branch provides horizon-specific evidence aggregation, relying solely on learned contributions may introduce instability when informative historical evidence is insufficient. To provide a stable forecasting baseline, PHL introduces an anchor prediction $A_{b,c,h}$ initialized as a persistence forecast.

The final prediction is obtained through horizon-wise gated fusion:

$$\hat{Y}_{b,c,h} = A_{b,c,h} + g_h(\hat{P}_{b,c,h} - A_{b,c,h}), \tag{9}$$

where g_h is a learnable horizon-dependent gate constrained within a stable range.

This formulation adaptively balances the stable anchor and the contribution-based prediction. For horizons where historical evidence provides reliable predictive signals, the gate allows the contribution branch to dominate; otherwise, the model retains the anchor prediction as a robust fallback. Consequently, the final forecast can be interpreted as a horizon-dependent combination between prior knowledge and learned historical evidence.

4 Experiments

4.1 Datasets and Baselines

We evaluate PHL on eight widely used benchmarks, including ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Traffic, and Solar Zhou et al. (2020); Lai et al. (2018). Following previous studies Murad et al. (2025); Qiu et al. (2025); Wu et al. (2026), we use an input length of 512 and evaluate four forecasting horizons $H \in \{96, 192, 336, 720\}$. The forecasting performance is measured by mean squared error (MSE) and mean absolute error (MAE). We compare PHL with representative state-of-the-art methods, including patch-based forecasters SEER Qiu et al. (2026), SRSNet Wu et al. (2026), xPatch Stitsyuk & Choi (2025)), frequency-based model Amplifier Fei et al. (2025), distribution-aware method DistDF Wang et al. (2025a), Transformer-based models (DUET Qiu et al. (2025), iTransformer Liu et al. (2024)), and other recent baselines. The detailed implementation configurations, inclZuding PHL-specific hyperparameters and training settings are provided in Appendix A.

Variant	ETTh1		ETTm1	
	MSE	MAE	MSE	MAE
Fixed wavelet only	0.4032	0.4241	0.3481	0.3763
Lifting wavelet only	0.4015	0.4227	0.3464	0.3762
w/o raw resolution	0.4008	0.4241	0.3481	0.3755
Uniform router	0.4069	0.4287	0.3529	0.3744
Prior-only router	0.4024	0.4247	0.3508	**0.3733**
Dense router	0.4039	0.4260	0.3468	0.3772
w/o recency prior	0.4098	0.4311	0.3531	0.3823
Full PHL	**0.4002**	**0.4214**	**0.3455**	0.3758

Table 2: Ablation study for PHL on ETTh1 and ETTm1.

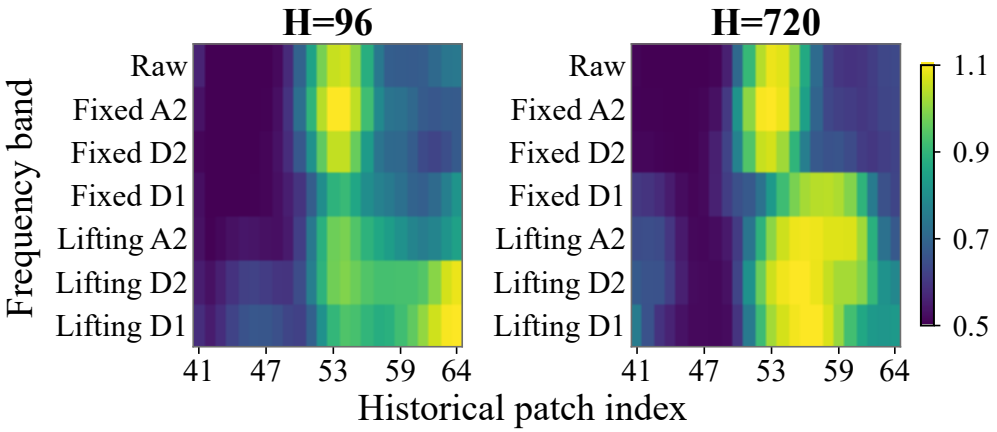


Figure 3: Patch contribution maps on Solar. Each heatmap averages absolute signed patch contributions over test batches and variables.

4.2 Main Results

Table 1 summarizes the forecasting performance of PHL and representative baselines over four prediction horizons ($H = 96, 192, 336, 720$), while detailed horizon-wise results are provided in Appendix B. PHL achieves the best or highly competitive performance across most benchmarks, demonstrating the effectiveness of horizon-specific contribution learning for long-term forecasting.

Compared with representative baselines, PHL achieves the best average MSE on all eight datasets and outperforms most existing methods in average MAE, demonstrating strong performance across the evaluated benchmarks. These results suggest that explicitly modeling horizon-specific contributions from multi-resolution historical patches can improve forecasting accuracy. Together with the ablation results, they support the effectiveness of the proposed patch-to-horizon contribution learning framework.

4.3 Ablation Study

To validate the effectiveness of key components, we conduct ablation studies on ETTh1 and ETTm1, with results summarized in Table 2.

The uniform router provides the most direct validation of our hypothesis. Replacing adaptive routing with uniform weighting increases MSE from 0.4002 to 0.4069 on ETTh1 and from 0.3455 to 0.3529 on ETTm1, confirming that historical patches have unequal predictive contributions and should be selectively aggregated according to forecasting horizons. Removing the recency prior causes the largest degradation, increasing MSE to 0.4098 and 0.3531 on ETTh1 and ETTm1, respectively. This demonstrates that the temporal prior provides an effective inductive bias for locating informative historical evidence. The multi-resolution design is also important: removing the raw branch or

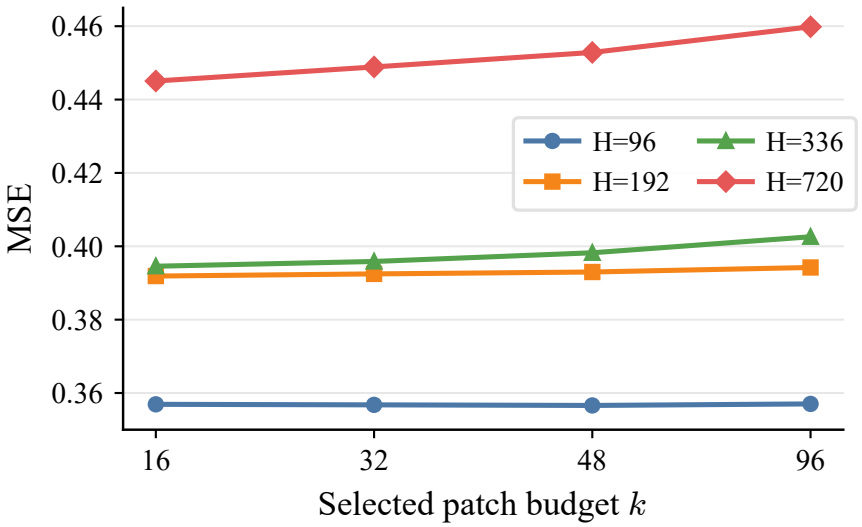


Figure 4: Sensitivity of the contribution bottleneck size k on ETTh1. A compact patch selection achieves better forecasting performance than dense aggregation.

using only a single wavelet representation consistently degrades performance, indicating that time-domain and resolution-aware evidence are complementary. Moreover, dense routing performs worse than sparse routing, suggesting that excessive historical evidence introduces redundancy.

4.4 Contribution Analysis

Beyond numerical improvements, PHL explicitly models how historical evidence contributes to different forecasting horizons. Figure 3 visualizes the learned contribution distributions over resolution-patch units on Solar.

Unlike uniform aggregation, PHL learns highly non-uniform contribution patterns across both temporal positions and resolution levels. Only a subset of patches receives dominant contributions, while many patches contribute marginally. This indicates that historical evidence has different predictive importance and supports the motivation of patch-to-horizon contribution learning.

Notably, PHL assigns contributions to resolution-patch units rather than individual timestamps, enabling joint modeling of temporal locality and frequency characteristics. Recent high-resolution patches can capture short-term variations, while low-resolution patches provide long-term trend information.

Table 3 quantifies the routing behavior on ETTh1 (H=192). Compared with uniform routing, PHL achieves lower entropy and assigns a larger proportion of contribution to the selected top-k patches. Removing the recency prior preserves sparse selection but substantially reduces recent-patch contributions, confirming that the prior provides a meaningful temporal bias rather than directly determining routing decisions. Appendix C further shows that the performance gains of PHL stem from learning where and how historical patches contribute to future predictions.

Appendix D further presents patch contribution maps alongside forecasting results. These visualizations show that the selected historical evidence changes across prediction steps, while longer horizons draw on a broader but still sparse set of patches.

4.5 Sensitivity Analysis

We further analyze the sensitivity of the contribution bottleneck and key design choices. Figure 4 shows the influence of the selected patch number k on ETTh1. Increasing k does not consistently

Variant	Entropy	Top-k Contribution	Recent Contribution
Full PHL	1.52	1.00	0.81
Uniform router	4.16	0.11	0.13
w/o recency prior	1.47	1.00	0.15
w/o sparse top-k	2.27	0.87	0.75

Table 3: Router diagnostics on ETTh1. Top-k and Recent Contribution denote the proportions of total contribution assigned to selected and recent patches, respectively.

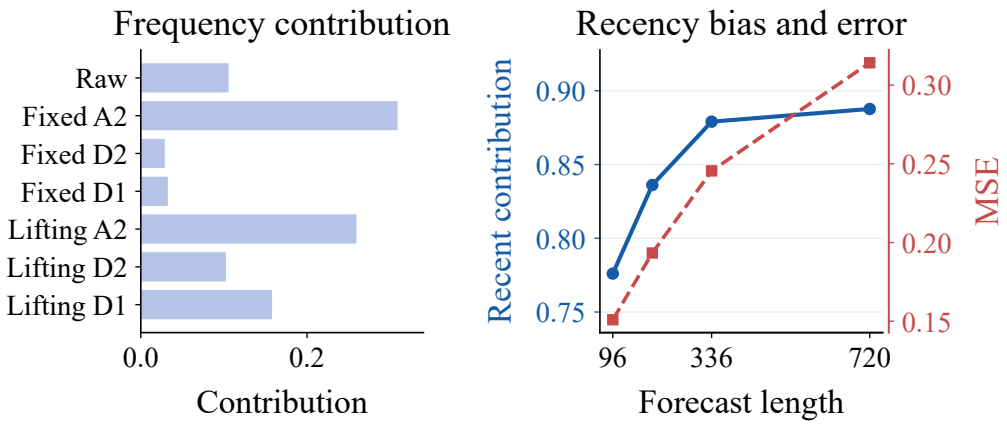


Figure 5: Frequency and recency structure of learned patch contributions on Weather. Left: the contribution mass is concentrated in approximation bands and learnable lifting bands, indicating that PHL uses both stable low-frequency structure and adaptive wavelet evidence. Right: recent-patch contribution remains high as the forecast length increases, while the forecasting error also grows with horizon length.

improve performance; overly large routing budgets may introduce redundant evidence and weaken the contribution bottleneck. The default setting $k = 48$ balances evidence coverage and sparsity.

Figure 5 shows that learned contributions concentrate on approximation and adaptive lifting resolutions, suggesting that the lifting branch enriches the evidence pool beyond fixed wavelet bases. Contributions also remain biased toward recent history while incorporating broader context at longer forecasting horizons. Together, these results support the use of diverse multi-resolution evidence and sparse, horizon-specific patch selection.

5 Conclusion

In this paper, we proposed Patch-Horizon Lifting (PHL), a contribution-aware framework that formulates long-term time series forecasting as patch-to-horizon contribution learning. PHL constructs multi-resolution historical evidence and uses horizon-specific routing to select relevant patches for each future step. It then estimates signed predictive contributions from these patches and aggregates them into a forecast, making the role of selected historical evidence explicit. Extensive experiments and ablation studies demonstrate that historical patches contribute unequally to future predictions, and that explicitly modeling these horizon-specific contributions improves forecasting accuracy. Rather than replacing representation learning, PHL complements it by making the predictive roles of historical patches explicit. More broadly, our results suggest that progress in long-term forecasting depends not only on how well history is represented, but also on how effectively the model identifies and uses the evidence relevant to each future horizon.

AI USE STATEMENT

We used generative AI tools to help identify relevant literature and improve the wording of selected passages. All suggested references were checked by the authors against the original sources, and all AI-assisted edits were reviewed and revised by the authors. We did not use generative AI tools for other research or manuscript preparation tasks. The authors take full responsibility for the content of this paper.

REFERENCES

Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 6989–6997, 2023.

Yu Chen, Nathalia Céspedes, and Payam Barnaghi. A closer look at transformers for time series forecasting: Understanding why they work and where they struggle. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 7763–7780, 2025.

Jingru Fei, Kun Yi, Wei Fan, Qi Zhang, and Zhendong Niu. Amplifier: Bringing attention to neglected low-energy components in time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11645–11653, 2025.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *International Conference on Machine Learning*, pp. 24893–24911. PMLR, 2025.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *10th International Conference on Learning Representations, ICLR 2022*. International Conference on Learning Representations, ICLR, 2022.

Dilfira Kudrat, Zongxia Xie, Yanru Sun, Tianyu Jia, and Qinghua Hu. Patch-wise structural loss for time series forecasting. In *International Conference on Machine Learning*, pp. 31841–31859. PMLR, 2025.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104. ACM, 2018.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting. In *International Conference on Machine Learning*, pp. 39815–39840. PMLR, 2025.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *10th International Conference on Learning Representations, ICLR 2022*, 2022a.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35: 9881–9893, 2022b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Jiecheng Lu, Xu Han, Yan Sun, and Shihao Yang. Wave: Weighted autoregressive varying gate for time series forecasting. In *International Conference on Machine Learning*, pp. 40464–40490. PMLR, 2025.

Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *ICLR*, 2024.

Luca Masserano, Abdul Fatir Ansari, Boran Han, Xiyuan Zhang, Christos Faloutsos, Michael W. Mahoney, Andrew Gordon Wilson, Youngsuk Park, Syama Sundar Rangapuram, Danielle C. Maddix, and Bernie Wang. Enhancing foundation models for time series forecasting via wavelet-based tokenization. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 43248–43275, 2025.

Md Mahmuddun Nabi Murad, Mehmet Aktukmak, and Yasin Yilmaz. Wpmixer: Efficient multi-resolution mixing for long-term time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 19581–19588, 2025.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. 2020.

Licheng Pan, Hao Wang, Haocheng Yang, Yuqi Li, Qingsong Wen, Xiaoxi Li, Zhichao Chen, Haoxuan Li, Zhixuan Chu, and Yuan Lu. Deep time-series forecasting needs kernelized moment balancing. *arXiv preprint arXiv:2602.00717*, 2026.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1185–1196, 2025.

Xiangfei Qiu, Xvyuan Liu, Tianen Shen, Xingjian Wu, Hanyin Cheng, Bin Yang, and Jilin Hu. Seer: Transformer-based robust time series forecasting via automated patch enhancement and replacement. *arXiv preprint arXiv:2602.00589*, 2026.

Artyom Stitsyuk and Jaesik Choi. xpatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20601–20609, 2025.

Wim Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM journal on mathematical analysis*, 29(2):511–546, 1998.

Peiwang Tang and Weitai Zhang. Unlocking the power of patch: Patch-based mlp for long-term time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12640–12648, 2025.

Suxin Tong and Jingling Yuan. Efficiently enhancing long-term series forecasting via ultra-long lookback windows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20912–20920, 2025.

Suxin Tong and Jingling Yuan. Efficiently enhancing long-term series forecasting via adaptive lookback with wavelets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25966–25974, 2026.

Hao Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zhichao Chen, Haoxuan Li, Qingsong Wen, and Zhouchen Lin. Distdf: Time-series forecasting needs joint-distribution wasserstein alignment. *arXiv preprint arXiv:2510.24574*, 2025a.

Hua Wang, Jinghao Lu, and Fan Zhang. Idealtsf: Can non-ideal data contribute to enhancing the performance of time series forecasting models? *arXiv preprint arXiv:2512.05442*, 2025b.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 38626–38652, 2024.

Yihang Wang, Yuying Qiu, Peng Chen, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series forecasting model with unified representation and adaptive transfer. In *International Conference on Machine Learning*, pp. 64127–64151. PMLR, 2025c.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Xingjian Wu, Xiangfei Qiu, Hanyin Cheng, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Enhancing time series forecasting through selective representation spaces: A patch perspective. *Advances in Neural Information Processing Systems*, 38:23328–23354, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *arXiv preprint arXiv:2012.07436*, 2020.

Tian Zhou, Ziqing Ma, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, Rong Jin, et al. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in neural information processing systems*, 35:12677–12690, 2022a.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022b.

A Dataset Statistics and Implementation Details

We evaluate PHL on eight widely used long-term forecasting benchmarks, including ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Traffic, and Solar Zhou et al. (2020); Lai et al. (2018). These datasets cover diverse forecasting scenarios, including energy systems, traffic networks, meteorological observations, and solar power generation. The detailed statistics of these datasets, including domains, sampling frequencies, sequence lengths, variable dimensions, and train/validation/test splits, are summarized in Table A1.

For all experiments, the input sequence length is fixed to 512, and the prediction horizons are selected from $H \in \{96, 192, 336, 720\}$. We employ mean squared error (MSE) as the optimization objective and report both MSE and mean absolute error (MAE) for evaluation. Following previous studies, all experiments use the standard train/validation/test splits and forecasting protocols for fair comparison.

The main configuration of PHL uses a patch length of 16, a stride of 8, top-k, with $k = 48$, and a score dimension of 64. For multi-resolution historical evidence construction, ETTh1 and Weather adopt a two-level db2 fixed wavelet decomposition, while ETTm1 uses a one-level db3 decomposition. The learnable lifting branch is initialized with a Haar-like split-predict-update transform and optimized jointly with the forecasting objective.

The implementation of PHL consists of aligned resolution-patch construction, contribution routing, patch-horizon expert forecasting, and gated anchor fusion modules. The raw and fixed wavelet branches generate temporally aligned patch representations, allowing the router to evaluate different resolution-patch units through a unified scoring mechanism. The learnable lifting branch employs depth-wise one-dimensional convolutions in the predict and update operators, which preserve channel-wise decomposition while enabling adaptive learning of temporal patterns.

The anchor predictor is initialized as a persistence forecast to provide a stable baseline during early optimization, before the patch-horizon estimator learn reliable signed contributions. During training, the forecasting loss is computed between the denormalized prediction and the ground-truth target. The router temperature, top-k value, and wavelet decomposition level are determined according to

validation performance. Unless otherwise specified, each dataset uses a unified configuration across different forecasting horizons rather than horizon-specific model designs.

Dataset	Scenario	Sampling	Length	Dim.	Split	Description
Electricity	Electricity load	1 hour	26,304	321	7:1:2	Hourly client-level electricity consumption from 2012 to 2014.
Traffic	Transportation	1 hour	17,544	862	7:1:2	Hourly road occupancy from 862 freeway sensors.
Weather	Meteorology	10 min	52,696	21	7:1:2	Weather measurements with 21 variables in 2020.
Solar	Renewable energy	10 min	52,560	137	6:2:2	Solar production from 137 photovoltaic plants.
ETTh1	Transformer temperature	1 hour	14,400	7	6:2:2	Hourly records from electricity transformer 1.
ETTh2	Transformer temperature	1 hour	14,400	7	6:2:2	Hourly records from electricity transformer 2.
ETTm1	Transformer temperature	15 min	57,600	7	6:2:2	Fifteen-minute records from transformer 1.
ETTm2	Transformer temperature	15 min	57,600	7	6:2:2	Fifteen-minute records from transformer 2.

Table A1: Statistics of long-term forecasting benchmarks. Length denotes the number of timestamps, Dim denotes the number of variables, and Split represents the train/validation/test ratio.

B Detailed Forecasting Results

This section provides detailed forecasting results under different prediction horizons. Table A2 reports the MSE and MAE values of PHL and baseline methods on eight benchmark datasets with forecasting lengths $H \in \{96, 192, 336, 720\}$. The average results reported in the main paper are computed over these four forecasting horizons.

As shown in Table A2, PHL consistently achieves competitive performance across different forecasting settings. On the ETT benchmarks, PHL obtains substantial improvements over existing methods, demonstrating the effectiveness of contribution-aware historical evidence modeling for datasets with complex temporal dependencies. The improvements become more evident as the forecasting horizon increases, suggesting that horizon-specific contribution learning enables the model to identify long-range relevant evidence rather than relying on fixed historical representations.

Compared with the strongest patch-based competitor SEER, PHL achieves consistent improvements on several benchmarks. For example, on ETTh1, PHL reduces the average MSE and MAE from 0.423 and 0.424 of SEER to 0.400 and 0.422, respectively. Similar improvements are observed on ETTm1, where PHL decreases the average MSE from 0.377 to 0.345 and the MAE from 0.381 to 0.376. These improvements indicate that contribution-aware aggregation can effectively identify more informative historical evidence compared with conventional patch representation learning.

Overall, PHL achieves state-of-the-art or highly competitive performance across most benchmarks. In particular, PHL obtains the best average MSE on seven out of eight datasets and consistently outperforms recent patch-based forecasting methods, including SEER, SRSNet, and xPatch. These results demonstrate that explicitly modeling horizon-specific patch contributions provides additional benefits beyond improving historical representations.

C Contribution Concentration Anlysis

Figure A1 analyzes the relationship between learned contribution distributions and forecasting performance on ETTh1. The x-axis reports the contribution entropy, which measures how contribution weights are distributed across historical patches. Lower entropy indicates that contributions are concentrated on a small number of informative patches, whereas higher entropy corresponds to a more diffuse or nearly uniform allocation. The dashed vertical line denotes the entropy of an idealized

		PHL (Ours)		SEER 2026		DistDF 2026		SRSNet 2025		DUET 2025		xPatch 2025		Amplifier 2025		iTransformer 2024	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.357**	0.389	0.366	**0.385**	0.373	0.393	0.383	0.395	0.377	0.393	0.378	0.390	0.376	0.393	0.386	0.405
	192	**0.393**	**0.411**	0.417	0.416	0.428	0.425	0.433	0.422	0.429	0.425	0.433	0.420	0.442	0.430	0.441	0.436
	336	**0.398**	**0.424**	0.454	0.435	0.466	0.445	0.476	0.446	0.471	0.446	0.484	0.445	0.478	0.446	0.487	0.458
	720	**0.453**	0.462	0.456	0.459	**0.453**	**0.453**	0.474	0.471	0.496	0.480	0.480	0.462	0.501	0.479	0.503	0.491
	Avg	**0.400**	**0.421**	0.423	0.424	0.430	0.429	0.442	0.433	0.443	0.436	0.444	0.429	0.449	0.437	0.454	0.448
ETTh2	96	**0.253**	**0.322**	0.281	0.329	0.287	0.336	0.296	0.345	0.296	0.345	0.287	0.332	0.298	0.347	0.297	0.349
	192	**0.304**	**0.360**	0.358	0.379	0.358	0.381	0.369	0.392	0.368	0.389	0.360	0.382	0.378	0.401	0.380	0.400
	336	**0.308**	**0.371**	0.401	0.413	0.408	0.421	0.413	0.425	0.411	0.422	0.417	0.421	0.420	0.437	0.428	0.432
	720	**0.379**	0.421	0.384	**0.417**	0.416	0.435	0.425	0.444	0.412	0.434	0.412	0.433	0.452	0.460	0.427	0.445
	Avg	**0.311**	**0.369**	0.356	0.384	0.367	0.393	0.376	0.402	0.372	0.398	0.369	0.392	0.389	0.411	0.383	0.407
ETTm1	96	**0.292**	0.343	0.308	**0.337**	0.316	0.357	0.319	0.358	0.324	0.354	0.316	0.343	0.318	0.356	0.334	0.368
	192	**0.323**	**0.362**	0.361	0.369	0.358	0.380	0.359	0.381	0.369	0.379	0.369	0.369	0.362	0.381	0.377	0.391
	336	**0.355**	**0.384**	0.391	0.391	0.392	0.404	0.391	0.404	0.403	0.402	0.401	0.392	0.393	0.404	0.426	0.420
	720	**0.411**	**0.415**	0.450	0.429	0.448	0.437	0.470	0.436	0.463	0.437	0.461	0.429	0.460	0.442	0.491	0.459
	Avg	**0.346**	**0.376**	0.377	0.381	0.378	0.394	0.385	0.395	0.390	0.393	0.387	0.383	0.383	0.396	0.407	0.410
ETTm2	96	**0.164**	0.254	0.172	**0.249**	0.174	0.256	0.181	0.267	0.174	0.255	0.174	0.252	0.178	0.261	0.180	0.264
	192	**0.220**	**0.292**	0.238	0.294	0.239	0.298	0.243	0.306	0.243	0.302	0.240	0.297	0.243	0.303	0.250	0.309
	336	**0.275**	**0.327**	0.297	0.334	0.300	0.338	0.306	0.346	0.304	0.341	0.302	0.335	0.305	0.344	0.311	0.348
	720	**0.357**	**0.379**	0.395	0.390	0.397	0.394	0.407	0.399	0.399	0.397	0.403	0.393	0.393	0.397	0.412	0.407
	Avg	**0.254**	**0.313**	0.276	0.317	0.277	0.321	0.284	0.329	0.280	0.324	0.280	0.319	0.280	0.326	0.288	0.332
Weather	96	**0.151**	0.201	0.162	**0.198**	0.164	0.209	0.167	0.214	0.163	0.202	0.166	0.202	0.165	0.210	0.174	0.214
	192	**0.193**	**0.240**	0.211	**0.240**	0.212	0.252	0.215	0.255	0.218	0.252	0.210	0.242	0.212	0.253	0.224	0.254
	336	**0.246**	**0.280**	0.268	0.284	0.270	0.295	0.270	0.294	0.274	0.294	0.267	0.284	0.267	0.293	0.278	0.296
	720	**0.314**	**0.331**	0.344	0.336	0.348	0.345	0.346	0.344	0.349	0.343	0.344	0.335	0.344	0.342	0.358	0.349
	Avg	**0.226**	**0.263**	0.246	0.265	0.248	0.275	0.250	0.277	0.251	0.273	0.247	0.266	0.247	0.275	0.258	0.278
Solar	96	0.189	0.236	**0.185**	**0.199**	-	-	0.216	0.258	0.200	0.207	0.234	0.245	0.186	0.232	0.203	0.237
	192	**0.206**	0.247	0.220	**0.222**	-	-	0.247	0.280	0.228	0.233	0.265	0.262	0.231	0.264	0.233	0.261
	336	**0.214**	0.254	0.238	**0.233**	-	-	0.268	0.294	0.262	0.244	0.301	0.280	0.234	0.263	0.248	0.273
	720	**0.221**	0.257	0.241	**0.241**	-	-	0.268	0.290	0.258	0.249	0.308	0.284	0.238	0.265	0.249	0.273
	Avg	**0.208**	0.249	0.221	**0.224**	-	-	0.250	0.281	0.237	0.233	0.277	0.268	0.222	0.256	0.233	0.262
Electricity	96	**0.132**	0.227	0.142	**0.225**	0.137	0.235	0.161	0.252	0.145	0.233	0.160	0.244	0.149	0.245	0.148	0.240
	192	**0.147**	0.241	0.158	**0.240**	0.159	0.257	0.172	0.261	0.163	0.248	0.169	0.253	0.165	0.260	0.162	0.253
	336	**0.164**	**0.257**	0.174	**0.257**	0.178	0.272	0.190	0.279	0.175	0.262	0.185	0.268	0.176	0.271	0.178	0.269
	720	0.202	0.290	**0.201**	**0.282**	0.208	0.300	0.231	0.313	0.204	0.291	0.221	0.300	0.204	0.296	0.225	0.317
	Avg	**0.161**	0.254	0.169	**0.251**	0.172	0.267	0.189	0.276	0.172	0.259	0.184	0.266	0.174	0.268	0.178	0.270
Traffic	96	0.390	0.271	0.392	**0.233**	**0.380**	0.262	0.471	0.295	0.407	0.252	0.475	0.279	0.450	0.295	0.395	0.268
	192	**0.402**	0.276	0.417	**0.240**	0.407	0.275	0.480	0.300	0.431	0.262	0.486	0.277	0.489	0.311	0.417	0.276
	336	**0.412**	0.280	0.430	**0.247**	0.429	0.284	0.496	0.306	0.456	0.269	0.500	0.280	0.484	0.321	0.433	0.283
	720	**0.447**	0.300	0.448	**0.261**	0.452	0.297	0.531	0.328	0.509	0.292	0.537	0.295	0.517	0.333	0.467	0.302
	Avg	**0.413**	0.282	0.422	**0.245**	0.417	0.279	0.494	0.307	0.451	0.269	0.500	0.283	0.485	0.315	0.428	0.282

Table A2: Detailed forecasting results on long-term time series forecasting benchmarks. Results are reported as MSE/MAE for prediction lengths $H \in \{96, 192, 336, 720\}$. Avg denotes the average performance over four horizons. Lower values indicate better performance.

uniform allocation. The y-axis reports the increase in MSE relative to the full PHL model, where larger values indicate greater performance degradation.

The uniform router lies closest to the uniform-allocation line and exhibits the highest contribution entropy, indicating an almost uniform contribution distribution across historical patches and a clear degradation in forecasting performance. In contrast, removing the recency prior preserves concentrated contributions but results in the largest increase in MSE, suggesting that concentration alone is insufficient. Instead, contributions must be assigned to historical patches that are most relevant to the target forecasting horizon. Overall, these results provide empirical evidence that accurate long-term forecasting requires horizon-aware, non-uniform patch contribution learning rather than uniformly aggregating historical evidence.

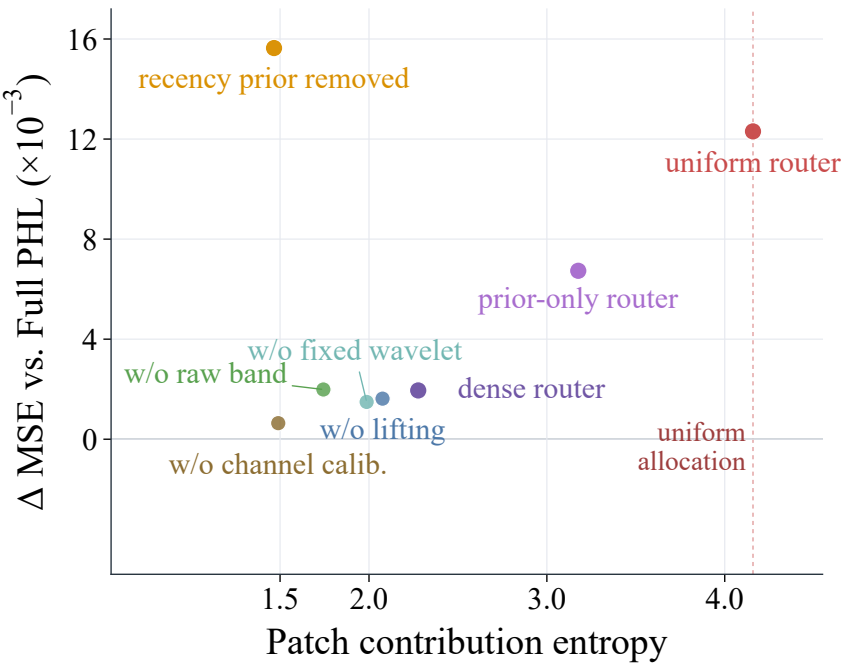
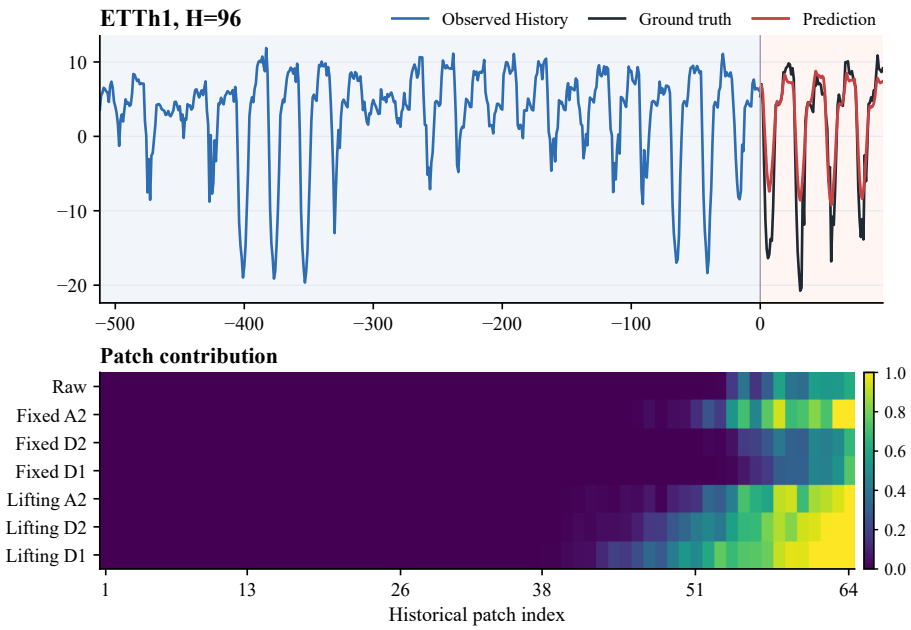


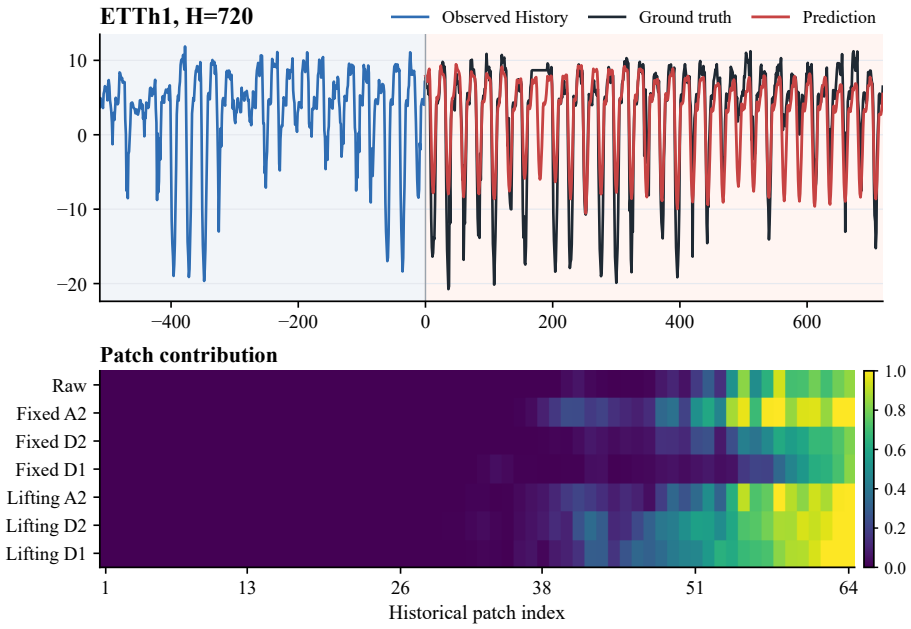
Figure A1: Contribution concentration analysis on ETTh1 (H=192). Each point denotes an ablated PHL variant. The vertical dashed line indicates the entropy of uniform contribution assignment. Uniform router exhibits the highest contribution entropy, approaching uniform allocation, whereas removing the recency prior results in the largest increase in MSE despite maintaining concentrated contributions. These observations highlight that accurate forecasting depends not only on sparse contribution allocation but also on assigning contributions to horizon-relevant historical patches.

D Horizon-wise Contribution Analysis

Figure A2 compares PHL on ETTh1 under short- (Fig. A2a, H=96) and long-horizon forecasting (Fig. A2b, H=720). In both cases, the predicted sequence closely follows the ground truth, demonstrating that PHL effectively captures temporal dynamics across different forecasting horizons. The corresponding contribution maps show that the forecasts are consistently supported by only a sparse subset of historical patches rather than uniformly aggregating the entire look-back window. Moreover, the selected evidence evolves with the forecasting horizon. For H=96, contributions are concentrated near the prediction boundary, indicating that short-term forecasting mainly relies on recent observations. As the horizon increases to H=720, the contribution distribution gradually expands toward more distant historical patches while remaining sparse, suggesting that PHL adaptively enlarges the effective historical context for longer-term prediction. These observations provide additional evidence that PHL learns horizon-specific patch contributions instead of a shared historical representation.



(a) H=96



(b) H=720

Figure A2: Comparison of learned patch contribution evidence on ETTh1 under different forecasting horizons. (a) H=96 and (b) H=720. Each subfigure contains the forecast (top) and the corresponding patch contribution evidence (bottom). Longer forecasting horizons require a broader historical context while preserving sparse patch selection.