PEAKTS: FORECASTING HIGH-VALUE REGIONS VIA RISK-CONDITIONED DISTRIBUTION CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series forecasting models often achieve strong average performance yet remain unreliable in critical high-value regions, where rare large values are particularly difficult to predict accurately. A key limitation is that existing methods typically lack explicit mechanisms for modeling historical patterns associated with future high-value outcomes. To address this issue, we propose a probability-distribution correction framework, **PEAKTS**, for high-value time-series forecasting. **PEAKTS** predicts an ordered categorical distribution over future value bins and derives risk-sensitive statistics, including the distribution mean and upper quantiles, to adaptively correct deterministic forecasts. We further introduce a Risk-Conditioned Peak Memory module that stores general and peak-precursor prototypes and retrieves relevant patterns from time-, difference-, and frequency-domain views to modulate the predictive distribution. To better learn sparsely represented high-value regions, we employ a tail-aware ordinal objective with frequency-based threshold reweighting and upper-tail weighting. Experiments on multiple forecasting benchmarks demonstrate consistent improvements in high-value forecasting while maintaining competitive overall forecasting performance. Our code is available at **Anonymous Github**.

1 INTRODUCTION

High-value forecasting is a critical problem in time-series prediction, as many real-world applications care not only about average accuracy but also about whether future peaks can be correctly anticipated. Examples include peak electricity demand, traffic congestion, financial risk, and abnormal sensor monitoring, where rare high-value outcomes often have disproportionate practical impact (Zhou et al., 2021; Huang et al., 2026; Wang et al., 2026b). However, accurately predicting such values remains challenging because high-value events are usually sparse, irregular, and embedded in highly imbalanced temporal distributions (Yang et al., 2021; Camacho & Bacao, 2024; Huang et al., 2026).

Most existing forecasting models are designed to optimize average prediction accuracy over all future steps and variables, commonly using point-wise losses such as MSE or MAE (Zeng et al., 2023; Nie et al., 2023; Wang et al., 2024). While these objectives are effective for general forecasting, they are dominated by the more frequent low- and medium-value observations and provide limited emphasis on rare high-value regions (Yang et al., 2021; Ren et al., 2022; Avelino et al., 2024; Wang et al., 2026b). Consequently, forecasting models that perform well on average may still struggle to accurately capture rare high-value observations. High-value windows are often preceded by informative temporal patterns, such as rapid upward trends, increased volatility, or changes in frequency-domain structures (Li et al., 2025; Huang et al., 2026). However, standard forecasting models usually encode input sequences into generic representations, without explicitly distinguishing ordinary contexts from peak-indicative contexts or reusing historical patterns that are predictive of high-value outcomes. Consequently, a model may achieve strong overall performance while still failing to recognize input conditions that signal upcoming high-value events.

To address these limitations, we propose **PEAKTS**, a risk-conditioned distribution correction framework for high-value time-series forecasting. Rather than relying solely on deterministic point predictions, **PEAKTS** models an ordered categorical distribution over future value bins, from which risk-sensitive statistics, such as the distributional mean and upper quantiles, are derived to refine the raw forecast. This distribution-guided correction allows the model to adaptively adjust its predic-

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.

tions when the estimated future distribution indicates elevated high-value risk. To further capture informative historical precursors of future peaks, we introduce a **Risk-Conditioned Peak Memory (RCPM)** module. RCPM maintains separate prototype banks for general temporal contexts and peak-precursor patterns, and retrieves relevant prototypes from complementary time-, difference-, and frequency-domain representations. The retrieved patterns are then used to adaptively modulate the predictive distribution according to the current risk context, enabling **PEAKTS** to better recognize and respond to conditions associated with future high-value outcomes.

This paper makes four main contributions:

- We propose **PEAKTS**, a distribution-guided forecasting framework that leverages ordered predictive distributions to refine deterministic forecasts for high-value regions.
- We introduce RCPM, which retrieves general and peak-precursor prototypes from multiple temporal views to enhance peak-aware distribution modeling.
- We develop a tail-risk learning objective that strengthens supervision for sparsely represented high-value regions through ordinal and peak-aware regularization.
- Extensive experiments demonstrate that **PEAKTS** consistently improves high-value forecasting performance while maintaining competitive overall accuracy.

2 Related Work

Time-Series Forecasting Deep time-series forecasting has progressed from attention-based long-range modeling to architectures that more explicitly capture multiscale temporal structures, cross-variable dependencies, and heterogeneous predictive patterns. Early Transformer-based methods such as Informer (Zhou et al., 2021) and Autoformer (Wu et al., 2021) improve long-range forecasting through sparse attention and decomposition, while DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024), and TimeMixer (Wang et al., 2024) explore linear modeling, patch-wise representations, variate-wise attention, and multiscale mixing. More recent studies further investigate frequency-decomposed representations with TimeKAN (Huang et al., 2025), long-context unified forecasting with Timer-XL (Liu et al., 2025b), non-stationarity-aware dependency modeling with TimeBridge (Liu et al., 2025a), retrieval-augmented forecasting with RAFT (Han et al., 2025), selective patch representation spaces with SRSNet (Wu et al., 2025), and feature-aware temporal modeling with FeTS (Wang et al., 2026d). These advances primarily target general forecasting performance across the full prediction horizon, whereas our work focuses on balancing strong overall accuracy with improved prediction in high-value regions.

High-Value and Extreme Time-Series Forecasting Forecasting rare high-value or extreme observations has received comparatively less attention. Ding et al. (Ding et al., 2019) combine an extreme-value-inspired loss with a memory network to better model extreme events in time series. DeepExtrema (Galib et al., 2022) integrates neural networks with the generalized extreme value distribution to forecast block maxima, while DeepGPD (Wilson et al., 2022) combines deep learning with the generalized Pareto distribution for geospatio-temporal extremes. More recently, TXtreme (Yadav & Thakkar, 2025) explicitly distinguishes normal and extreme regimes, and APAL (Hari et al., 2026) introduces asymmetric peak-aware weighting for peak-critical forecasting. M2FMoE (Huang et al., 2026) combines frequency-domain mixtures of experts in the Fourier and wavelet domains with multi-resolution fusion to capture both regular and extreme patterns. Different from these approaches, **PEAKTS** combines an ordered predictive distribution with risk-conditioned prototype retrieval, using historical peak precursors to modulate the distribution and subsequently guide value correction for high-value forecasting.

3 Method

Let $x_t \in \mathbb{R}^C$ denote the observation at time step t, where C is the number of variables. Given a sequence of past observations within a look-back window of length T, denoted by $\mathbf{X} = \{x_{t-T+1}, x_{t-T+2}, \ldots, x_t\} \in \mathbb{R}^{T\times C}$, the goal of the forecasting model is to predict the future sequence $\hat{\mathbf{Y}} = \{\hat{x}_{t+1}, \hat{x}_{t+2}, \ldots, \hat{x}_{t+P}\} \in \mathbb{R}^{P\times C}$, where P denotes the forecasting horizon. Beyond overall forecasting performance, **we aim to improve forecasting accuracy in high-value**

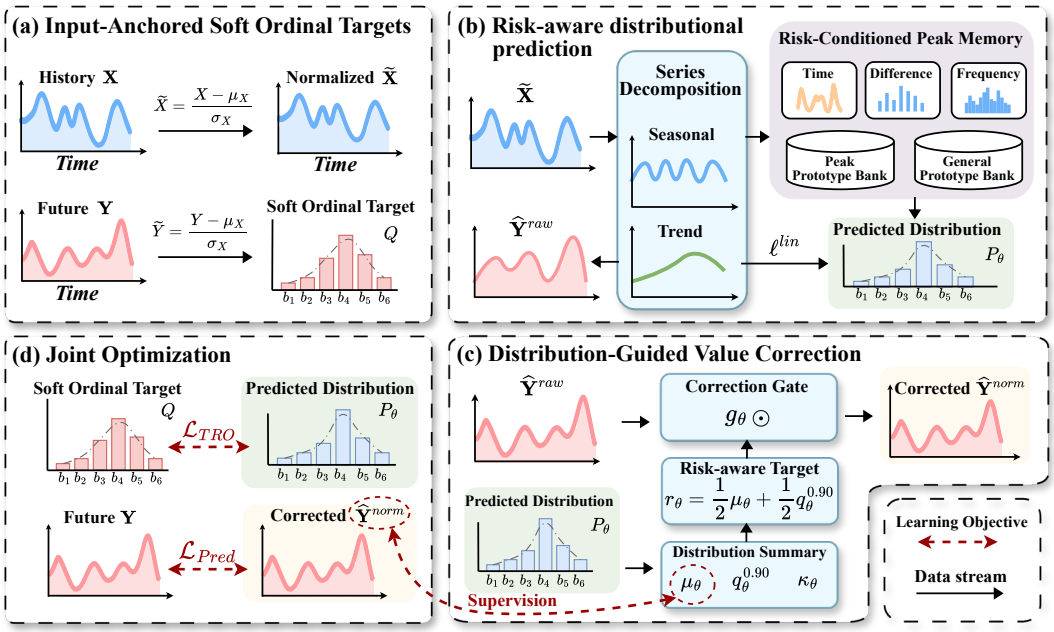


Figure 1: Overview of **PEAKTS**. (a) Normalizes the input and constructs soft ordinal targets to provide scale-consistent distributional supervision. (b) Predicts the raw forecast and risk-aware distribution to capture temporal patterns and peak-related information. (c) Corrects the raw forecast using distribution-derived risk statistics to reduce high-value underestimation. (d) Jointly optimizes distribution and value predictions to improve both peak awareness and forecasting accuracy.

regions while maintaining strong overall predictive performance, where rare but important observations often have disproportionate practical impact. As shown in Figure 1, **PEAKTS** consists of three main components: a distributional temporal predictor, a risk-conditioned peak memory module, and a tail-aware training objective.

3.1 INPUT-ANCHORED NORMALIZATION AND SOFT ORDINAL TARGETS

We adopt an input-anchored normalization strategy to keep the target representation consistent with the input scale. For each sample and variable, normalization statistics are computed only from the observed look-back window $\mathbf{X}$. Both the input sequence $\mathbf{X}$ and the future target sequence $\mathbf{Y}$ are then mapped to this same scale, yielding $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$.

The normalized targets are converted into soft distributions over K ordered bins. Let the bin-edge support be $\mathcal{B} = \{b_0, \ldots, b_K\}$, initialized uniformly on $[-1, 1]$ and slightly extended at both ends by half a bin width. For a normalized scalar target $\tilde{Y}_{p,c}$ from $\tilde{\mathbf{Y}}$, where $p = 1, \ldots, P$ and $c = 1, \ldots, C$, the target probability assigned to bin k is defined as

$$q_{p,c,k} = \frac{\Phi((b_{k+1} - \tilde{Y}_{p,c})/\sigma) - \Phi((b_k - \tilde{Y}_{p,c})/\sigma)}{\Phi((b_K - \tilde{Y}_{p,c})/\sigma) - \Phi((b_0 - \tilde{Y}_{p,c})/\sigma) + \varepsilon}, \quad k = 0, \ldots, K-1, \tag{1}$$

where Φ denotes the Gaussian cumulative distribution function, σ controls the degree of label smoothing, and ε is a small constant for numerical stability. This operation produces an ordered soft target distribution $\mathbf{q}_{p,c} \in \mathbb{R}^K$ for each future step and variable.

3.2 DISTRIBUTIONAL TEMPORAL PREDICTOR

We first introduce the temporal modeling module that serves as the forecasting backbone of **PEAKTS**. Given the normalized input $\tilde{\mathbf{X}}$, we use a decomposition-based temporal backbone to extract seasonal and trend representations from $\tilde{\mathbf{X}}$, denoted as $\mathbf{X}^s$ and $\mathbf{X}^t$. These representations are fed into two

parallel branches. The probability branch produces initial logits over K ordered bins, while the value branch produces a raw normalized forecast:

$$\boldsymbol{\ell}^{\text{lin}} = W_s^p \mathbf{X}^s + W_t^p \mathbf{X}^t \in \mathbb{R}^{P\times C\times K}, \qquad \hat{\mathbf{Y}}^{\text{raw}} = W_s^v \mathbf{X}^s + W_t^v \mathbf{X}^t \in \mathbb{R}^{P\times C}. \tag{2}$$

where $\boldsymbol{\ell}^{\text{lin}}$ denotes the initial probability logits, and $\hat{\mathbf{Y}}^{\text{raw}}$ denotes the direct normalized value prediction. The initial logits $\boldsymbol{\ell}^{\text{lin}}$ are subsequently modulated by the risk-conditioned peak memory to incorporate peak-related historical information.

3.3 Risk-conditioned Peak Memory for Distributional Modulation

High-value future windows are often preceded by distinctive temporal patterns. To capture such peak-related precursors, we introduce a Risk-Conditioned Peak Memory (RCPM) module, which retrieves historical patterns from general and peak prototype banks to modulate the initial probability logits $\boldsymbol{\ell}^{\text{lin}}$. RCPM maintains two learnable prototype banks: a general bank for ordinary contexts and a peak bank for patterns associated with future high-value windows. The two banks are initialized using a training-set future-maximum threshold u_ρ. Given a normalized input sequence, RCPM encodes each variable from three complementary views: time, difference, and frequency domains:

$$\mathbf{e}_{c,1} = E_{\text{time}}(\tilde{\mathbf{X}}_{:,c}), \qquad \mathbf{e}_{c,2} = E_{\text{diff}}(\Delta\tilde{\mathbf{X}}_{:,c}), \qquad \mathbf{e}_{c,3} = E_{\text{freq}}(|\mathcal{F}(\tilde{\mathbf{X}}_{:,c})|). \tag{3}$$

The three views capture complementary peak-related information: the time view represents overall level and shape, the difference view highlights local variations, and the frequency view captures periodic patterns. RCPM combines these view embeddings with route statistics $\mathbf{s}_c$, including level, slope, volatility, and peak gap, to learn adaptive view weights $\boldsymbol{\pi}_c$. For each view, relevant prototypes are then retrieved from both the general and peak banks via top-N_{ret} similarity search:

$$\mathbf{m}_{c,v}^b = \sum_{m\in \text{TopN}_{\text{ret}}(\mathbf{e}_{c,v}^\top \mathbf{k}_{v,m}^b)} \alpha_{c,v,m}^b \mathbf{v}_{v,m}^b, \qquad b \in \{\text{gen}, \text{peak}\}. \tag{4}$$

Here, $\mathbf{k}_{v,m}^b$ and $\mathbf{v}_{v,m}^b$ denote the key and value of the m-th prototype in bank b for view v, and $\alpha_{c,v,m}^b$ is the attention weight over the selected prototypes. To make retrieval sensitive to high-value risk, RCPM further predicts a coarse peak risk score $\hat{r}_c$ from the multi-view context. This risk score is combined with the similarity to peak prototypes, denoted by $c_{c,\text{peak}}$, to compute the general/peak bank weights:

$$[\omega_{c,\text{gen}}, \omega_{c,\text{peak}}] = \text{softmax}\left(H([\mathbf{e}_{c,1}, \mathbf{e}_{c,2}, \mathbf{e}_{c,3}, \mathbf{s}_c]) + [0, \alpha_{\text{sim}} c_{c,\text{peak}} + \alpha_{\text{risk}} \,\text{logit}\, \hat{r}_c]\right). \tag{5}$$

When the current input is close to peak prototypes or receives a high predicted peak risk, the peak bank receives a larger weight; otherwise, the retrieval relies more on the general bank. The retrieved contents are then aggregated across both views and banks to form the final memory content for variable c:

$$\mathbf{m}_{c,\text{mem}} = \sum_{v=1}^{3} \pi_{c,v} \left(\omega_{c,\text{gen}} \mathbf{m}_{c,v}^{\text{gen}} + \omega_{c,\text{peak}} \mathbf{m}_{c,v}^{\text{peak}}\right). \tag{6}$$

This memory content is transformed into two input-dependent modulation terms: a bias term and a scale term. Concretely, $\mathbf{m}_{c,\text{mem}}$ is split into a raw bias $\mathbf{B}_{c,\text{raw}}$ and a raw scale $\mathbf{R}_{c,\text{scale}}$, while a lightweight gate γ_c controls the strength of memory intervention:

$$\mathbf{B}_{c,\text{mem}} = \gamma_c \mathbf{B}_{c,\text{raw}}, \qquad \mathbf{S}_{c,\text{mem}} = 1 + \alpha_s \gamma_c \tanh(\mathbf{R}_{c,\text{scale}}). \tag{7}$$

Here, α_s controls the scale modulation strength. After stacking all variables, we obtain $\mathbf{S}_{\text{mem}}, \mathbf{B}_{\text{mem}} \in \mathbb{R}^{P\times C\times K}$, which are used to modulate the initial probability logits:

$$\boldsymbol{\ell}_\theta = \boldsymbol{\ell}^{\text{lin}} \odot \mathbf{S}_{\text{mem}} + \mathbf{B}_{\text{mem}}. \tag{8}$$

The memory-enhanced logits $\boldsymbol{\ell}_\theta$ are then normalized along the bin dimension to obtain the ordered predictive distribution $\mathbf{p}_\theta \in \mathbb{R}^{P\times C\times K}$. Each element $p_{\theta,p,c,k}$ denotes the predicted probability that the normalized future value $\tilde{Y}_{p,c}$ falls into the k-th bin $[b_k, b_{k+1})$. Through this modulation, RCPM reshapes the predictive distribution according to retrieved peak-related patterns, making the probability branch more sensitive to high-value risks.

3.4 DISTRIBUTION-GUIDED VALUE CORRECTION

After obtaining the memory-refined predictive distribution, we use its upper-tail information to refine the raw deterministic forecast. While $\hat{\mathbf{Y}}^{\mathrm{raw}}$ provides a stable point estimate, it may insufficiently capture large target values in high-value regions. Therefore, we derive risk-sensitive statistics from the predictive distribution to guide the final value prediction.

We first summarize the ordered predictive distribution into continuous statistics. Let $m_k = (b_k + b_{k+1})/2$ denote the center of the k-th bin. The distributional mean is computed as $\mu_{\theta,p,c} = \sum_{k=0}^{K-1} p_{\theta,p,c,k} m_k$. In addition to the mean, we extract an upper quantile $q^{0.90}_{\theta,p,c}$ from the cumulative distribution by linear interpolation within the first bin whose cumulative probability exceeds 0.90. This upper quantile serves as a risk-sensitive estimate of potential high-value outcomes. We also compute the normalized entropy confidence

$$\kappa_{\theta,p,c} = 1 + \frac{\sum_{k=0}^{K-1} p_{\theta,p,c,k} \log(p_{\theta,p,c,k} + \varepsilon)}{\log K}, \tag{9}$$

where a larger $\kappa_{\theta,p,c}$ indicates a sharper and more confident predictive distribution. Based on these distributional summaries, a lightweight correction gate determines how much the probability branch should adjust the raw forecast. For each variable c, the gate takes the normalized input sequence, the distributional mean, the upper quantile, and the confidence as input:

$$\mathbf{g}_{\theta,:,c} = \sigma\left(G\left([\tilde{\mathbf{X}}_{:,c}, \boldsymbol{\mu}_{\theta,:,c}, \mathbf{q}^{0.90}_{\theta,:,c}, \boldsymbol{\kappa}_{\theta,:,c}]\right)\right), \tag{10}$$

where $G(\cdot)$ is a gating network. We then define a risk-aware distributional target by combining the distributional mean and the upper quantile: $\mathbf{r}_\theta = \frac{1}{2}\boldsymbol{\mu}_\theta + \frac{1}{2}\mathbf{q}^{0.90}_\theta$. The final normalized forecast is obtained by moving the raw prediction toward this risk-aware target:

$$\hat{\mathbf{Y}}^{\mathrm{norm}} = \hat{\mathbf{Y}}^{\mathrm{raw}} + \alpha \mathbf{g}_\theta \odot \left(\mathbf{r}_\theta - \hat{\mathbf{Y}}^{\mathrm{raw}}\right), \tag{11}$$

where $\alpha = 0.8$ controls the correction strength. This formulation preserves the stability of point forecasting while allowing the model to adaptively refine predictions when the predictive distribution indicates elevated upper-tail risk.

Proposition 3.1. *For an arbitrary prediction step and variable, consider the correction in Eq. (11), where $\lambda = \alpha g_\theta \in [0, 1]$. If the raw prediction lies below the normalized ground truth and the risk-aware distributional target remains between the raw prediction and the ground truth, i.e., $\hat{y}^{\mathrm{raw}} \le r_\theta \le \tilde{y}$, then*

$$|\tilde{y} - \hat{y}^{\mathrm{norm}}| \le |\tilde{y} - \hat{y}^{\mathrm{raw}}|. \tag{12}$$

3.5 TAIL-RISK DISTRIBUTIONAL LEARNING OBJECTIVE

To train **PEAKTS** for both accurate value prediction and high-value sensitivity, we introduce a **Tail-Risk Distributional Learning** (TRDL) objective. It combines a tail-risk ordinal loss for supervising the ordered predictive distribution with value-level losses for optimizing the final corrected forecast. Given the predicted ordered distribution $\mathbf{p}_\theta \in \mathbb{R}^{P\times C\times K}$ and the Gaussian-soft target distribution $\mathbf{q} \in \mathbb{R}^{P\times C\times K}$, we compare their cumulative probabilities over ordered thresholds. For each prediction step p and variable c, the cumulative probabilities up to threshold t are

$$P_{\theta,p,c,t} = \sum_{k=0}^{t} p_{\theta,p,c,k}, \qquad Q_{p,c,t} = \sum_{k=0}^{t} q_{p,c,k}, \qquad t = 0, \ldots, K-2. \tag{13}$$

This cumulative formulation preserves the ordinal structure of value bins and measures whether the model assigns appropriate probability mass below or above each threshold. To emphasize high-value regions, we estimate a data-adaptive exceedance frequency η_t for each threshold from the normalized training targets. Rarely exceeded upper thresholds receive larger weights:

$$w_t = \frac{(\eta_t + \epsilon_f)^{-\beta}}{\frac{1}{K-1}\sum_{j=0}^{K-2}(\eta_j + \epsilon_f)^{-\beta}}, \qquad \eta_t = \Pr_{\tilde{\mathbf{Y}}\sim\mathcal{D}_{\mathrm{tr}}}\left(\tilde{Y} > b_{t+1}\right), \tag{14}$$

where ϵ_f is a small constant and β controls the strength of frequency-aware reweighting. The Tail-Risk Ordinal Loss is then defined as

$$\mathcal{L}_{\mathrm{TRO}} = \frac{1}{PC\sum_t w_t} \sum_{p=1}^{P} \sum_{c=1}^{C} \sum_{t=0}^{K-2} w_t a_{p,c,t} \ell_{\mathrm{BCE}} \left(P_{\theta,p,c,t}, Q_{p,c,t} \right), \tag{15}$$

where $\ell_{\mathrm{BCE}}(\cdot,\cdot)$ is the binary cross-entropy between cumulative probabilities. The optional factor $a_{p,c,t}$ introduces asymmetric emphasis on upper-tail distribution errors:

$$a_{p,c,t} = 1 + \alpha_{\mathrm{under}} \mathbb{I}[w_t > 1] \left[P_{\theta,p,c,t} - Q_{p,c,t} \right]_+ . \tag{16}$$

For upper thresholds, $P_{\theta,p,c,t} > Q_{p,c,t}$ indicates that the predicted distribution assigns excessive probability mass below the corresponding threshold, motivating stronger supervision in the upper tail. Setting $\alpha_{\mathrm{under}} = 0$ recovers the symmetric ordinal loss. Finally, the full TRDL objective jointly optimizes the ordered distribution and the corrected value forecast:

$$\mathcal{L}_{\mathrm{TRDL}} = \left\| \hat{\mathbf{Y}}^{\mathrm{norm}} - \tilde{\mathbf{Y}} \right\|_1 + \lambda_{\mathrm{ord}} \mathcal{L}_{\mathrm{TRO}} + \lambda_{\mathrm{cons}} \left\| \hat{\mathbf{Y}}^{\mathrm{norm}} - \mathrm{sg}(\boldsymbol{\mu}_\theta) \right\|_1 . \tag{17}$$

Here, $\mathrm{sg}(\cdot)$ denotes stop-gradient. The ordinal loss trains the predictive distribution, the value loss directly supervises the final corrected forecast, and the consistency loss keeps the corrected prediction aligned with the probability-derived mean.

4 Experiment

4.1 Experiment Setup

Datasets We evaluate **PeakTS** on nine real-world time-series benchmarks: ETTm1, ETTm2, ETTh1, ETTh2, Weather, Electricity, AQWan, CzeLan, and ZafNoo. These datasets cover commonly used forecasting scenarios, including energy and environmental measurements. For standard forecasting evaluation, we use a fixed look-back window of $T = 336$ and consider four prediction horizons, $P \in \{96, 192, 336, 720\}$. For high-value-region evaluation, we fix the prediction horizon to $P = 96$ and evaluate forecasting performance over the Top-25%, Top-50%, Top-75%, and full target regions.

Baseline We compare **PeakTS** against 11 competitive forecasting methods: DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024), TimeMixer (Wang et al., 2024), TimeFilter (Hu et al., 2025), TimeEmb (Xia et al., 2025), SEED (Xiong et al., 2026), OCE-TS Wang et al. (2026c), M2FMoE (Huang et al., 2026), DistDF (Wang et al., 2026a), and TeCh (Yu et al., 2026). The baselines include both widely used long-term forecasting models and recent approaches that strengthen temporal representation learning from different perspectives.

Implementation Details All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB memory, with all models implemented in PyTorch and trained using the Adam optimizer. Model performance is evaluated using both MSE and MAE. To avoid discarding incomplete mini-batches, we set `drop_last=False` throughout training and evaluation. The initial learning rate is chosen from $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$, and the hidden dimension is chosen from $\{32, 64, 128, 256, 512\}$ according to validation performance.

4.2 Main Results

Due to space limitations, we report the results averaged over different prediction horizons, while the complete results are provided in Table 4.

As shown in Figure 2, our method achieves strong overall performance, with particularly consistent gains in high-value regions, with particularly clear improvements in MAE. Existing methods primarily optimize point-wise accuracy and typically pay limited attention to distributional characteristics and rare high-value outcomes, which can result in degraded accuracy in high-value regions. This improvement can be attributed to the probability-guided correction mechanism, which uses distributional information to adaptively refine deterministic forecasts. In addition, the risk-conditioned peak memory captures informative historical patterns associated with high-value outcomes, allowing the

Table 1: Forecasting performance across different value regions.

Dataset	Region	DLinear		PatchTST		iTransformer		TimeMixer		TimeFilter		TimeEmb		SEED		OCE-TS		M2FMoE		DistDF		TeCh		Ours	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	25%	0.410	0.405	0.406	0.419	0.388	0.404	0.378	0.398	0.375	0.396	0.371	0.396	0.378	0.395	0.688	0.611	0.421	0.423	0.414	0.409	0.590	0.554	**0.361**	**0.378**
	50%	0.340	0.363	0.345	0.382	0.331	0.370	0.331	0.370	0.328	0.365	0.329	0.369	0.328	0.362	0.609	0.568	0.379	0.391	0.346	0.366	0.463	0.477	**0.309**	**0.345**
	75%	0.318	0.352	0.321	0.369	0.318	0.365	0.317	0.363	0.312	0.356	0.312	0.361	0.312	0.354	0.572	0.548	0.355	0.380	0.326	0.357	0.421	0.454	**0.293**	**0.337**
	All	0.300	0.344	0.301	0.358	0.304	0.358	0.306	0.359	0.298	0.350	0.299	0.355	0.298	0.348	0.545	0.536	0.348	0.375	0.308	0.349	0.382	0.431	**0.282**	**0.331**
ETTm2	25%	0.209	0.305	0.210	0.314	0.225	0.322	0.210	0.307	0.214	0.311	0.229	0.324	0.211	0.306	2.314	0.979	0.247	0.342	0.210	0.304	0.446	0.547	**0.200**	**0.289**
	50%	0.164	0.267	0.168	0.276	0.179	0.280	0.166	0.265	0.167	0.269	0.178	0.279	0.164	0.263	2.516	0.996	0.193	0.298	0.164	0.261	0.359	0.485	**0.156**	**0.248**
	75%	0.155	0.256	0.162	0.267	0.166	0.266	0.160	0.253	0.158	0.257	0.170	0.266	0.154	0.250	2.610	1.008	0.183	0.286	0.154	0.248	0.319	0.455	**0.150**	**0.238**
	All	0.167	0.262	0.181	0.276	0.181	0.274	0.176	0.261	0.170	0.262	0.182	0.271	0.167	0.256	2.685	1.029	0.189	0.286	0.164	0.253	0.307	0.437	**0.162**	**0.245**
ETTh1	25%	0.479	0.436	0.508	0.467	0.481	0.447	0.488	0.442	0.484	0.444	0.485	0.450	0.482	0.440	0.712	0.674	0.624	0.529	0.472	0.431	0.645	0.565	**0.471**	**0.425**
	50%	0.447	0.432	0.476	0.462	0.463	0.447	0.468	0.441	0.462	0.443	0.470	0.452	0.462	0.439	0.760	0.703	0.580	0.516	0.446	0.431	0.595	0.550	**0.441**	**0.423**
	75%	0.410	0.413	0.435	0.439	0.429	0.431	0.426	0.422	0.425	0.425	0.431	0.433	0.422	0.420	0.677	0.651	0.540	0.495	0.411	0.413	0.545	0.525	**0.405**	**0.406**
	All	0.373	0.396	0.394	0.418	0.394	0.415	0.390	0.406	0.389	0.408	0.397	0.418	0.385	0.403	0.623	0.621	0.507	0.482	0.376	0.397	0.497	0.504	**0.368**	**0.389**
ETTh2	25%	0.283	0.381	0.327	0.420	0.390	0.450	0.339	0.413	0.342	0.418	0.356	0.432	0.351	0.426	1.251	0.809	0.401	0.462	0.296	0.384	0.926	0.756	**0.271**	**0.363**
	50%	0.305	0.376	0.340	0.405	0.375	0.417	0.345	0.394	0.337	0.395	0.351	0.407	0.351	0.403	1.277	0.810	0.392	0.441	0.301	0.367	1.317	0.928	**0.289**	**0.352**
	75%	0.249	0.334	0.282	0.363	0.292	0.356	0.273	0.341	0.264	0.340	0.280	0.353	0.273	0.343	1.200	0.776	0.326	0.397	0.237	0.316	1.339	0.956	**0.231**	**0.306**
	All	0.291	0.357	0.321	0.382	0.327	0.374	0.316	0.362	0.303	0.360	0.318	0.371	0.307	0.360	1.244	0.785	0.373	0.418	0.277	0.337	1.521	1.023	**0.273**	**0.330**
Weather	25%	0.317	0.277	0.302	0.264	0.311	0.257	0.303	0.249	0.294	0.250	0.299	0.251	0.290	0.246	42.87	2.722	**0.283**	0.256	0.318	0.269	0.299	0.269	0.311	**0.241**
	50%	0.248	0.268	0.220	0.243	0.228	0.238	0.227	0.234	0.214	0.231	0.215	0.232	0.211	0.229	39.91	2.633	**0.210**	0.244	0.250	0.263	0.215	0.247	0.226	**0.222**
	75%	0.204	0.250	0.179	0.226	0.186	0.219	0.184	0.214	0.175	0.215	0.175	0.215	**0.173**	0.214	38.06	2.603	0.177	0.233	0.203	0.238	0.178	0.233	0.183	**0.204**
	All	0.174	0.236	0.153	0.213	0.159	0.207	0.154	0.202	0.150	0.204	0.149	0.203	**0.147**	0.202	36.91	2.580	0.154	0.219	0.175	0.225	0.151	0.220	0.155	**0.192**
Electricity	25%	0.185	0.269	0.180	0.263	0.183	0.270	0.183	0.265	0.220	0.318	0.180	0.268	0.181	0.261	0.844	0.573	0.185	0.272	0.185	0.268	0.247	0.347	**0.177**	**0.259**
	50%	0.167	0.254	0.163	0.249	0.170	0.260	0.166	0.249	0.197	0.286	0.169	0.257	0.165	0.247	0.824	0.561	0.169	0.258	0.168	0.254	0.243	0.349	**0.162**	**0.246**
	75%	0.153	0.247	0.149	0.241	0.155	0.252	0.158	0.249	0.183	0.274	0.153	0.249	0.151	0.240	0.825	0.557	0.155	0.251	0.154	0.246	0.225	0.338	**0.147**	**0.239**
	All	0.140	0.238	0.136	0.232	0.143	0.244	0.145	0.240	0.168	0.265	0.140	0.241	**0.134**	**0.230**	0.819	0.551	0.142	0.242	0.141	0.237	0.215	0.333	0.138	0.231
AQWan	25%	1.519	0.538	1.594	0.546	1.506	0.529	1.521	0.527	1.524	0.532	1.515	0.524	1.537	0.525	3.119	1.045	1.684	0.596	1.510	0.529	1.580	0.562	**1.504**	**0.515**
	50%	0.990	0.498	1.018	0.495	0.987	0.489	0.992	0.488	1.002	0.495	0.990	0.487	1.001	0.486	2.850	1.013	1.103	0.538	0.992	0.491	1.019	0.512	**0.984**	**0.476**
	75%	0.847	0.501	0.843	0.489	0.844	0.493	0.850	0.491	0.854	0.495	0.847	0.490	0.846	0.485	2.807	1.031	0.922	0.530	0.855	0.495	0.848	0.508	**0.840**	**0.478**
	All	0.719	0.478	0.716	0.468	0.716	0.470	0.720	0.468	0.724	0.472	0.717	0.466	0.716	0.462	2.728	1.027	0.808	0.506	0.726	0.472	0.720	0.487	**0.710**	**0.455**
CzeLan	25%	0.413	0.383	1.022	0.598	0.364	0.315	0.372	0.312	**0.361**	0.313	0.371	0.315	0.365	0.310	19.36	2.044	0.543	0.483	0.371	0.311	0.964	0.606	**0.361**	**0.294**
	50%	0.292	0.334	0.846	0.531	0.247	0.267	0.247	0.261	0.242	0.262	0.251	0.264	0.243	0.259	19.45	1.992	0.410	0.427	0.246	0.259	0.810	0.554	**0.241**	**0.233**
	75%	0.244	0.312	0.764	0.497	0.204	0.249	0.198	0.239	**0.195**	0.243	0.200	0.240	**0.195**	0.240	19.11	1.953	0.346	0.398	0.198	0.238	0.736	0.526	**0.195**	**0.214**
	All	0.227	0.305	0.769	0.498	0.183	0.240	0.181	0.240	0.177	0.235	0.180	0.231	0.178	0.232	18.58	1.938	0.325	0.391	0.179	0.230	0.727	0.517	**0.175**	**0.208**
ZafNoo	25%	0.633	0.477	0.660	0.486	0.642	0.474	0.627	0.461	0.617	0.464	**0.614**	0.456	**0.614**	0.461	2.737	1.059	0.644	0.492	0.628	0.465	0.659	0.501	0.646	**0.451**
	50%	0.558	0.458	0.579	0.470	0.565	0.456	0.555	0.446	0.554	0.453	0.549	0.442	**0.547**	0.447	2.935	1.099	0.570	0.477	0.560	0.452	0.582	0.486	0.574	**0.437**
	75%	0.485	0.432	0.494	0.439	0.496	0.434	0.486	0.424	0.489	0.432	0.478	0.418	**0.477**	0.422	3.219	1.122	0.501	0.454	0.498	0.432	0.495	0.453	0.500	**0.411**
	All	0.435	0.412	0.442	0.418	0.450	0.415	0.440	0.406	0.441	0.413	0.432	0.399	**0.427**	0.401	3.251	1.110	0.455	0.437	0.448	0.413	0.440	0.428	0.451	**0.391**

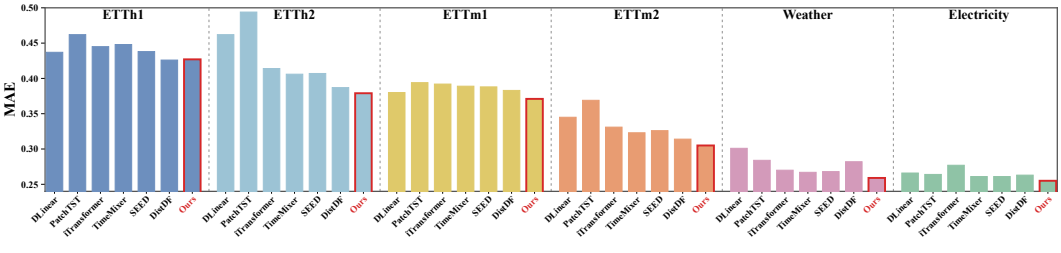


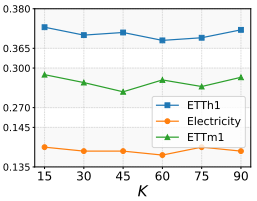
Figure 2: Overall performance is compared using the average results across four prediction lengths, with the complete results reported in Table 4.

model to make more appropriate corrections under different temporal conditions. We note that the relatively poor performance of OCE-TS results from our leakage-free reimplementation, where its original data processing procedure was modified to avoid information leakage and ensure a fair comparison under the same experimental protocol. Overall, these results demonstrate that the proposed framework improves forecasting accuracy and robustness while maintaining competitive forecasting performance.

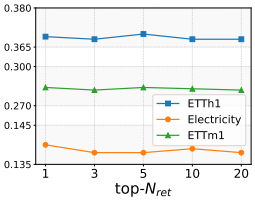
Table 1 further reports forecasting performance across four evaluation regions (Top-25%, Top-50%, Top-75%, and All) with $P = 96$, where the Top-x% subsets are defined by ranking all ground-truth future observations in the test set by value. Overall, forecasting errors generally increase as the evaluation region narrows from all observations to the Top-25% subset, which indicates that aggregate performance can obscure the difficulty of predicting the upper tail. Moreover, existing methods exhibit noticeable performance difference across datasets and region thresholds, whereas our method remains consistently competitive, demonstrating its ability to capture high-value patterns under different region definitions. Notably, it also maintains strong performance over the full evaluation region, which shows that improved high-value forecasting does not compromise overall forecasting accuracy. This superiority benefits from the complementary roles of probability-guided correction, which emphasizes upper-tail deviations, and risk-conditioned peak memory, which leverages historical peak patterns to enhance high-value prediction.

Table 2: Ablation study of **PEAKTS**.

Variant	ETTm1				ETTh1				Electricity			
	Top 25%		All		Top 25%		All		Top 25%		All	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
w/o PGC	0.396	0.394	0.293	0.338	0.476	0.431	0.370	0.393	0.182	0.265	0.140	0.237
w/o RCPM	0.388	0.391	0.291	0.337	0.477	0.429	0.371	0.392	0.180	0.266	0.139	0.235
w/o Peak Bank	0.376	0.386	0.286	0.334	0.473	0.430	0.369	0.390	0.181	0.262	0.139	0.232
w/o Tail Reweighting	0.382	0.383	0.288	0.333	0.475	0.427	0.370	0.391	0.179	0.264	0.140	0.233
w/o Asymmetric Tail Shaping	0.370	0.381	0.284	0.332	0.472	0.428	0.369	0.390	0.178	0.260	0.138	0.232
w/o Consistency	0.368	0.380	0.285	0.333	0.474	0.426	0.369	0.388	0.180	0.261	0.138	0.231
PEAKTS	0.361	0.378	0.282	0.331	0.471	0.425	0.368	0.389	0.177	0.259	0.138	0.231

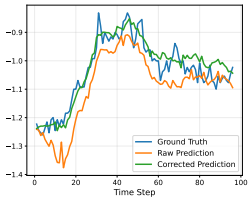


(a) Number of bins K.

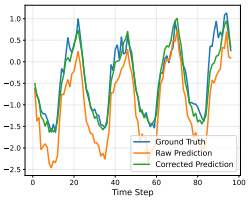


(b) # Retrieved prototypes.

Figure 3: Hyperparameter analysis.



(a) ETTm1 dataset.



(b) Electricity dataset.

Figure 4: Prediction correction analysis.

4.3 Ablation Study

To assess the contribution of each component in **PEAKTS**, we conduct ablation studies with the following variants: (a) '**w/o PGC**' removes Probability-Guided Correction and uses the raw deterministic prediction as the final forecast; (b) '**w/o RCPM**' removes the Risk-Conditioned Peak Memory module; (c) '**w/o Peak Bank**' removes the peak prototype bank; (d) '**w/o Tail Reweighting**' assigns equal weights to all ordinal thresholds; (e) '**w/o Asymmetric Tail Shaping**' removes the asymmetric upper-tail weighting by setting $\alpha_{\text{under}} = 0$; and (f) '**w/o Consistency**' removes the consistency loss.

Table 2 presents the ablation results, from which we make the following observations. (1) Removing PGC leads to a noticeable performance degradation, particularly in the Top-25% region, demonstrating that upper-tail distributional information provides effective guidance for high-value prediction correction. (2) Removing RCPM results in poorer performance in the Top-25% region, indicating that historical peak-related patterns provide valuable information for accurately forecasting high-value observations. (3) Removing the peak bank also degrades performance, which suggests that peak-specific prototypes capture information that cannot be fully represented by general prototypes. (4) Without tail reweighting, the Top-25% errors increase across all datasets, confirming the benefit of assigning greater importance to rare upper thresholds. (5) Removing asymmetric tail shaping mainly affects the Top-25% metrics, showing that asymmetric upper-tail supervision improves forecasting accuracy in high-value regions. (6) Removing the consistency loss results in lower high-value forecasting accuracy, although its effect on the overall metrics varies across datasets. Overall, the complete **PEAKTS** achieves the best overall performance on all datasets.

5 Hyperparameter Analysis

We analyze the sensitivity of **PEAKTS** to the number of bins K, the number of retrieved prototypes N_{ret}, and the loss weights λ_{ord} and λ_{cons}. Specifically, as shown in Figure 3a, $K = 45$ achieves the best performance on ETTm1, whereas $K = 60$ performs best on ETTh1 and Electricity. A finer discretization initially improves distribution modeling, but an excessively large K fragments the target distribution and complicates probability estimation. Figure 3b shows that $N_{\text{ret}} = 3$ provides the best overall performance, while further increases in N_{ret} yield no clear improvement because additional prototypes may be less relevant and introduce noise.

5.1 Effect of probability-guided prediction correction

Figure 4 demonstrates the effect of the probability-guided correction mechanism introduced in Section 3.4 on representative windows from ETTm1 and Electricity. On ETTm1, the correction brings the prediction closer to the ground truth around rising segments and local peaks. The improvement is more pronounced on Electricity, where the raw prediction exhibits noticeable deviations around several high-value regions, while the corrected prediction better captures both the periodic dynamics and peak magnitudes. These results illustrate that the proposed mechanism adaptively refines deterministic forecasts using the predictive mean, upper quantile, and confidence while preserving the temporal structure learned by the base predictor.

5.2 Analysis of Risk-Conditioned Memory Routing

Figure 5 shows how RCPM distributes weight between the general and peak memory banks. The general bank dominates in lower-value regions, while the peak-bank weight gradually increases toward the upper tail and eventually becomes dominant. The transition suggests that the router responds to different levels of peak-related evidence rather than relying on a fixed switching rule. This separation reflects the distinct roles of the two memory banks, with the general bank representing ordinary temporal contexts and the peak bank providing patterns associated with future high values. Overall, RCPM adapts its memory retrieval according to the estimated risk of each input.

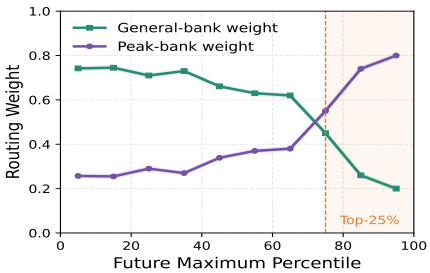


Figure 5: Risk-conditioned memory routing.

Figure 6 visualizes the learned prototypes and query-to-prototype retrieval behavior of RCPM. The general and peak prototype banks occupy largely distinct regions of the embedding space, which suggests that their prototype keys specialize in ordinary temporal contexts and peak-precursor patterns, respectively. The Bottom-25% and Top-25% queries also show different distributions, indicating that the multi-view query representations retain information related to future high-value risk. The retrieval paths further show that high-value queries retrieve peak prototypes when peak-related evidence is present. The retrieved content can therefore provide input-dependent peak information for modulating the predictive distribution, consistent with the risk-conditioned retrieval mechanism in RCPM.

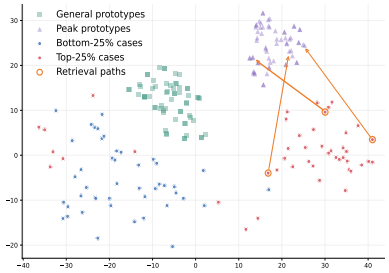


Figure 6: Prototype topology and retrieval paths learned by RCPM.

6 Conclusion

In this work, we study high-value time-series forecasting, where accurate prediction of rare but important large-value observations is required in addition to strong overall forecasting performance. We propose **PeakTS**, a risk-conditioned distribution correction framework that augments deterministic forecasting with an ordered predictive distribution. By retrieving general and peak-related historical prototypes through the Risk-Conditioned Peak Memory, **PeakTS** captures informative precursors of future high-value events and uses the resulting distributional information to adaptively correct conservative forecasts. We further introduce a tail-risk distributional learning objective to strengthen supervision for sparsely represented upper-tail regions. Extensive experiments on nine real-world forecasting benchmarks demonstrate that **PeakTS** consistently improves prediction accuracy in high-value regions while maintaining competitive overall forecasting performance.

Future work will extend **PeakTS** with adaptive value discretization and online prototype updating to support evolving temporal patterns. Another promising direction is to incorporate application-specific risk preferences for more flexible high-value forecasting in peak-critical scenarios.

REFERENCES

Juscimara G. Avelino, George D. C. Cavalcanti, and Rafael M. O. Cruz. Resampling strategies for imbalanced regression: A survey and empirical analysis. *Artificial Intelligence Review*, 57, 2024. doi: 10.1007/s10462-024-10724-3.

Lu'is Camacho and Fernando Bacao. WSMOTER: A novel approach for imbalanced regression. *Applied Intelligence*, 54(19):8789–8799, 2024. doi: 10.1007/s10489-024-05608-6.

Daizong Ding, Mi Zhang, Xudong Pan, Min Yang, and Xiangnan He. Modeling extreme events in time series prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, pp. 1114–1122, 2019. doi: 10.1145/3292500.3330896.

Asadullah Hill Galib, Andrew McDonald, Tyler Wilson, Lifeng Luo, and Pang-Ning Tan. Deepextrema: A deep learning approach for forecasting block maxima in time series data. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2980–2986, 2022. doi: 10.24963/ijcai.2022/413.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21774–21797. PMLR, 2025.

Theivaprakasham Hari, Yanan Xin, Winnie Daamen, Serge Paul Hoogendoorn, and Sascha Hoogendoorn-Lanser. Asymmetric peak-aware loss for peak-critical time series forecasting. *arXiv preprint arXiv:2607.14871*, 2026.

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. *arXiv preprint arXiv:2501.13041*, 2025.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wTLc79YNbh.

Yaohui Huang, Runmin Zou, Yun Wang, Laeeq Aslam, and Ruipeng Dong. M2FMoE: Multiresolution multi-view frequency mixture-of-experts for extreme-adaptive time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22075–22083, 2026. doi: 10.1609/aaai.v40i26.39362.

Quan Li, Wenchao Yu, Suhang Wang, Minhua Lin, Lingwei Chen, Wei Cheng, and Haifeng Chen. xTime: Extreme event prediction with hierarchical knowledge distillation and expert fusion. In *Proceedings of the 25th IEEE International Conference on Data Mining (ICDM)*, pp. 1380–1389, 2025. doi: 10.1109/ICDM65498.2025.00147.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39815–39840. PMLR, 2025a.

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*, 2024.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025b.

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.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pp. 1185–1196, 2025.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7926–7935, 2022. doi: 10.1109/CVPR52688.2022.00777.

Hao Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zi Chan, Qingsong Wen, Haoxuan Li, and Zhouchen Lin. Distdf: Time-series forecasting needs joint-distribution wasserstein alignment. In *International Conference on Learning Representations*, volume 2026, pp. 131512–131535, 2026a.

Jialou Wang, Jacob Sanderson, and Wai Lok Woo. Extreme-aware time-series forecasting via weak-label-guided mixture of experts. *Sensors*, 26(5):1571, 2026b. doi: 10.3390/s26051571.

Jieting Wang, Huimei Shi, Feijiang Li, and Xiaolei Shang. Beyond mse: Ordinal cross-entropy for probabilistic time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26269–26276, 2026c. doi: 10.1609/aaai.v40i31.39831.

Le Wang, Jianyong Chen, and Songbai Liu. Fets: A feature-aware framework for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26328–26336, 2026d. doi: 10.1609/aaai.v40i31.39838.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024.

Tyler Wilson, Pang-Ning Tan, and Lifeng Luo. Deepgpd: A deep learning approach for modeling geospatio-temporal extreme events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4245–4253, 2022. doi: 10.1609/aaai.v36i4.20344.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 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. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Mingyuan Xia, Chunxu Zhang, Zijian Zhang, Hao Miao, Qidong Liu, Yuanshao Zhu, and Bo Yang. Timeemb: A lightweight static-dynamic disentanglement framework for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 76227–76258, 2025.

Feng Xiong, Zongxia Xie, Yanru Sun, Haoyu Wang, and Jianhong Lin. Seed: Spectral entropy-guided evaluation of spatial-temporal dependencies for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27153–27161, 2026.

Hemant Yadav and Amit Thakkar. Txtreme: Transformer-based extreme value prediction framework for time series forecasting. *Discover Applied Sciences*, 7(2):98, 2025. doi: 10.1007/s42452-025-06478-4.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11842–11851. PMLR, 2021.

Guoqi Yu, Juncheng Wang, Chen Yang, Jing Qin, Angelica Aviles-Rivero, and Shujun Wang. Decentralized attention fails centralized signals: Rethinking transformers for medical time series. In *International Conference on Learning Representations*, volume 2026, pp. 108935–108955, 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

A Theoretical Analysis of Distribution-Guided Correction

Proposition A.1. *For an arbitrary prediction step and variable, consider the correction in Eq. (11), where $\lambda = \alpha g_\theta \in [0, 1]$. If the raw prediction underestimates the normalized ground truth and the risk-aware distributional target provides a non-overshooting correction, i.e., $\hat{y}^{\mathrm{raw}} \le r_\theta \le \tilde{y}$, then*

$$|\tilde{y} - \hat{y}^{\mathrm{norm}}| \le |\tilde{y} - \hat{y}^{\mathrm{raw}}|. \tag{18}$$

Proof. Consider an arbitrary prediction step and variable, and omit the corresponding indices for simplicity. According to Eq. (11), the corrected prediction is

$$\begin{aligned} \hat{y}^{\mathrm{norm}} &= \hat{y}^{\mathrm{raw}} + \lambda\,(r_\theta - \hat{y}^{\mathrm{raw}}) \\ &= \hat{y}^{\mathrm{raw}} + \lambda r_\theta - \lambda \hat{y}^{\mathrm{raw}} \\ &= (1-\lambda)\hat{y}^{\mathrm{raw}} + \lambda r_\theta. \end{aligned} \tag{19}$$

Since $\lambda \in [0, 1]$, the coefficients $1 - \lambda$ and λ are both nonnegative and sum to one. Therefore, $\hat{y}^{\mathrm{norm}}$ is a convex combination of the raw prediction $\hat{y}^{\mathrm{raw}}$ and the risk-aware distributional target r_θ. Under the assumption $\hat{y}^{\mathrm{raw}} \le r_\theta$, the displacement of the corrected prediction from the raw prediction is

$$\hat{y}^{\mathrm{norm}} - \hat{y}^{\mathrm{raw}} = \lambda\,(r_\theta - \hat{y}^{\mathrm{raw}}) \ge 0, \tag{20}$$

while the remaining distance from the corrected prediction to the distributional target is

$$r_\theta - \hat{y}^{\mathrm{norm}} = (1-\lambda)\,(r_\theta - \hat{y}^{\mathrm{raw}}) \ge 0. \tag{21}$$

These two inequalities imply

$$\hat{y}^{\mathrm{raw}} \le \hat{y}^{\mathrm{norm}} \le r_\theta. \tag{22}$$

Combining this relation with the non-overshooting assumption $r_\theta \le \tilde{y}$ yields

$$\hat{y}^{\mathrm{raw}} \le \hat{y}^{\mathrm{norm}} \le r_\theta \le \tilde{y}. \tag{23}$$

Hence, both the raw and corrected predictions remain on the same side of the normalized ground truth. In particular, $\tilde{y} - \hat{y}^{\mathrm{raw}} \ge 0$ and $\tilde{y} - \hat{y}^{\mathrm{norm}} \ge 0$, so the absolute values can be removed without changing their signs. Therefore,

$$\begin{aligned} &|\tilde{y} - \hat{y}^{\mathrm{raw}}| - |\tilde{y} - \hat{y}^{\mathrm{norm}}| \\ &= (\tilde{y} - \hat{y}^{\mathrm{raw}}) - (\tilde{y} - \hat{y}^{\mathrm{norm}}) \\ &= \hat{y}^{\mathrm{norm}} - \hat{y}^{\mathrm{raw}} \\ &= \lambda\,(r_\theta - \hat{y}^{\mathrm{raw}}) \ge 0. \end{aligned} \tag{24}$$

Therefore,

$$|\tilde{y} - \hat{y}^{\mathrm{norm}}| \le |\tilde{y} - \hat{y}^{\mathrm{raw}}|. \tag{25}$$

□

B PSEUDOCODE OF PEAKTS

Algorithm 1: Training pipeline for **PEAKTS**

Input: Historical sequence $\mathbf{X} \in \mathbb{R}^{T \times C}$, future target $\mathbf{Y} \in \mathbb{R}^{P \times C}$.
Output: Forecasting result $\widehat{\mathbf{Y}}$.
 # 1. Input Normalization and Ordinal Target Construction.
Normalize $\mathbf{X}$ and $\mathbf{Y}$ using statistics computed from the historical window.
Convert the normalized future targets into soft ordinal distributions.
 # 2. Temporal Prediction.
Decompose the normalized input into seasonal and trend components.
Generate the raw forecast $\widehat{\mathbf{Y}}^{\text{raw}}$ and initial distribution logits $\boldsymbol{\ell}^{\text{lin}}$.
 # 3. Risk-Conditioned Peak Memory.
Extract multi-view representations from time, difference, and frequency domains.
Retrieve relevant patterns from the general and peak prototype banks.
Use the retrieved memory to modulate the initial distribution logits.
 # 4. Predictive Distribution Estimation.
Obtain the ordered predictive distribution $\mathbf{p}_\theta$ from the memory-modulated logits.
Derive the distributional mean, upper quantile, and confidence from $\mathbf{p}_\theta$.
 # 5. Distribution-Guided Forecast Correction.
Construct a risk-aware distributional target from the distributional statistics.
Adaptively correct $\widehat{\mathbf{Y}}^{\text{raw}}$ using the learned correction gate to obtain $\widehat{\mathbf{Y}}$.
 # 6. Joint Optimization.
Compute the tail-risk ordinal loss for distribution learning.
Compute the forecasting loss and consistency loss.
Jointly optimize the complete model.
return $\widehat{\mathbf{Y}}$.

C EXPERIMENTAL SETUP

C.1 DATASETS

We evaluate the proposed **PEAKTS** on nine real-world time-series forecasting datasets, including four ETT datasets, Electricity, Weather, AQWan, CzeLan, and ZafNoo (Wu et al., 2021; Qiu et al., 2025). These datasets cover different application domains, sampling frequencies, and temporal patterns, enabling a comprehensive evaluation of forecasting performance in high-value regions. Table 3 summarizes the dataset statistics, while detailed descriptions of each dataset are provided as follows:

Table 3: Dataset statistics.

Dataset	#Variables	#Timesteps	Frequency	Domain	Split
ETTh1	7	14,400	1 hour	Electricity	6:2:2
ETTh2	7	14,400	1 hour	Electricity	6:2:2
ETTm1	7	57,600	15 mins	Electricity	6:2:2
ETTm2	7	57,600	15 mins	Electricity	6:2:2
Weather	21	52,696	10 mins	Meteorology	7:1:2
Electricity	321	26,304	1 hour	Electricity	7:1:2
AQWan	11	35,064	1 hour	Environment	7:1:2
CzeLan	11	19,934	30 mins	Nature	7:1:2
ZafNoo	11	19,225	30 mins	Nature	7:1:2

- **ETT**$*$: The ETT dataset contains four subsets, including ETTh1 and ETTh2 with hourly observations, and ETTm1 and ETTm2 with 15-minute observations. These datasets record multiple measurements from electricity transformers and are widely used for long-term time-series forecasting.
- **Weather**: The Weather dataset contains meteorological measurements collected every 10 minutes from 2020 to 2021 at the Max Planck Institute for Biogeochemistry in Germany. It includes 21 variables related to weather and environmental conditions.
- **Electricity (ECL)**: The ECL dataset records the hourly electricity consumption of 321 customers. Each variable corresponds to the electricity usage of one customer, resulting in a multivariate time series with diverse consumption patterns.

- **AQWan**: The AQWan dataset contains hourly air-quality measurements collected from a monitoring station over four years. It includes 11 variables related to air pollution and meteorological conditions.
- **CzeLan**: The CzeLan dataset comes from the Sapflux project and is sampled every 30 minutes. It contains 11 variables related to sap flow and environmental conditions.
- **ZafNoo**: The ZafNoo dataset is also collected from the Sapflux project with a 30-minute sampling interval. It contains 11 variables describing sap flow and the surrounding environmental conditions.

C.2 METRICS

We use mean squared error (MSE) and mean absolute error (MAE) to evaluate forecasting performance as follows:

$$\begin{aligned} \text{MSE} &= \frac{1}{P \times C} \sum_{p=1}^{P} \sum_{c=1}^{C} \left(x_{t+p,c} - \hat{x}_{t+p,c} \right)^2, \\ \text{MAE} &= \frac{1}{P \times C} \sum_{p=1}^{P} \sum_{c=1}^{C} \left| x_{t+p,c} - \hat{x}_{t+p,c} \right|. \end{aligned} \tag{26}$$

C.3 BASELINE MODELS

We compare **PEAKTS** with 11 representative time-series forecasting methods.

- **DLinear** (Zeng et al., 2023): DLinear decomposes each series into trend and seasonal components and forecasts them with separate linear layers.
- **PatchTST** (Nie et al., 2023): PatchTST divides each univariate series into patches and uses a channel-independent Transformer to capture temporal dependencies among patch tokens.
- **iTransformer** (Liu et al., 2024): iTransformer represents each variable as a variate token and applies self-attention across variables to capture multivariate correlations.
- **TimeMixer** (Wang et al., 2024): TimeMixer decomposes multi-scale time series into seasonal and trend components and mixes them across resolutions to capture complex temporal patterns at different forecasting horizons.
- **TimeFilter** (Hu et al., 2025): TimeFilter constructs patch-level spatial-temporal graphs and adaptively removes irrelevant correlations while preserving informative dependencies.
- **TimeEmb** (Xia et al., 2025): TimeEmb separates time-invariant and time-varying components through a global embedding bank and lightweight frequency-domain filters.
- **SEED** (Xiong et al., 2026): SEED uses spectral entropy to balance channel-independent and channel-dependent modeling. Signed graphs and contextual spatial features further capture complex inter-variable relationships under varying temporal conditions.
- **OCE-TS** (Wang et al., 2026c) OCE-TS reformulates forecasting as ordinal classification and replaces MSE with ordinal cross-entropy. It predicts distributions over ordered intervals to preserve ordinal relationships, improve outlier robustness, and quantify uncertainty.
- **M2FMoE**: (Huang et al., 2026) M2FMoE combines Fourier- and wavelet-domain experts with multi-resolution fusion to model both regular patterns and rare extreme variations.
- **DistDF** (Wang et al., 2026a): DistDF aligns the joint distributions of predicted and target sequences through a differentiable Wasserstein discrepancy and accounts for dependencies among future values in multi-step forecasting tasks.
- **TeCh** (Yu et al., 2026): TeCh replaces decentralized self-attention with a centralized interaction strategy. It uses a global core token and adaptive tokenization to capture temporal and channel dependencies with linear computational complexity.

Table 4: Full forecasting performance comparison on nine time-series datasets with the look-back length fixed to $T = 336$ and prediction horizons $P \in \{96, 192, 336, 720\}$. Performance is evaluated using MSE and MAE, where lower values indicate better performance. "Avg" denotes the average result over the four prediction horizons. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method		DLinear (2023)		PatchTST (2023)		iTransformer (2024)		TimeMixer (2024)		TimeFilter (2025)		TimeEmb (2025)		SEED (2026)		OCE-TS (2026)		M2FMoE (2026)		DistDF (2026)		TeCh (2026)		PEAKTS (Ours)	
Year																									
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	0.300	0.344	0.301	0.358	0.304	0.358	0.306	0.359	0.298	0.350	0.299	0.355	0.298	0.348	0.545	0.536	0.348	0.375	0.308	0.349	0.382	0.431	**0.282**	**0.331**
	192	0.335	0.367	0.341	0.380	0.340	0.379	0.334	0.372	0.338	0.373	0.341	0.381	0.341	0.373	0.580	0.556	0.416	0.423	0.345	0.370	0.410	0.446	**0.326**	**0.358**
	336	0.370	0.388	0.376	0.403	0.374	0.399	0.375	0.399	0.374	0.391	0.374	0.400	0.378	0.395	0.596	0.566	0.436	0.434	0.380	0.392	0.489	0.478	**0.364**	**0.381**
	720	0.425	0.420	0.428	0.433	0.432	0.431	0.425	0.426	0.425	0.422	0.426	0.426	0.441	0.435	0.644	0.595	0.632	0.555	0.434	0.421	0.591	0.546	**0.423**	**0.415**
	Avg	0.358	0.380	0.362	0.394	0.363	0.392	0.360	0.389	0.359	0.384	0.360	0.391	0.365	0.388	0.591	0.563	0.458	0.447	0.367	0.383	0.468	0.475	**0.349**	**0.371**
ETTm2	96	0.167	0.262	0.181	0.276	0.181	0.274	0.176	0.261	0.170	0.262	0.182	0.271	0.167	0.256	2.685	1.029	0.189	0.286	0.164	0.253	0.307	0.437	**0.162**	**0.245**
	192	0.229	0.309	0.269	0.344	0.243	0.313	0.232	0.298	0.233	0.304	0.244	0.311	0.230	0.303	2.813	1.044	0.282	0.353	0.221	0.291	0.414	0.522	**0.218**	**0.286**
	336	0.297	0.358	0.376	0.401	0.291	0.342	0.292	0.339	0.290	0.343	0.295	0.345	0.288	0.343	2.770	1.052	0.515	0.509	0.276	0.327	0.595	0.598	**0.268**	**0.318**
	720	0.439	0.450	0.457	0.455	0.378	0.396	0.371	0.392	0.372	0.400	0.386	0.399	0.372	0.400	2.662	1.064	0.478	0.472	0.371	0.384	2.195	1.257	**0.349**	**0.371**
	Avg	0.283	0.345	0.321	0.369	0.273	0.331	0.268	0.323	0.266	0.327	0.277	0.332	0.264	0.326	2.733	1.047	0.366	0.405	0.258	0.314	0.878	0.704	**0.249**	**0.305**
ETTh1	96	0.373	0.396	0.394	0.418	0.394	0.415	0.390	0.406	0.389	0.408	0.397	0.418	0.385	0.403	0.623	0.621	0.507	0.482	0.376	0.397	0.497	0.504	**0.368**	**0.389**
	192	0.415	0.419	0.437	0.445	0.431	0.437	0.433	0.434	0.431	0.432	0.424	0.432	0.421	0.424	0.643	0.631	0.592	0.516	**0.410**	0.416	0.594	0.564	**0.410**	**0.415**
	336	0.436	0.441	0.462	0.464	0.448	0.451	0.469	0.459	0.452	0.449	0.442	0.445	0.449	0.444	0.665	0.649	0.706	0.599	**0.435**	**0.432**	0.664	0.607	0.438	**0.432**
	720	0.475	0.493	0.526	0.522	0.462	0.478	0.488	0.491	0.478	0.483	0.468	0.475	0.473	0.482	0.689	0.671	0.848	0.663	**0.446**	**0.459**	0.762	0.665	0.461	0.472
	Avg	0.425	0.437	0.455	0.462	0.434	0.445	0.445	0.448	0.438	0.443	0.433	0.443	0.432	0.438	0.655	0.643	0.663	0.565	**0.417**	**0.426**	0.629	0.585	0.419	0.427
ETTh2	96	0.291	0.357	0.321	0.382	0.327	0.374	0.316	0.362	0.303	0.360	0.318	0.371	0.307	0.360	1.244	0.785	0.373	0.418	0.277	0.337	1.521	1.023	**0.273**	**0.330**
	192	0.381	0.416	0.451	0.465	0.388	0.411	0.367	0.398	0.366	0.399	0.375	0.409	0.365	0.401	1.251	0.803	0.438	0.444	0.343	0.380	1.504	0.948	**0.337**	**0.373**
	336	0.460	0.469	0.521	0.503	0.400	0.424	0.402	0.425	0.396	0.421	0.395	0.429	0.398	0.421	1.338	0.826	0.618	0.548	**0.367**	0.403	1.518	0.989	0.369	**0.397**
	720	0.731	0.607	0.787	0.627	0.420	0.447	0.412	0.439	0.415	0.445	0.432	0.456	0.414	0.444	3.069	1.216	1.312	0.716	0.392	0.429	1.080	0.804	**0.372**	**0.414**
	Avg	0.466	0.462	0.520	0.494	0.384	0.414	0.374	0.406	0.370	0.406	0.380	0.416	0.371	0.407	1.726	0.908	0.685	0.532	0.345	0.387	1.406	0.941	**0.338**	**0.379**
Weather	96	0.174	0.236	0.153	0.213	0.159	0.207	0.154	0.202	0.150	0.204	0.149	0.203	**0.147**	0.202	36.91	2.580	0.154	0.219	0.175	0.225	0.151	0.220	0.155	**0.192**
	192	0.216	0.275	0.200	0.261	0.205	0.250	**0.190**	**0.235**	0.199	0.249	0.194	0.245	0.194	0.244	37.43	2.620	0.202	0.269	0.218	0.261	0.198	0.268	0.200	**0.235**
	336	0.264	0.316	0.259	0.305	0.255	0.289	0.257	0.292	0.251	0.291	0.248	0.286	**0.246**	0.287	38.54	2.685	0.267	0.310	0.266	0.296	0.252	0.309	0.252	**0.277**
	720	0.335	0.375	0.331	0.358	**0.322**	0.335	0.327	0.339	**0.322**	0.340	0.325	0.337	0.327	0.340	39.99	2.764	0.341	0.377	0.335	0.344	0.328	0.372	0.330	**0.331**
	Avg	0.247	0.301	0.236	0.284	0.235	0.270	0.232	0.267	0.231	0.271	0.229	0.268	**0.229**	0.268	38.21	2.662	0.241	0.294	0.249	0.282	0.232	0.292	0.234	**0.259**
Electricity	96	0.140	0.238	0.136	0.232	0.143	0.244	0.145	0.240	0.168	0.265	0.140	0.241	**0.134**	**0.230**	0.819	0.551	0.142	0.242	0.141	0.237	0.215	0.333	0.138	0.231
	192	0.154	0.251	0.155	0.252	0.162	0.262	**0.152**	0.245	0.180	0.277	0.158	0.257	0.153	0.248	0.983	0.621	0.155	0.252	0.156	0.250	0.253	0.377	**0.152**	**0.242**
	336	0.169	0.268	0.172	0.270	0.180	0.280	0.168	0.262	0.196	0.293	0.175	0.274	0.174	0.270	1.210	0.765	0.174	0.274	0.172	0.267	0.227	0.344	**0.167**	**0.256**
	720	0.214	0.308	**0.205**	0.300	0.226	0.321	0.208	0.295	0.242	0.339	0.214	0.306	0.206	0.296	1.442	0.803	0.213	0.302	0.211	0.298	0.248	0.348	**0.205**	**0.289**
	Avg	0.169	0.266	0.167	0.264	0.178	0.277	0.168	0.261	0.197	0.294	0.172	0.270	0.167	0.261	1.114	0.685	0.171	0.268	0.170	0.263	0.236	0.351	**0.166**	**0.255**
AQWan	96	0.719	0.478	0.716	0.468	0.716	0.470	0.720	0.468	0.724	0.472	0.717	0.466	0.716	0.462	2.728	1.027	0.808	0.506	0.726	0.472	0.720	0.487	**0.710**	**0.455**
	192	0.762	0.499	0.751	0.488	0.754	0.487	0.762	0.489	0.757	0.487	0.758	0.484	0.755	0.481	2.786	1.051	1.183	0.544	0.771	0.492	0.755	0.503	**0.748**	**0.475**
	336	0.784	0.516	**0.767**	0.501	0.780	0.499	0.780	0.497	0.789	0.500	0.778	0.495	0.779	0.495	2.722	1.041	0.847	0.538	0.792	0.503	0.796	0.523	0.777	**0.484**
	720	0.854	0.546	**0.824**	0.525	0.850	0.527	0.850	0.522	0.845	0.523	0.843	0.520	0.849	0.522	2.909	1.073	0.888	0.555	0.870	0.533	0.827	0.531	0.858	**0.514**
	Avg	0.780	0.510	**0.765**	0.496	0.775	0.496	0.778	0.494	0.779	0.496	0.774	0.491	0.775	0.490	2.786	1.048	0.932	0.536	0.790	0.500	0.775	0.511	0.773	**0.482**
CzeLan	96	0.227	0.305	0.769	0.498	0.183	0.240	0.181	0.240	0.177	0.235	0.180	0.231	0.178	0.232	18.58	1.938	0.325	0.391	0.179	0.230	0.727	0.517	**0.175**	**0.208**
	192	0.297	0.353	1.102	0.599	0.212	0.263	0.215	0.258	0.219	0.268	0.213	0.254	0.211	0.259	20.31	1.990	0.413	0.431	0.211	0.254	0.957	0.624	**0.209**	**0.231**
	336	0.386	0.405	1.458	0.693	0.238	0.286	0.234	0.275	0.237	0.293	0.236	0.276	**0.230**	0.284	24.97	2.226	0.991	0.606	0.244	0.283	1.264	0.717	0.244	**0.257**
	720	0.534	0.481	1.835	0.782	**0.270**	0.315	0.274	0.310	0.276	0.325	0.274	0.305	0.289	0.332	25.06	2.286	0.695	0.545	0.300	0.324	1.447	0.723	0.292	**0.293**
	Avg	0.361	0.386	1.291	0.643	**0.226**	0.276	0.226	0.271	0.227	0.280	**0.226**	0.267	0.227	0.277	22.23	2.110	0.606	0.493	0.234	0.273	1.099	0.645	0.230	**0.247**
ZafNoo	96	0.435	0.412	0.442	0.418	0.450	0.415	0.440	0.406	0.441	0.413	0.432	0.399	**0.427**	0.401	3.251	1.110	0.455	0.437	0.448	0.413	0.440	0.428	0.451	**0.391**
	192	0.487	0.447	0.487	0.446	0.509	0.451	0.504	0.449	0.511	0.453	0.495	0.439	**0.486**	0.443	3.532	1.157	0.518	0.481	0.513	0.450	0.489	0.464	0.515	**0.433**
	336	0.521	0.467	**0.520**	0.465	0.563	0.481	0.552	0.471	0.560	0.480	0.545	0.466	0.541	0.473	4.130	1.241	0.559	0.495	0.562	0.476	0.521	0.480	0.559	**0.459**
	720	**0.549**	0.489	0.562	0.489	0.614	0.509	0.607	0.503	0.613	0.505	0.629	0.519	0.587	0.497	5.137	1.355	0.618	0.531	0.615	0.504	0.554	0.499	0.611	**0.487**
	Avg	**0.498**	0.454	0.503	0.455	0.534	0.464	0.526	0.457	0.531	0.463	0.525	0.456	0.510	0.454	4.013	1.216	0.538	0.486	0.535	0.461	0.501	0.468	0.534	**0.443**

D FORECASTING VISUALIZATION

Figures 7 and 8 visualize the forecasting results on ETTm1 and ECL. While most baselines capture the overall temporal patterns, they show noticeable deviations around high-value regions, such as peak underestimation or over-smoothed predictions. Such deviations are observed under both the varying dynamics of ETTm1 and the pronounced periodic patterns of ECL. In contrast, **PEAKTS** follows the ground truth more closely, particularly around high-value regions, while preserving the overall temporal trend. These results demonstrate the effectiveness of **PEAKTS** in capturing high-value dynamics across different temporal patterns.

E HYPERPARAMETER ANALYSIS

E.1 SENSITIVITY TO THE LOSS WEIGHTS

We examine the effects of the ordinal loss weight $\lambda_{\text{ord}} \in \{0.1, 0.5, 1, 5, 10\}$ and the consistency loss weight $\lambda_{\text{cons}} \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. As shown in Fig. 9, the best overall performance is obtained with $\lambda_{\text{ord}} = 1$ and $\lambda_{\text{cons}} = 0.01$. A small λ_{ord} provides insufficient

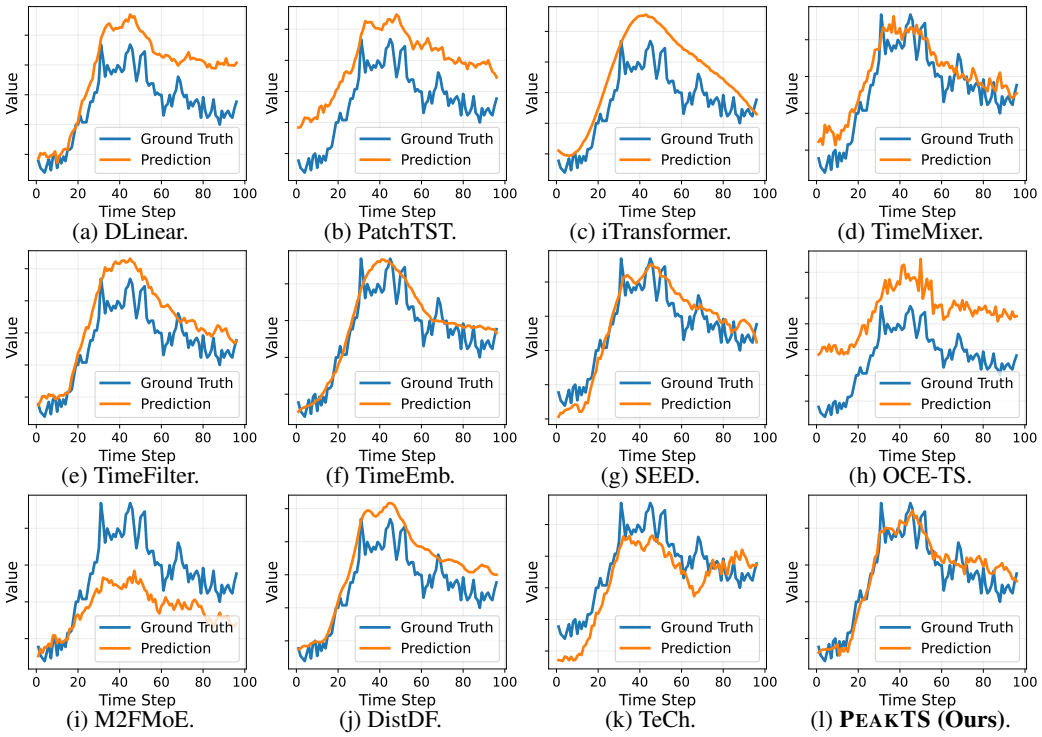


Figure 7: Prediction visualization results on the ETTm1 dataset.

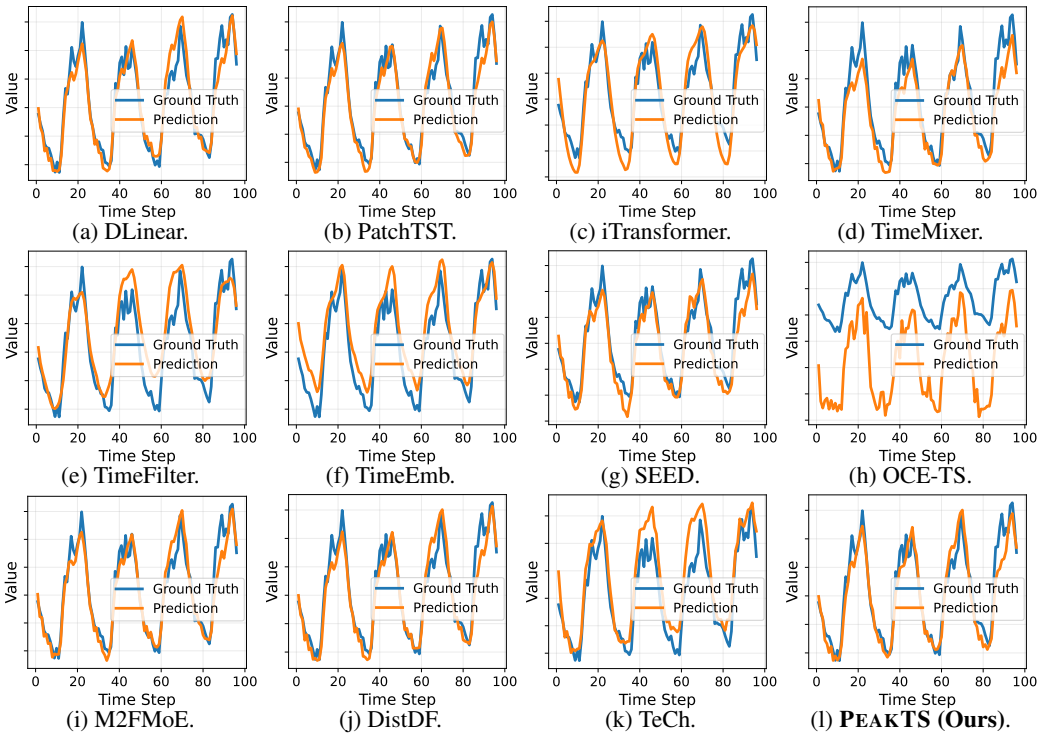
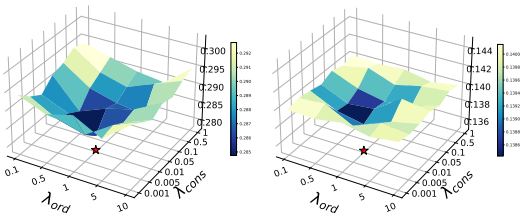
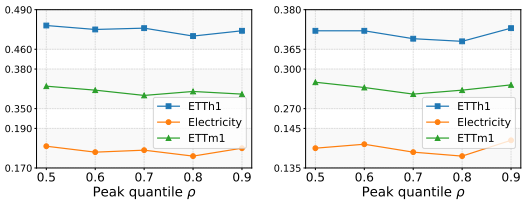


Figure 8: Prediction visualization results on the Electricity dataset.



(a) ETTm1 dataset. (b) Electricity dataset.

Figure 9: Loss weights λ_{ord} and λ_{cons}.



(a) Top-25% region. (b) All regions.

Figure 10: Sensitivity to ρ.

Table 5: Comparison of different temporal-view configurations in **PEAKTS**.

Variant	ETTm1				ETTh1				Electricity			
	Top 25%		All		Top 25%		All		Top 25%		All	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Time Only	0.374	0.385	0.287	0.334	0.476	0.429	0.371	0.392	0.181	0.264	0.140	0.234
Time+Diff	0.365	0.380	0.284	0.332	0.473	0.427	0.369	0.390	0.180	0.262	0.139	0.233
Time+Freq	0.368	0.382	0.285	0.333	0.474	0.428	0.370	0.391	0.178	0.260	0.139	0.232
Full	0.361	0.378	0.282	0.331	0.471	0.425	0.368	0.389	0.177	0.259	0.138	0.231

supervision for the probability branch, whereas an excessively large value places too much emphasis on distribution learning at the expense of value prediction. For λ_{cons}, a moderate value helps align the final prediction with the distributional mean. However, a large consistency weight may over-constrain the corrected prediction and limit its ability to fit the ground truth. These results show that balancing the ordinal, value, and consistency objectives is important for accurate forecasting.

E.2 Sensitivity to the Peak Quantile

Figure 10 presents the forecasting performance under different values of the peak quantile ρ. The results on both the top-25% region and the full test set show that values between 0.7 and 0.8 generally provide better performance. In particular, $\rho = 0.8$ achieves the lowest MSE on ETTh1 and Electricity, while $\rho = 0.7$ gives the best result on ETTm1. A smaller ρ may include too many ordinary patterns in the peak prototype bank, making peak patterns less distinguishable. A larger ρ, however, reduces the number of available peak samples and limits the diversity of the learned prototypes.

F Effect of Multi-View Temporal Representations

As shown in Table 5, incorporating either the difference-domain or frequency-domain view consistently improves performance over using the time-domain view alone. The difference-domain view yields larger gains on ETTm1 and ETTh1, while the frequency-domain view performs better on Electricity, reflecting its pronounced periodic patterns. Combining all three views achieves the lowest errors across all datasets and evaluation regions, indicating that the views capture complementary temporal characteristics. The larger improvements in the Top-25% regions further suggest that multi-view representations are particularly helpful for high-value forecasting.

G AI Use Statement

In this work, we used generative AI tools for language editing to improve the clarity, grammar, and readability of the manuscript. All AI-assisted edits were carefully reviewed and verified by the authors. We take full responsibility for the final content of this work.