Attention-Based Feature Online Conformal Prediction for Time Series

Anonymous authors
Paper under double-blind review

Abstract

Online conformal prediction (OCP) wraps around any pre-trained predictor to produce prediction sets with coverage guarantees that hold irrespective of temporal dependencies or distribution shifts. However, standard OCP faces two key limitations: it operates in the output space using simple nonconformity (NC) scores, and it treats all historical observations uniformly when estimating quantiles. This paper introduces attention-based feature OCP (AFOCP), which addresses both limitations through two key innovations. First, AFOCP operates in the feature space of pre-trained neural networks, leveraging learned representations to construct more compact prediction sets by concentrating on task-relevant information while suppressing nuisance variation. Second, AFOCP incorporates a multi-head attention mechanism that adaptively weights historical observations based on their relevance to the current test point, effectively handling non-stationarity and distribution shifts. We provide theoretical guarantees showing that AFOCP maintains long-term coverage while achieving smaller long-term time-averaged prediction sets than standard OCP under mild regularity conditions. Extensive experiments on synthetic and real-world time series datasets demonstrate that AFOCP consistently reduces the prediction interval lengths by as much as 88% relative to OCP and yields shorter intervals than the online counterparts of representative offline CP designs for time series, while maintaining target coverage levels, validating the benefits of both feature-space calibration and attention-based adaptive weighting.

1 Introduction

1.1 Context and Motivation

Uncertainty quantification has become increasingly critical as machine learning systems are deployed in high-stakes applications such as autonomous vehicles, medical diagnosis, financial forecasting, and telecommunications (Gawlikowski et al., 2023; Angelopoulos & Bates, 2021; Simeone et al., 2025). While deep neural networks and language models have achieved remarkable predictive accuracy, they often produce overconfident predictions without reliable uncertainty estimates (Guo et al., 2017; Kadavath et al., 2022). Traditional approaches to uncertainty quantification, such as Bayesian methods (Gal & Ghahramani, 2016; Simeone, 2022; Huang et al., 2024) and ensemble techniques (Lakshminarayanan et al., 2017; Abbasli et al., 2025), either require strong distributional assumptions or significant computational overhead, limiting their practical applicability in streaming and resource-constrained settings.

Conformal prediction (CP) (Vovk et al., 2005; Shafer & Vovk, 2008) offers an attractive alternative aiming at calibrating the uncertainty level of existing, pre-trained, models. This is done by augmenting a model's output with prediction sets or intervals that meet a desired miscoverage level. Specifically, given a target miscoverage rate α, for any given pre-trained model, CP constructs prediction sets that contain the true label with probability at least $1-\alpha$, without making any assumptions about the underlying data distribution beyond exchangeability. This framework has gained significant attention in recent years due to its model-agnostic nature and rigorous theoretical foundations (Angelopoulos et al., 2024).

However, the exchangeability assumption – that data points can be reordered without changing their joint distribution – is frequently violated in time series and sequential prediction tasks. Temporal dependencies,

concept drift, and distribution shifts are inherent characteristics of many real-world applications, including financial markets, weather forecasting, and sensor networks. When exchangeability fails, standard CP methods can suffer from miscalibration, leading to prediction sets that either under-cover, failing to contain the true value, or over-cover, producing excessively large, uninformative intervals.

To address non-exchangeability, online conformal prediction (OCP) (Gibbs & Candes, 2021) adapts the conformal framework by continuously updating miscoverage levels, or prediction confidence thresholds, through feedback mechanisms. Unlike CP, whose coverage guarantees are probabilistic, OCP provides deterministic long-term coverage guarantees. However, in practice, it still faces two important limitations:

1. *Processing in the output space:* OCP typically leverages simple confidence scores in the output space, such as absolute prediction errors. This approach fails to leverage the rich semantic representations learned by modern deep neural networks, potentially leading to overly conservative prediction intervals.

2. *Uniform weighting:* OCP treats all historical observations *uniformly* when producing the current prediction set, ignoring the fact that some past data points may be more relevant than others for predicting uncertainty at the given time step.

The goal of this work is to address these limitations while preserving the online coverage guarantees of OCP.

1.2 Related Work

Conformal prediction in non-exchangeable and time-series settings. Building on the theoretical foundational framework of Vovk et al. (2005), the literature on CP has developed along several directions, including extensions to inductive (Papadopoulos et al., 2002) and cross-validation-based methodologies (Lei et al., 2018; Cohen et al., 2024). More recent advances have extended CP to non-exchangeable settings. For example, Barber et al. (2023) developed weighted CP (WCP) methods with explicit bounds on coverage gaps under distribution drift, while Oliveira et al. (2024) proved that inductive CP remains approximately valid for non-exchangeable processes (see also Tibshirani et al. (2019); Bhattacharyya & Barber (2024)).

For time series, Zecchin et al. (2024) and Lindemann et al. (2023) proposed methods that obtain coverage properties on average over time series, while Xu & Xie (2021) and Xu & Xie (2023) introduced methodologies that can leverage existing predictors to achieve asymptotic valid conditional coverage under given technical assumptions. A complementary line of work studies cross-sectional and longitudinal validity for time-series regression (Lin et al., 2022). Another line developed conformalized probabilistic forecasting via quantile regression, producing prediction intervals that adapt to heteroscedasticity while retaining validity guarantees (Romano et al., 2019; Jensen et al., 2024).

A central question in WCP is how to assign weights to historical observations. Early designs rely on handcrafted weights, such as likelihood ratios (Tibshirani et al., 2019) or exponential recency decay (Barber et al., 2023). Subsequent work has explored learning the weights from data. Auer et al. (2023) proposed a WCP method that uses an attention mechanism based on modern Hopfield networks to identify similar error regimes in the output space. This approach relies on attention modules pre-trained offline on a held-out calibration set, and operates within a split-conformal framework without updating the miscoverage level during deployment. Lee et al. (2025) adopts a kernel-based weighting scheme as an alternative to attention. Heurich et al. (2026) considers a weighting architecture that combines sparse retrieval, a mixture-of-experts (MoE) gate, and a hypernetwork for large-scale time-series foundation-model benchmarks. Across these designs, the weighting module is learned offline, and calibration happens in output space.

Finally, Teng et al. (2022) introduced feature CP (FCP), which extends conventional CP to operate in semantic feature spaces by leveraging the inductive bias of deep representation learning. They demonstrate provable improvements over output-space methods under reasonable assumptions, such as a stable feature space and a smooth prediction head. Building on this feature-space perspective, Chen et al. (2024) combined FCP with attention-based weighting in the spirit of Auer et al. (2023), similarly relying on an offline-trained attention module and a fixed miscoverage level.

Online conformal prediction and online learning under non-stationarity. Gibbs & Candes (2021) introduced OCP, which continuously re-estimates miscoverage parameters using gradient descent to maintain long-term coverage under distribution shift. Follow-up work refined the miscoverage update through automatic step-size tuning (Gibbs & Candès, 2024; Zaffran et al., 2022), proportional-integral-derivative mechanisms (Angelopoulos et al., 2023), localization mechanisms (Zecchin & Simeone, 2024), and strongly adaptive properties (Bhatnagar et al., 2023).

A separate line of work extended the OCP framework along structural dimensions: multi-model selection from a candidate pool under distribution shift (Hajihashemi & Shen, 2025), selective inference when prediction intervals are only required at certain time steps (Sale & Ramdas, 2025), feature-augmented residual predictors that feed intermediate representations as auxiliary inputs while keeping calibration in the output space (Huang & Qiu, 2025), and the extension to settings with multiple parallel time series, where calibration draws on observations from other related series at the same time step (Tu & Giesecke, 2026). Across these directions, calibration is typically performed in the output space, and learning-based weighting, when introduced, is trained offline rather than updated online. To our knowledge, no prior work integrates feature-space calibration, online-trained attention weights, and online miscoverage updates within a single framework, which is the gap addressed by AFOCP.

More broadly, adaptation under non-stationarity and distribution shift has been studied in related sequential settings. Recent work investigated adaptation under evolving conditions for sequential and time-series problems, often by leveraging temporal structure and incoming data (Su et al., 2024; Li et al., 2024; Yang et al., 2025). Further work analyzed degradation under evolving distributions and studied selective reuse of past information (Talbot et al., 2025; Wang et al., 2025). These works motivate incorporating data-dependent notions of relevance when leveraging historical observations in sequential settings.

1.3 Main Contributions

In this paper, we aim to calibrate a pre-trained machine learning model to generate a prediction set that includes the ground-truth label for a sufficiently large fraction of time. Specifically, as shown in Figure 1, we consider an online setting in which input-label pairs $\{(X_t, Y_t)\}_{t=1,2,\ldots}$ arrive sequentially as a time series over discrete time t. At time $t = 1, 2, \ldots$, given a test input X_t, a pre-trained model $\mu(\cdot)$, and the historical data pairs $\{(X_\tau, Y_\tau)\}_{\tau=1}^{t-1}$, our goal is to construct a prediction set that contains the true label Y_t for a fraction at least $100(1-\alpha)\%$ of the time.

As reviewed above, this goal can be attained by leveraging OCP and variants. This paper introduces *attention-based feature OCP* (AFOCP), a novel framework that enhances OCP through two key innovations:

1. *Feature-space OCP for compact prediction sets:* We extend OCP to construct prediction sets in the feature space of pre-trained neural networks. By computing confidence scores using learned representations rather than output-space predictions, AFOCP exploits the inductive bias of deep learning models to construct more informative prediction sets. Specifically, the feature extractor allows AFOCP to concentrate on task-relevant information, while suppressing nuisance variation.

2. *Attention-based adaptive weighting for non-stationary data:* We incorporate an attention mechanism that learns to assign relevance weights to historical observations based on their similarity to the current test point in feature space. Unlike standard OCP, which treats all data in the calibration window uniformly, our attention-based approach emphasizes past observations from similar distributional regimes. The attention weights are learned online through an autoregressive prediction task, where the model minimizes the error in predicting current nonconformity (NC) scores from past scores, weighted by feature similarity. This enables AFOCP to adapt to distribution shifts and temporal dependencies without requiring explicit change-point detection or regime identification.

Overall, our main contributions are as follows:

- We propose FOCP, an online calibration approach operating in the learned feature space of a pre-trained predictor. We further generalize FOCP to AFOCP by replacing uniform score aggregation with multi-head attention-based weighting over recent observations.
- We establish deterministic long-term coverage guarantees in the online setting and show that FOCP and AFOCP can attain smaller long-term time-averaged prediction sets than output-space OCP, or its attention-based counterpart, AOCP, under some regularity assumptions.
- On synthetic and real-world time-series benchmarks, AFOCP consistently achieves the target coverage, while markedly reducing the prediction interval length relative to OCP and to the online counterparts of representative offline CP designs for time series. Ablations are provided to disentangle the roles of feature-space calibration and attention-based weighting, showing how their relative gains vary with the feature dimension and calibration window length.

The remainder of this paper is organized as follows. Section 2 reviews OCP. Section 3 introduces the AFOCP framework, including feature-based NC scores, attention-based weighting, and theoretical guarantees. Section 4 presents experimental results on synthetic and real-world datasets. Section 5 concludes with discussion and future directions.

2 Online Conformal Prediction

OCP (Gibbs & Candes, 2021) extends the traditional CP framework (Vovk et al., 2005) by incorporating an online update mechanism. In this setting, CP defines an NC score to measure the dissimilarity between the model's prediction $\mu(X)$ for an input $X \in \mathcal{X} \subseteq \mathbb{R}^{D_{\text{in}}}$ and any candidate label $Y \in \mathcal{Y} \subseteq \mathbb{R}^{D_{\text{out}}}$, where D_{in} and D_{out} denote the dimensions of input and output variables, respectively. For regression tasks, a common choice for the NC score is the absolute error, i.e.,

$$s(X, Y) = \|Y - \mu(X)\| . \tag{1}$$

To adapt the method for streaming data, at time t, we calculate the $(1-\alpha_t)$-quantile of the NC scores using a sliding window of the most recent L observations $\{(X_{t-\tau}, Y_{t-\tau})\}_{\tau=1}^{L}$. The prediction set for input X_t is then constructed as

$$\Gamma_t^{\text{OCP}}(X_t) = \left\{ Y \in \mathcal{Y} : s\left(X_t, Y\right) \leq Q_{1-\alpha_t}\left(\sum_{\tau=1}^{L} \frac{1}{L+1} \delta_{s(X_{t-\tau}, Y_{t-\tau})} + \frac{1}{L+1} \delta_{+\infty} \right) \right\}, \tag{2}$$

where $Q_a(\cdot)$ denotes the a-quantile of its argument and δ_b represents the Dirac delta function centered at b. We assume that $\alpha_1 \in [0, 1]$ and that the quantile function Q_a is non-decreasing, with $Q_a = -\infty$ for $a < 0$ and $Q_a = \infty$ for $a > 1$.

For each prediction set $\Gamma_t^{\text{OCP}}(X_t)$ to achieve coverage probability $1 - \alpha$, i.e., $\mathbb{P}\left(Y_t \in \Gamma_t^{\text{OCP}}(X_t)\right) \geq 1 - \alpha$, the data $\{(X_\tau, Y_\tau)\}_{\tau=1}^{t}$ must be exchangeable, and we assume a fixed $\alpha_t = \alpha$ (Vovk et al., 2005). However, this exchangeability assumption typically does not hold for time-series data. To address this, OCP further updates the unreliability level α_t in the quantile in (2) using the online rule (Gibbs & Candes, 2021),

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t), \tag{3}$$

where γ denotes the step size, and the discrete error err_t is defined as

$$\text{err}_t = \begin{cases} 1 & \text{if } Y_t \notin \Gamma_t^{\text{OCP}}(X_t), \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This online update rule ensures that the predicted coverage probability converges to the desired level in the long run. The following theorem provides a reliability guarantee for OCP.

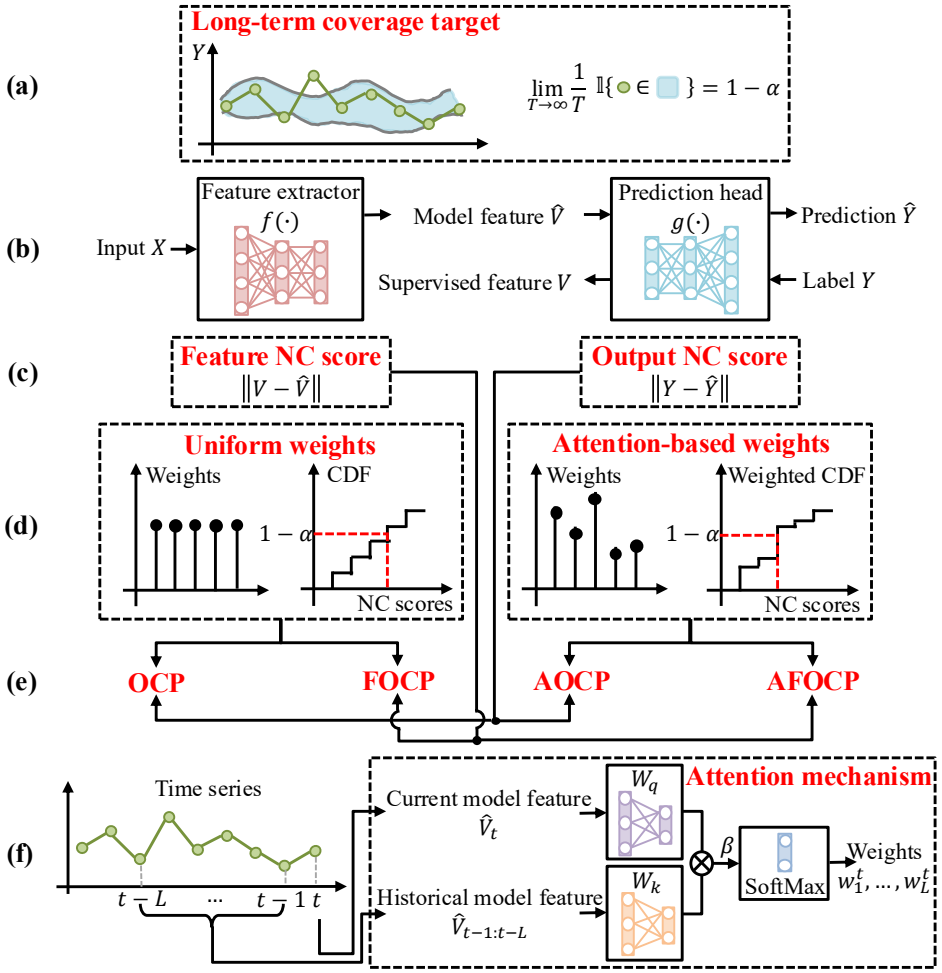


Figure 1: Overview of AFOCP and related baselines. (a) The goal of this work is to calibrate pre-trained predictors by augmenting their outputs with prediction sets that contain the true label Y for a fraction at least $100(1-\alpha)\%$ of the time. (b) For any input X, the pre-trained model $\mu(X) = g \circ f(X)$ maps inputs X through the feature extractor $f(\cdot)$ and the prediction head $g(\cdot)$. (c) Nonconformity (NC) scores can be evaluated in the *output* or *feature* spaces. (d) The NC scores can be combined to evaluate empirical distributions and quantiles using either *uniform weights* or *attention-based weights*. (e) OCP (Gibbs & Candes, 2021) uses output scores with uniform weights; feature OCP (FOCP) uses feature scores, while retaining uniform weights; attention-based OCP (AOCP) keeps output scores but learns data-dependent weights via attention; and attention-based feature OCP (AFOCP) combines feature scores with attention-based weights. FOCP, AOCP, and AFOCP are introduced in this work, with AFOCP being the most general of the three. (f) The attention mechanism in AOCP and AFOCP compares the current model feature with past features to produce similarity-based weights that serve as data-dependent weights for calibration.

Theorem 1 (**Proposition 4.1 (Gibbs & Candes, 2021)**). *The average error over time $T \in \mathbb{N}$ satisfies the equality*

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{err}_t = \alpha + \frac{\alpha_1 - \alpha_{T+1}}{T\gamma}. \tag{5}$$

In particular, as $T \to \infty$, the error converges to the desired level α, i.e.,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathrm{err}_t = \alpha. \tag{6}$$

Proof: Expanding the recursion in (3) leads to (5). Since $\alpha_t \in [-\gamma, 1+\gamma]$ according to Lemma 4.1 in Gibbs & Candes (2021), the long-term coverage in (6) holds.

3 Attention-based Feature Online Conformal Prediction

In this paper, we introduce AFOCP, a novel variant of OCP that leverages compact data representations in the feature space in order to obtain more compact prediction sets and to learn relevance weights over historical data to inform an attention-based quantile estimate.

3.1 Feature Online Conformal Prediction

As done in Teng et al. (2022) to introduce feature (offline) CP, we focus on pre-trained prediction models $\mu(\cdot)$, kept fixed throughout the online procedure, that can be decomposed as

$$\mu(\cdot) = g \circ f(\cdot), \tag{7}$$

where the feature extractor $f(\cdot)$ maps the input X to the latent feature $\hat{V} = f(X) \in \mathbb{R}^D$ with D denoting the dimension of the feature, and the prediction head $g(\cdot)$ transforms these features into output predictions $\hat{Y} = g(\hat{V})$.

To any data pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, we can thus associate two, generally different, features:

1) Model feature: The model feature is obtained by running the model in inference mode through the feature extractor $f(\cdot)$ as

$$\hat{V} = f(X) \tag{8}$$

2) Supervised feature: The supervised feature is obtained by running the model backward, starting from the label Y through the inverse of the prediction head $g(\cdot)$ as

$$V \in g^{-1}(Y). \tag{9}$$

Since function $g(\cdot)$ is generally many-to-one, any vector V in the inverse image $g^{-1}(Y)$ can be selected in (9).

The NC score function in the feature space is then defined as the difference between model feature and supervised feature as (Teng et al., 2022)

$$s^{\mathrm{f}}(X, Y) = \inf_{V \in g^{-1}(Y)} \left\| V - \hat{V} \right\|, \tag{10}$$

where the infimum operator is over all feature vectors in the inverse image (9) of the prediction head. We discuss in Section 3.3 how to evaluate the scores (10) in practice.

Given a new test input X_t at time t, the proposed AFOCP scheme constructs a prediction set $\Gamma_t^{\mathrm{AFOCP}}(X_t)$ based on scores in the feature space for the latest observed L data pairs $\{(X_{t-\tau}, Y_{t-\tau})\}_{\tau=1}^{L}$. Specifically, the prediction set for input X_t is constructed as

$$\Gamma_t^{\mathrm{AFOCP}}(X_t) = \left\{ Y \in \mathcal{Y} : s^{\mathrm{f}}(X_t, Y) \leq Q_{1-\alpha_t} \left(\sum_{\tau=1}^{L} w_t^{\tau} \delta_{s^{\mathrm{f}}(X_{t-\tau}, Y_{t-\tau})} + w_t^{L+1} \delta_{+\infty} \right) \right\}, \tag{11}$$

where the quantile function $Q_a(\cdot)$ and point mass δ_b are defined as in (2), while the weights $\{w_t^{\tau}\}_{\tau=1}^{L+1}$ are discussed next. Compared to the conventional OCP predictor in (2), AFOCP has the following distinguishing characteristics:

1) It leverages feature-based NC scores to evaluate the prediction set (11). This typically yields shorter prediction sets because the learned representation $f(X)$ concentrates task-relevant information and suppresses nuisance variation, which reduces the dispersion of feature-based NC scores (Teng et al., 2022). A smooth prediction head g maps this reduction to the output, while preserving the target coverage. See Theorem 2 for a rigorous justification.

2) It introduces normalized weights $\{w_t^\tau\}_{\tau=1}^{L+1}$, with weight w_t^τ assigned to the data point $X_{t-\tau}$ for $\tau = 1, \ldots, L$. Ideally, a larger weight w_t^τ should be assigned to a data point $X_{t-\tau}$ that is likely to share a similar data distribution with the current test point X_t, thereby improving the accuracy of the quantile estimation (Tibshirani et al., 2019; Barber et al., 2023; Auer et al., 2023; Chen et al., 2024). The weights $\{w_t^\tau\}_{\tau=1}^{L+1}$ satisfy the normalization condition $\sum_{\tau=1}^{L+1} w_t^\tau = 1$ with each $w_t^\tau \in [0, 1]$. AFOCP applies a weight assignment strategy that follows an attention mechanism detailed in Section 3.2.

To maintain the long-term reliability, the unreliability level α_t in (11) is updated via an online update rule similar to (3), i.e.,

$$\alpha_{t+1} = \alpha_t + \lambda(\alpha - \mathrm{err}_t^{\mathrm{f}}), \tag{12}$$

where λ denotes the step size and the discrete error $\mathrm{err}_t^{\mathrm{f}}$ is defined as

$$\mathrm{err}_t^{\mathrm{f}} = \begin{cases} 1 & \text{if } Y_t \notin \Gamma_t^{\mathrm{AFOCP}}(X_t), \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

3.2 Online Update of Weights via Multi-Head Attention

In AFOCP, the weights $\{w_t^\tau\}_{\tau=1}^{L+1}$ in (11) are obtained via a multi-head attention mechanism based on an MoE architecture, which aims at capturing the feature-space similarity between the current feature vector $f(X_t)$ and the feature vectors of the L most recent observations, i.e.,

$$\hat{V}_{t-1:t-L} = [f(X_{t-1}), \ldots, f(X_{t-L})]^{\mathrm{T}}. \tag{14}$$

The mechanism is parameterized by a number $M \geq 1$ of experts.

For each expert $m \in \{1, \ldots, M\}$, we compute a per-expert, i.e., per-head, attention vector

$$\left[a_t^{(m),1}, \ldots, a_t^{(m),L}\right]^{\mathrm{T}} = \mathrm{attention}^{(m)}\big(f(X_t), \hat{V}_{t-1:t-L}\big) \in \mathbb{R}^{1\times L}, \tag{15}$$

where each attention coefficient $a_t^{(m),l}$ quantifies the similarity of the feature vectors $f(X_t)$ and $f(X_{t-l})$ under expert m. Specifically, given two sequences $\{U_1, \ldots, U_{L_q}\}$ and $\{V_1, \ldots, V_{L_k}\}$ with $U_i \in \mathbb{R}^D$ and $V_j \in \mathbb{R}^D$, the per-expert attention operator returns the $\mathbb{R}^{L_q \times L_k}$ matrix whose (i, j)-th entry is

$$\mathrm{attention}^{(m)}\big(\{U_1, \ldots, U_{L_q}\}, \{V_1, \ldots, V_{L_k}\}\big)_{(i,j)} = \frac{\exp\big(\beta\langle U_i W_q^{(m)}, V_j W_k^{(m)}\rangle + \Delta_{ij}^{(m)}\big)}{\sum_{j'=1}^{L_k} \exp\big(\beta\langle U_i W_q^{(m)}, V_{j'} W_k^{(m)}\rangle + \Delta_{ij'}^{(m)}\big)}, \tag{16}$$

where $W_q^{(m)} \in \mathbb{R}^{D\times D'}$ and $W_k^{(m)} \in \mathbb{R}^{D\times D'}$ are the per-expert learned query and key embedding matrices with latent dimension D'; $\beta > 0$ is a scaling factor; and $\langle\cdot,\cdot\rangle$ denotes the standard Euclidean inner product on $\mathbb{R}^{D'}$. Here $\mathcal{N}_i^{(m)} \subseteq \{1, \ldots, L_k\}$ indexes the K keys most similar to query i,

$$\mathcal{N}_i^{(m)} = \underset{j\in\{1,\ldots,L_k\}}{\text{Top-}K} \big\langle U_i W_q^{(m)}, V_j W_k^{(m)}\big\rangle, \tag{17}$$

that is, the indices of the K largest similarity scores, recomputed at every step so that different experts may attend to different keys. The mask $\Delta_{ij}^{(m)}$ equals 0 for $j \in \mathcal{N}_i^{(m)}$ and $-\infty$ otherwise, and the $-\infty$ entries

vanish under the softmax, so each query distributes its attention only over $\mathcal{N}_i^{(m)}$. In (15), these keys are the L previous feature vectors, so $\mathcal{N}_i^{(m)}$ is the set of lags l that expert m attends to.

The M per-expert attention vectors are then combined through a softmax gate

$$\pi_m\big(f(X_t)\big) = \frac{\exp\big(\langle \theta_m, f(X_t)\rangle\big)}{\sum_{m'=1}^{M} \exp\big(\langle \theta_{m'}, f(X_t)\rangle\big)},\ m = 1, \ldots, M, \tag{18}$$

parameterized by vector $\{\theta_m\}_{m=1}^M \subseteq \mathbb{R}^D$, yielding a single attention vector

$$a_t^\tau = \sum_{m=1}^{M} \pi_m\big(f(X_t)\big) a_t^{(m),\tau},\ \tau = 1, \ldots, L. \tag{19}$$

In the output space, the multi-expert case ($M > 1$) of this mechanism corresponds to a lightweight instance of Heurich et al. (2026), using its sparse retrieval and MoE gate but without the hypernetwork.

The combined attention coefficients are then re-normalized to obtain the weights used in (11) as

$$\big[w_t^1, \ldots, w_t^L\big]^{\mathrm{T}} = \frac{L}{L+1} \cdot \big[a_t^1, \ldots, a_t^L\big]^{\mathrm{T}}, \tag{20a}$$

$$\text{and } w_t^{L+1} = \frac{1}{L+1}, \tag{20b}$$

where weight w_t^{L+1} is assigned to the $+\infty$ point mass in (11). A larger weight w_t^τ indicates greater relevance of the τ-th latest historical feature $\hat{V}_{t-\tau}$ to $f(X_t)$.

The attention mechanism (16) depends on the per-expert embedding matrices $\{W_q^{(m)}, W_k^{(m)}\}_{m=1}^M$ together with the gate parameters $\{\theta_m\}_{m=1}^M$ in (18). We propose to optimize these parameters in an online fashion. To elaborate, for each time t, denote the NC scores evaluated on the most recent L pairs $\{(X_{t-\tau}, Y_{t-\tau})\}_{\tau=1}^L$ as

$$\begin{aligned} S_{t-1:t-L}^{\mathrm{f}} &= \big[s^{\mathrm{f}}(X_{t-1}, Y_{t-1}), \ldots, s^{\mathrm{f}}(X_{t-L}, Y_{t-L})\big]^{\mathrm{T}}, \\ &\text{with } S_t^{\mathrm{f}} = s^{\mathrm{f}}(X_t, Y_t). \end{aligned} \tag{21}$$

The training loss takes the general form

$$\mathcal{L}_t = \ell_t\big(S_t^{\mathrm{f}}, \{w_t^\tau\}_{\tau=1}^{L+1}\big) - \kappa H\big(\pi(f(X_t))\big), \tag{22}$$

combining a data-fitting term ℓ_t dependent on the current NC score S_t^{f} and on the attention weights $\{w_t^\tau\}_{\tau=1}^{L+1}$ with an entropy regularizer of strength $\kappa \geq 0$ on the combining distribution π used in (19). This term, given by the Shannon entropy $H(\pi) = -\sum_{m=1}^M \pi_m \log \pi_m$, discourages the combination (19) from collapsing onto a single expert. The loss ℓ_t may be chosen as the squared prediction error (Auer et al., 2023; Chen et al., 2024)

$$\ell_t^{\mathrm{sq}} = \big(S_t^{\mathrm{f}} - \hat{S}_t^{\mathrm{f}}\big)^2,\ \text{with } \hat{S}_t^{\mathrm{f}} = \sum_{\tau=1}^{L} a_t^\tau\, S_{t-\tau}^{\mathrm{f}}, \tag{23}$$

or as the pinball loss (Heurich et al., 2026)

$$\ell_t^{\mathrm{pb}} = \rho\big(S_t^{\mathrm{f}} - \hat{q}_t^{\mathrm{f}}\big),\ \text{with } \rho(u) = \max\{(1-\alpha)u, -\alpha u\} \text{ and } \hat{q}_t^{\mathrm{f}} = Q_{1-\alpha}\Big(\sum_{\tau=1}^{L} w_t^\tau \delta_{S_{t-\tau}^{\mathrm{f}}} + w_t^{L+1}\delta_{+\infty}\Big). \tag{24}$$

The former encourages the weighted mean of past NC scores to predict the current score, while the latter directly targets the weighted quantile used in (11) at prediction time. The attention parameters $\{W_q^{(m)}, W_k^{(m)}\}_{m=1}^M$ and $\{\theta_m\}_{m=1}^M$ are updated online by gradient descent to minimize the objective (22).

Algorithm 1 AFOCP

Input: Data stream $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$; pre-trained machine learning model $\mu(\cdot) = g \circ f(\cdot)$; target long-term miscoverage α; step sizes λ and η; scaling factor β; number of gradient steps N

1: **for** each time step t **do**
2: Compute weights $\{w_t^\tau\}_{\tau=1}^{L+1}$ via the attention mechanism (15)
3: Calculate the feature-based NC scores $\left\{s^{\mathrm{f}}(X_{t-\tau}, Y_{t-\tau})\right\}_{\tau=1}^{L}$ using (25)
4: Form the weighted empirical distribution in (11) and compute its $(1-\alpha_t)$-quantile
5: Construct the prediction set $\Gamma_t^{\mathrm{AFOCP}}(X_t)$ in (11) (see Section 3.1)
6: Update the miscoverage level α_t using (12)
7: Train the attention mechanism attention$(\cdot,\cdot)$ online by minimizing the loss (22) over $\{W_q^{(m)}, W_k^{(m)}\}_{m=1}^{M}$ and $\{\theta_m\}_{m=1}^{M}$

3.3 Construction of the Prediction Set

A practical challenge in constructing the prediction set $\Gamma_t^{\mathrm{AFOCP}}(X_t)$ in (11) lies in evaluating the feature-based NC score $s^{\mathrm{f}}(X, Y)$ in (10) due to the need to evaluate the infimum operator. Following Teng et al. (2022), we approximate the score $s^{\mathrm{f}}(X, Y)$ using gradient descent in the feature space. Initializing the solution V at the model feature $f(X)$ and using a step size $\eta > 0$ yields the update

$$V \leftarrow V - \eta \nabla_V \|g(V) - Y\|^2. \tag{25}$$

We stop the gradient descent after a fixed number N of iterations, and denote the final iterate by $\bar{V}$. The NC score is then approximated as $S^{\mathrm{f}} \approx \|\bar{V} - f(X)\|$. Note that this is an upper bound on the true NC score $s^{\mathrm{f}}(X, Y)$.

Another practical challenge is the composition of the set (11) given the most recent L feature-based NC scores $\left\{s^{\mathrm{f}}(X_{t-\tau}, Y_{t-\tau})\right\}_{\tau=1}^{L}$. In fact, when the label space $\mathcal{Y}$ is discrete and finite, as in classification tasks, one can construct the set (11) by enumerating the possible labels $Y \in \mathcal{Y}$. In contrast, when $\mathcal{Y}$ is continuous, as in regression tasks, it is not generally feasible to calculate the NC scores $s^{\mathrm{f}}(X_t, Y)$ across all candidates $Y \in \mathcal{Y}$. Following Teng et al. (2022), we instead adopt a band estimation strategy based on linear relaxation based perturbation analysis (LiRPA) (Xu et al., 2020). This scheme provides a certified upper-bound approximation of prediction set (11) via linear relaxation. The resulting interval is then a computationally efficient outer approximation of $\Gamma_t^{\mathrm{AFOCP}}(X_t)$. We refer to Teng et al. (2022) for details.

3.4 Special Cases of AFOCP

In order to isolate and evaluate the effectiveness of feature-based NC scores and attention-based weights, we also introduce two intermediate variants of OCP, namely, AOCP and FOCP, as illustrated in Figure 1. These intermediate variants also serve as online counterparts of representative offline CP methods recently developed for time series (Auer et al., 2023; Teng et al., 2022; Chen et al., 2024), as elaborated below.

3.4.1 AOCP

NC scores are computed in the output space as in OCP, but uniform weights are replaced by attention-based adaptive weights. The prediction set is defined as

$$\Gamma_t^{\mathrm{AOCP}}(X_t) = \left\{ Y \in \mathcal{Y} : s(X_t, Y) \le Q_{1-\alpha_t}\left(\sum_{\tau=1}^{L} w_t^\tau \delta_{s(X_{t-\tau}, Y_{t-\tau})} + w_t^{L+1} \delta_{+\infty} \right) \right\}. \tag{26}$$

Unlike OCP, which uses uniform weights for the last L NC scores in (2), AOCP assigns data-dependent weights via an attention mechanism. The weight generation follows the same procedure as AFOCP described in Section 3.2, except that the NC scores in the loss function (22) are computed using (1) in the output space instead of the feature space. AOCP can thus be regarded as an online version of the output-space attention-based WCP method of Auer et al. (2023), in which the attention module is updated online instead of being pre-trained, and the miscoverage level is updated via (3).

3.4.2 FOCP

The weighting scheme remains uniform as in OCP, but NC scores are evaluated in the semantic feature space. The prediction set is given by

$$\Gamma_t^{\mathrm{FOCP}}(X_t) = \left\{ Y \in \mathcal{Y} : s^{\mathrm{f}}(X_t, Y) \leq Q_{1-\alpha_t} \left(\sum_{\tau=1}^{L} \frac{1}{L+1} \delta_{s^{\mathrm{f}}(X_{t-\tau}, Y_{t-\tau})} + \frac{1}{L+1} \delta_{+\infty} \right) \right\}. \tag{27}$$

That is, FOCP replaces the output-level NC score calculation in (1) with the feature-level counterpart defined in (10). Following the same perspective, FOCP corresponds to an online version of feature-space CP (Teng et al., 2022), with the fixed split-conformal calibration set replaced by a sliding window and the miscoverage level updated via (3).

The full AFOCP scheme accordingly corresponds to an online version of the feature-space attention-based WCP method of Chen et al. (2024), with the attention module updated online via (22) rather than pre-trained, and the miscoverage level updated via (3). The resulting online procedure admits both a long-term coverage guarantee and a time-averaged interval-length improvement, analyzed in Section 3.5.

3.5 Theoretical Guarantees

We now demonstrate that the proposed AFOCP scheme, summarized in Algorithm 1, ensures the desired long-term coverage, while being provably more efficient in the long run than the vanilla OCP reviewed in Section 2. First, similar to Theorem 1, we have the following reliability guarantee.

Corollary 1. *The average error over $T \in \mathbb{N}$ obtained by FOCP and AFOCP satisfies the equality*

$$\frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{f}} = \alpha + \frac{\alpha_1 - \alpha_{T+1}}{T\lambda}. \tag{28}$$

In particular, we have the limit

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{f}} = \alpha. \tag{29}$$

Proof: The proof follows directly from Theorem 1.

Second, to compare efficiency in terms of prediction interval length, we proceed under informal assumptions analogous to those underlying Theorem 4 in Teng et al. (2022), together with a mild temporal-dependence condition on the data stream. While a formal statement can be found in Appendix A, these assumptions essentially require that: (*i*) the output-space quantiles induced by the feature-space construction remain close to those based directly on output-space lengths; (*ii*) the mapping from feature space to output space amplifies deviations between individual lengths and their quantiles; and (*iii*) the resulting output-space quantiles are stable across calibration windows, with fluctuations diminishing as the window grows. This yields the following result.

Theorem 2. *Under the regularity conditions of Teng et al. (2022); Chen et al. (2024) together with a mild condition on the temporal dependence of the data stream (see Appendix A), the time-averaged prediction set sizes of FOCP and AFOCP over $T \in \mathbb{N}$ can be upper bounded by those of OCP and AOCP, respectively, as*

$$\frac{1}{T} \sum_{t=1}^{T} \left| \Gamma_t^{\mathrm{AFOCP/FOCP}}(X_t) \right| \leq \frac{1}{T} \sum_{t=1}^{T} \left| \Gamma_t^{\mathrm{AOCP/OCP}}(X_t) \right| + \Delta(T), \tag{30}$$

where $\Delta(T) = O(T^{-1/2})$. In particular, we have the long-term limit

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \left| \Gamma_t^{\mathrm{AFOCP/FOCP}}(X_t) \right| \leq \limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \left| \Gamma_t^{\mathrm{AOCP/OCP}}(X_t) \right|. \tag{31}$$

Proof: Appendix A.

3.6 Computational Complexity

We analyze the per-step computational complexity of AFOCP relative to OCP. Both methods share a forward pass of the base predictor μ in (7) on the test point and an $O(L \log L)$ quantile computation over the calibration window. Following the shallow two-layer architecture of Teng et al. (2022) and Chen et al. (2024), one forward-backward pass through the prediction head g costs $O(D^2)$, and one LiRPA call on g incurs the same order of complexity. Beyond the shared cost, AFOCP introduces three additional steps: (i) for each of the L historical samples, N gradient descent iterations through the prediction head g to approximate the feature-space NC score in (25), contributing $O(LND^2)$; (ii) a single LiRPA-based band estimation step (Xu et al., 2020) per test point, which maps the feature-space quantile back to the output space, as detailed in Section 3.3, contributing $O(D^2)$; and (iii) one online gradient update of the attention parameters $\{W_q^{(m)}, W_k^{(m)}\}_{m=1}^M$ in (16) and $\{\theta_m\}_{m=1}^M$ in (18) on the loss (22), which is dominated by ML similarity computations of cost $O(DD')$ each, contributing $O(MLDD')$. The update of α_t in (3) is $O(1)$.

Summing these contributions, the per-step complexity of AFOCP scales as

$$O\left(L \log L + LND^2 + MLDD'\right), \tag{32}$$

i.e., linearly in the window length L, the number of inner gradient steps N for NC score approximation, the number of experts M, and the embedding dimension D', and quadratically in the feature dimension D. In exchange, AFOCP yields tighter prediction intervals as established in Theorem 2 and confirmed by the experiments in Section 4.

4 Experimental Results

In this section, we empirically evaluate AFOCP and the baselines OCP, FOCP, and AOCP on one synthetic and four real-world time-series datasets. As noted in Section 3, these four schemes correspond to online versions of four representative CP designs for time series, namely *(i)* OCP (Gibbs & Candes, 2021); *(ii)* feature-space CP (Teng et al., 2022) as FOCP; *(iii)* output-space attention-based WCP (Auer et al., 2023) as AOCP; and *(iv)* feature-space attention-based WCP (Chen et al., 2024) as AFOCP. Their comparison thus simultaneously isolates the contributions of feature-space calibration and attention-based weighting, and benchmarks AFOCP against the online versions of these existing designs. We report time-averaged coverage and time-averaged prediction interval length to verify the long-term coverage guarantees and to assess the efficiency of the resulting prediction sets. By varying the calibration window length and feature dimension, we also isolate the impact of attention-based adaptive weighting and feature-space calibration.

4.1 Setting

4.1.1 Datasets

Following the evaluation setting in Chen et al. (2024), we analyze the performance of OCP, FOCP, AOCP and AFOCP (see Section 3) on synthetic data and four real-world benchmark time-series datasets:

- **Synthetic data (Auer et al., 2023):** We generate a multivariate time series of length 1500 with alternating segments of variable lengths with different inputs and noise distributions. In particular, at each time step t, the input $X_t \in \mathbb{R}^{50}$ and target $Y_t \in \mathbb{R}^{50}$ are related as $Y_t = 10 + HX_t + \varepsilon_t$, where $H \in \mathbb{R}^{50\times 50}$ has i.i.d. entries drawn from $\mathcal{N}(0, 1/50)$. Each segment length is drawn uniformly from the set $\{40, 41, 42, \ldots, 80\}$. In a segment, the input X_t is a vector of all entries equal to 3 and the noise is $\varepsilon_t \sim \mathcal{N}(0, X_t/2\mathbf{I}_{50})$; in the next segment, the next input X_t is a vector of all entries equal to 21 and the noise ε_t has i.i.d. entries sampled from $\mathcal{U}(-X_t, X_t)$.
- **Air quality (Zhang et al., 2017; Auer et al., 2023):** This dataset provides hourly air quality and meteorological measurements from the Tiantan station in Beijing from March 2013 to February 2017, totaling 35064 timestamps. At each time step t, we predict the current particulate-matter concentration from an 11-dimensional input X_t built from other pollutant indicators, temperature,

pressure, dew point, precipitation, wind speed, and wind direction encoded as two orthogonal components scaled to $[-1, 1]$. The scalar target Y_t alternates by contiguous segments between PM10 and PM2.5, with segment lengths uniformly sampled from the set $\{40, 41, 42, \ldots, 80\}$, modeling a single sensor that intermittently measures the two particulate indicators in separate intervals.

- **Electricity (Harries et al., 1999; Barber et al., 2023):** This dataset comprises half-hourly records of electricity price, demand, and transfer for New South Wales and Victoria from 7 May 1996 to December 1998. Inputs X_t are given by the state-level electricity prices and demands. The prediction target Y_t is transfer, defined as the amount of electricity exchanged between the two states. We retain observations in the time slot 09 : 00–12 : 00 and discard an initial transient period with constant transfer, yielding 3444 time points.
- **Bike-sharing (Teng et al., 2022):** This dataset contains daily records from an urban bike-sharing system from 2011 to 2012, comprising 731 days. At each time step t, the input X_t has 16 features capturing calendar context, such as year, month, weekday, holiday status, working day status, as well as weather conditions, such as air temperature, perceived temperature, humidity, wind speed, and a coarse weather category. The target Y_t is the total number of rentals per day.
- **Wind speed (Xu & Xie, 2021; Dong et al., 2021):** This dataset contains wind speed measurements from wind farms operated by the Midcontinent Independent System Operator in the United States, sampled every 15 minutes over one week in September 2020 and comprising 764 timestamps. At each time step t, we use ten input features X_t and a two-component target Y_t, with the inputs capturing contemporaneous values and short-term lags, and the targets describing future wind conditions.

For datasets exceeding 2000 samples, we deterministically downsample to 2000 evenly spaced observations while preserving temporal order, keeping the sequential online evaluation computationally tractable. We use an 85%/15% train/test split and report results averaged over five random seeds for robustness.

4.1.2 Model training

We train a two-stage neural network (7) for regression, where both the feature extractor $f(\cdot)$ and the prediction head $g(\cdot)$ are two-layer fully connected networks with ReLU activations and hidden size D. The model is trained using the mean squared error (MSE) loss, optimized by Adam with a learning rate of 5×10^{-4}, a weight decay of 10^{-6}, and a batch size of 64 for 10 epochs.

We further pre-train an attention module that assigns dynamic weights to historical observations based on feature similarity. The input embeddings, produced by $f(\cdot)$, have dimension D and are mapped into query and key vectors of dimension $D' = 32$ with scaling $\beta = 1/\sqrt{D'}$, as in (16). The module is pre-trained on the training set over a sliding window of length L by optimizing (22) with Adam (learning rate 10^{-3}, weight decay 10^{-6}) for 80 epochs. During deployment, after each prediction the window slides forward by one step and the module is fine-tuned online on the updated window (learning rate 10^{-4}), enabling real-time adaptation to distribution shifts. The miscoverage level is updated online with step size $\gamma = \lambda = 0.05$, and each feature-space NC score (25) is approximated by $N = 80$ gradient steps of step size $\eta = 10^{-3}$.

Unless otherwise specified, AOCP and AFOCP use single-head attention ($M = 1$). We additionally report results for $M = 5$ experts. Each expert uses a top-$K = 15$ sparse support set $\mathcal{N}_i^{(m)}$ and ℓ_2-normalized query and key projections measuring cosine similarity, with an entropy coefficient $\kappa = 0.01$ in (22) following Heurich et al. (2026). We train $M = 1$ with the squared error (23) and $M = 5$ with the pinball loss (24).

4.1.3 Evaluation metrics

We adopt two standard metrics in OCP: time-averaged coverage and time-averaged prediction interval length, which assess reliability and efficiency, respectively.

Consider a multi-dimensional response $Y_t = (Y_t^{(1)}, \ldots, Y_t^{(i)}, \ldots, Y_t^{(I)}) \in \mathbb{R}^I$ and its prediction set $\Gamma_t^*(X_t) \subseteq \mathbb{R}^I$, where the length along each dimension forms a vector $|\Gamma_t^*(X_t)| \in \mathbb{R}^I$, with $* \in$

Table 1: Time-averaged coverage (%) and time-averaged interval length at the end of the test sequence under the default settings $L = 100$, $D = 50$, $\alpha = 0.1$, averaged over five random seeds. The shortest length within each of the output-space and feature-space methods is highlighted in bold.

	Output-space methods						Feature-space methods					
	OCP		AOCP		AOCP ($M = 5$)		FOCP		AFOCP		AFOCP ($M = 5$)	
Dataset	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
Synthetic	90.67	8.277	90.54	**5.969**	89.96	7.342	89.69	2.749	89.79	2.130	89.96	**2.125**
Electricity	91.34	2.022	90.67	**1.651**	90.54	1.971	91.07	0.205	90.54	0.169	90.54	**0.154**
Air quality	90.54	4.200	89.93	3.318	90.40	**2.820**	90.67	0.547	89.87	**0.443**	90.47	0.541
Bike sharing	98.15	3.022	91.30	2.729	93.52	**2.678**	96.85	0.313	89.63	**0.283**	91.67	0.314
Wind	93.10	4.805	90.17	**3.910**	89.20	4.347	92.21	0.499	90.73	0.418	90.62	**0.400**

$\{\text{OCP}, \text{AOCP}, \text{FOCP}, \text{AFOCP}\}$. The two metrics are defined as

$$\text{Time-averaged coverage} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left\{Y_t \in \Gamma_t^*\left(X_t\right)\right\}, \tag{33a}$$

$$\text{Time-averaged interval length} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{I}\sum_{i=1}^{I} \left|\Gamma_t^*\left(X_t\right)\right|_{(i)}, \tag{33b}$$

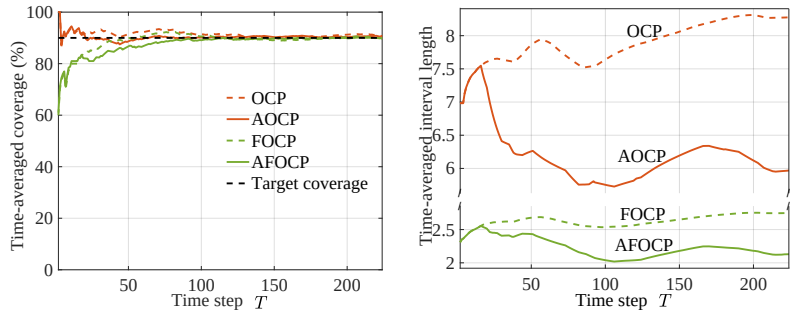
where the subscript (i) in (33b) denotes the i-th dimension of vector $\Gamma_t^*(X_t)$, i.e., the prediction interval corresponding to response $Y_t^{(i)}$. The indicator function $\mathbb{1}\{\cdot\}$ in (33a) takes 1 if all I entries of Y_t lie within the prediction interval, and (33b) also averages interval length across dimensions.

4.2 Performance Evaluation

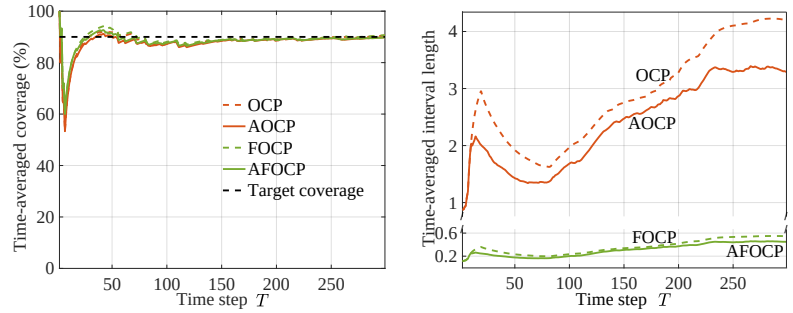
We first compare the four methods OCP, FOCP, AOCP, and AFOCP, with AOCP and AFOCP additionally evaluated at $M = 5$ experts, choosing the window length as $L = 100$, the feature dimension as $D = 50$, and the target miscoverage rate as $\alpha = 0.1$. Table 1 reports the time-averaged coverage and interval length at the end of the test sequence, averaged over five random seeds, with the shortest length within each space in bold. Across datasets, the long-term coverage of all methods remains close to the target level, as predicted by Corollary 1. In terms of efficiency, the attention-based schemes consistently produce substantially shorter intervals than OCP and FOCP at comparable coverage. Within each space, however, $M = 1$ and $M = 5$ do not dominate each other. AOCP is shortest on three datasets versus two for AOCP ($M = 5$) in the output space, and AFOCP ($M = 5$) is shortest on three versus two for AFOCP in the feature space. All pairwise gaps between $M = 1$ and $M = 5$ remain within a factor of 1.3 across the five datasets.

The additional sparse top-K support and MoE gate introduce further trainable parameters whose benefit is reported to scale with the size of the calibration stream (Heurich et al., 2026). The calibration windows used here (at most 2000 points) are too small for these components to bring stable gains over single-head attention. The two attention designs therefore yield comparable interval lengths in both spaces. Given that AFOCP with $M = 5$ offers no consistent length advantage over AFOCP with $M = 1$ while requiring substantially more trainable parameters and higher per-step inference cost (Section 3.6), the subsequent analyses focus on AOCP and AFOCP under the default $M = 1$. The conclusions on window length and feature dimension below concern the attention mechanism and apply equally to $M = 5$.

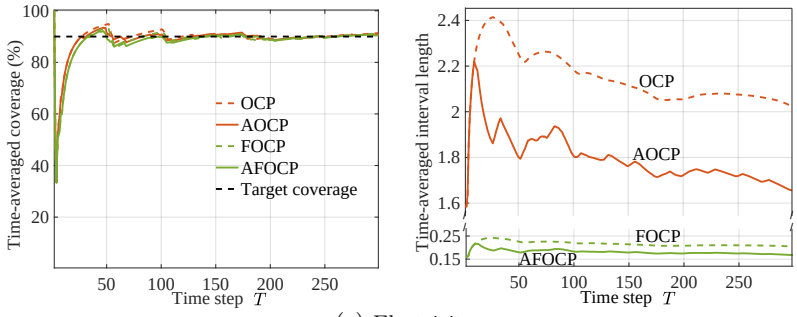
We evaluate OCP, AOCP, FOCP, and AFOCP on the five datasets described in Section 4.1.1 by showing the time-averaged coverage in (33a) and the time-averaged interval length in (33b) as a function of time T in Figure 2. For all datasets, we choose window length as $L = 100$, the feature dimension as $D = 50$, and the target miscoverage rate as $\alpha = 0.1$ (dashed line). Across datasets, the long-term coverage of all methods converges to the target level, confirming the theoretical results in Corollary 1. In terms of efficiency, feature-space calibration (FOCP and AFOCP) yields substantially shorter time-averaged interval lengths than output-space calibration (OCP and AOCP), highlighting the advantage of operating in the space of



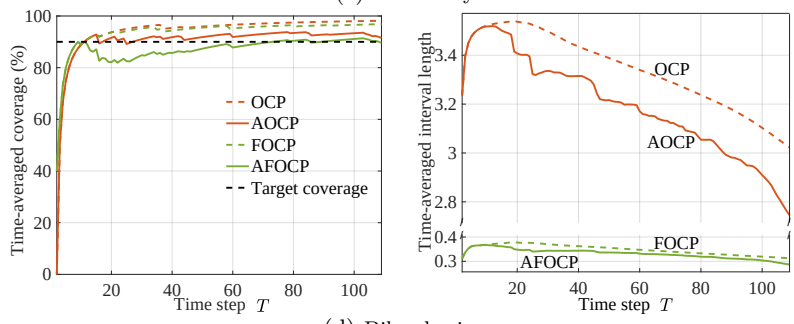
(a) Synthetic data



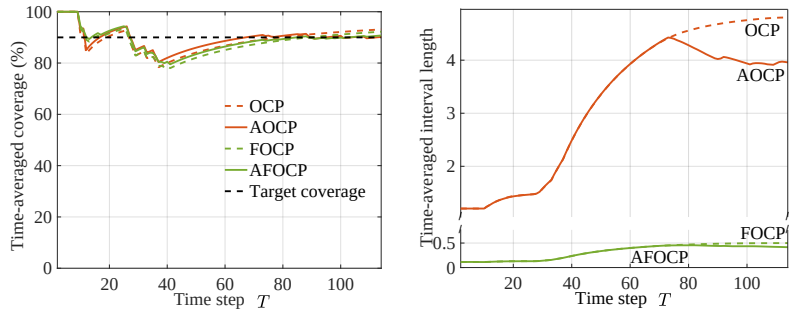
(b) Air quality



(c) Electricity

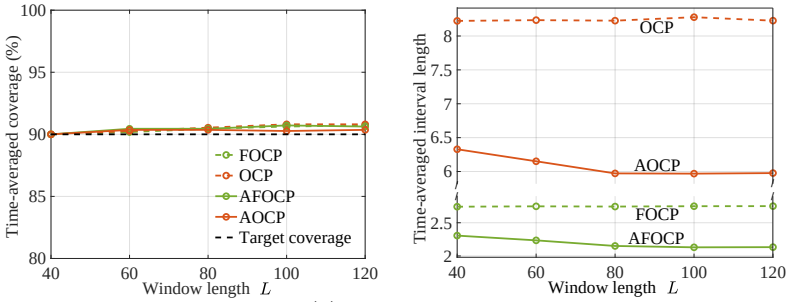


(d) Bike-sharing

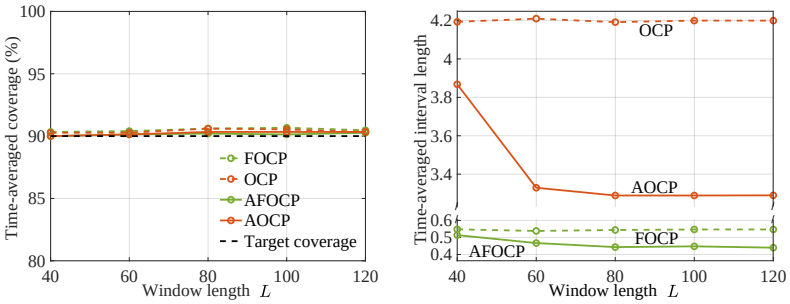


(e) Wind

Figure 2: Time-averaged coverage (left) and time-averaged interval length (right) of OCP, FOCP, AOCP, and AFOCP versus time T across various datasets, with window length $L = 100$, feature dimension $D = 50$, and target miscoverage rate $\alpha = 0.1$.



(a) Synthetic data



(b) Air quality

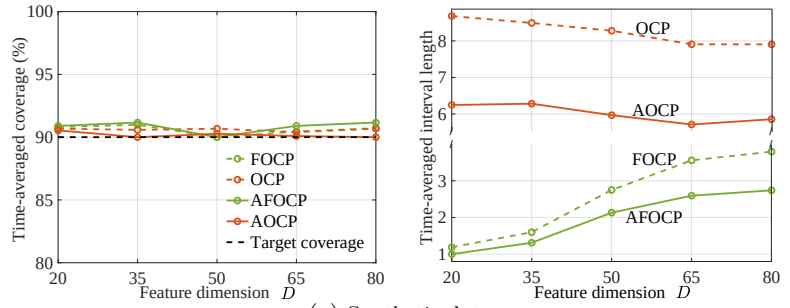
Figure 3: Time-averaged coverage (left) and time-averaged interval length (right) of OCP, FOCP, AOCP, and AFOCP versus window length L for synthetic data and air quality datasets with feature dimension $D = 50$ and target miscoverage rate $\alpha = 0.1$.

task-aligned representations. Incorporating attention for adaptive weighting (AOCP and AFOCP) further reduces interval lengths compared to uniform weighting (OCP and FOCP). The latter gain is particularly pronounced for the synthetic dataset, which presents a higher degree of non-stationarity as compared to the other datasets.

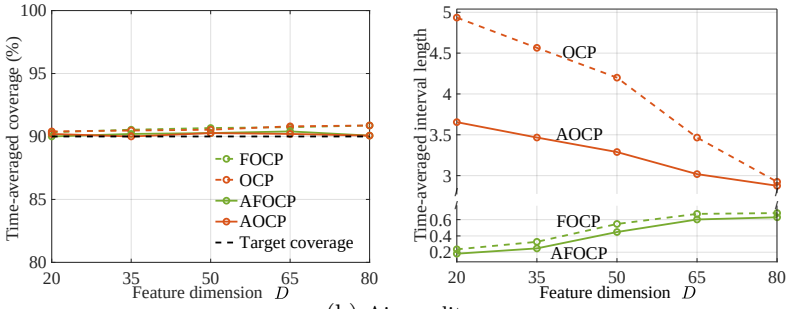
Figure 3 and Figure 4 report the performance at time T equal to the test sequence length for each dataset of OCP, AOCP, FOCP, and AFOCP versus the window length L with $D = 50$ and $\alpha = 0.1$, and versus the feature dimension D with $L = 100$ and $\alpha = 0.1$. For the window-length analysis, we include only the synthetic and air-quality datasets, which share a similar structure in alternating segments, making the results easier to interpret (see Section 4.1.1). As seen in the figures, coverage remains close to the nominal level across all settings. Furthermore, feature-space calibration (FOCP and AFOCP) consistently achieves shorter converged intervals than output-space calibration (OCP and AOCP), while attention-based weighting (AOCP and AFOCP) provides additional reductions over each uniform weighting baseline (OCP and FOCP).

As also shown in Figure 3, increasing the window size L decreases the time-averaged interval length for AOCP and AFOCP up to $L = 80$, after which further gains are negligible. This threshold coincides with the maximum segment length over which the process is approximately stationary, as defined in Section 4.1.1. Once the window covers a full segment, adding more history provides minimal gains. The use of attention is particularly important when the window tends to contain a mix of different segments, i.e., when $L > 80$, as in this case, attention can learn to assign larger weights to features within the same segment.

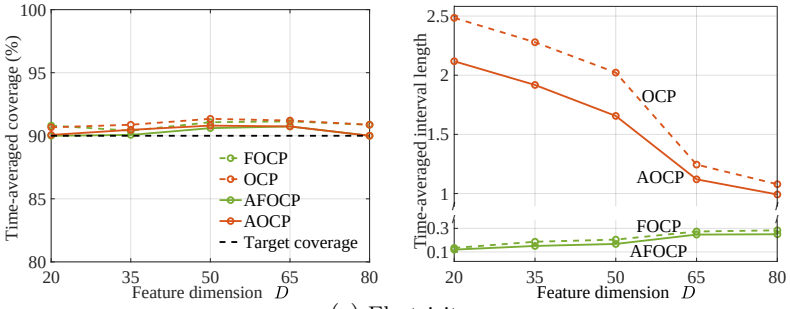
Finally, as shown in Figure 4, increasing the feature space dimension D widens the intervals for feature-space calibration (FOCP and AFOCP), but narrows them for output-space calibration (OCP and AOCP). In the former case, a larger dimension D has the dominant effect of introducing additional nuisance variations and amplifying the approximation error in the inverse mapping used for feature-level NC scores, thus increasing score dispersion and the resulting quantiles. In the latter case, a larger dimension D benefits performance due to improved predictive fit of the underlying model, reducing residual errors and thus the output-level quantiles.



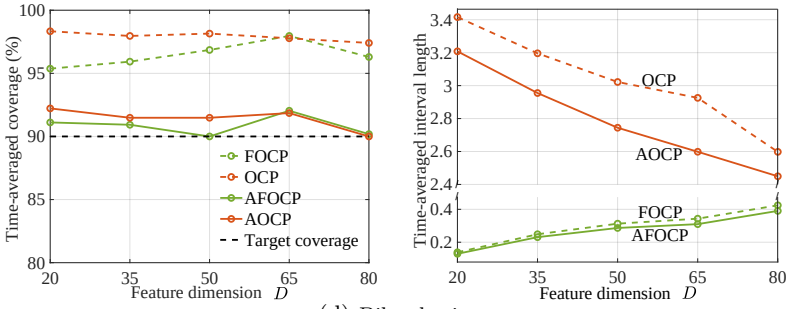
(a) Synthetic data



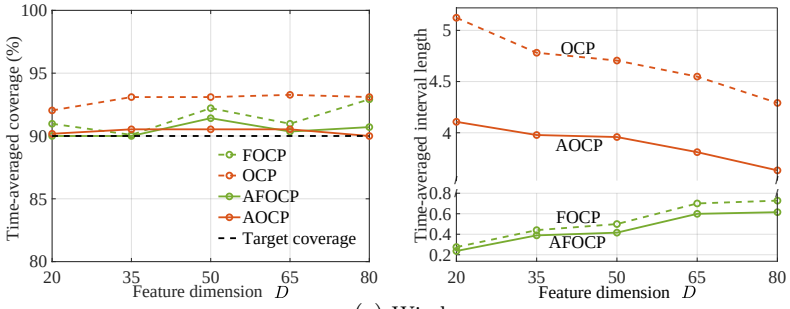
(b) Air quality



(c) Electricity



(d) Bike-sharing



(e) Wind

Figure 4: Time-averaged coverage (left) and time-averaged interval length (right) of OCP, FOCP, AOCP, and AFOCP versus feature dimension D across various datasets with window length $L = 100$ and target miscoverage rate $\alpha = 0.1$.

Complementary analyses are provided in the Appendix, covering the effect of different choices of training loss (Appendix B), per-seed variability of the long-term metrics (Appendix C), rolling-window local behavior (Appendix D), coverage behavior around regime changes (Appendix E), and a comparison of AFOCP with its offline feature-space CP counterpart (Appendix F)

5 Conclusions

This work introduced AFOCP, a principled extension of OCP that calibrates uncertainty in the learned feature space and adaptively reweights historical observations via an attention mechanism. By shifting calibration from the output space to task-aligned representations and replacing uniform aggregation with relevance weights, AFOCP reduces nuisance variation to concentrate NC scores in feature space and adapts the quantile to the current test point. The result is a simple, modular, and effective approach to reliable and efficient uncertainty quantification for non-stationary time series.

We established two guarantees. First, AFOCP attains the target long-term coverage. Second, under mild regularity conditions, AFOCP yields smaller long-term time-averaged prediction sets than its output-space counterpart. Experiments on synthetic and real-world time series support the theory: AFOCP maintains nominal coverage while substantially reducing the prediction interval length relative to OCP and to the online counterparts of representative offline CP designs for time series. The intermediate variants FOCP and AOCP further isolate the contributions of feature-space calibration and attention-based weighting.

Future work includes richer attention architectures such as cross-attention and multi-scale designs (Vaswani et al., 2017), adaptive selection of calibration history, and more efficient streaming implementations. Beyond attention, the proposed framework can in principle accommodate alternative weight-learning schemes, including kernel-based reweighting (Lee et al., 2025) and query-conditioned retrieval metrics in the spirit of Heurich et al. (2026), while the OCP-style update in (3) may be replaced by proportional-integral-derivative controllers (Angelopoulos et al., 2023) for improved responsiveness under sudden distribution shifts. Furthermore, extending the analysis to broader model classes, weaker assumptions, higher-dimensional inputs such as images and video, panel data with multiple correlated time series (Tu & Giesecke, 2026), structured outputs, and multivariate coverage criteria (Principato et al., 2026) is a promising direction.

References

Toghrul Abbasli, Kentaroh Toyoda, Yuan Wang, Leon Witt, Muhammad Asif Ali, Yukai Miao, Dan Li, and Qingsong Wei. Comparing uncertainty measurement and mitigation methods for large language models: A systematic review. *arXiv preprint arXiv:2504.18346*, 2025.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1–14, 2023.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56027–56074, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 2337–2363, 2023.

Anirban Bhattacharyya and Rina Foygel Barber. Group-weighted conformal prediction. *arXiv preprint arXiv:2401.17452*, 2024.

Richard C Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Baiting Chen, Zhimei Ren, and Lu Cheng. Conformalized time series with semantic features. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121449–121474, 2024.

Kfir M Cohen, Sangwoo Park, Osvaldo Simeone, and Shlomo Shamai Shitz. Cross-validation conformal risk control. In *IEEE International Symposium on Information Theory*, pp. 250–255, 2024.

Zhaoyu Dong, Haiwang Zhang, Shengyu Zhu, Yao Xie, and Pascal Van Hentenryck. Multi-resolution spatio-temporal prediction with application to wind power generation. *arXiv preprint arXiv:2108.13285*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56:1513–1589, 2023.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(22):1–33, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Erfan Hajihashemi and Yanning Shen. Multi-model online conformal prediction with graph-structured feedback. *Transactions on Machine Learning Research*, 2025.

Michael Harries et al. Splice-2 comparative evaluation: Electricity pricing. 1999.

Manuel Heurich, Maximilian Granz, and Tim Landgraf. Regime-aware retrieval for efficient conformal prediction. *arXiv preprint arXiv:2605.08857*, 2026.

Jiayi Huang, Sangwoo Park, and Osvaldo Simeone. Calibrating bayesian learning via regularization, confidence minimization, and selective inference. *arXiv preprint arXiv:2404.11350*, 2024.

Xiannan Huang and Shuhan Qiu. Feature fitted online conformal prediction for deep time series forecasting model. *arXiv preprint arXiv:2505.08158*, 2025.

Vilde Jensen, Filippo Maria Bianchi, and Stian Normann Anfinsen. Ensemble conformalized quantile regression for probabilistic time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9014–9025, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonghyeok Lee, Chen Xu, and Yao Xie. Kernel-based optimally weighted conformal time-series prediction. In *International Conference on Learning Representations*, 2025.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Zijian Li, Ruichu Cai, Tom ZJ Fu, Zhifeng Hao, and Kun Zhang. Transferable time-series forecasting under causal conditional shift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):1932–1949, 2024.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Conformal prediction intervals with temporal dependence. *Transactions on Machine Learning Research*, 2022.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8(8):5116–5123, 2023.

Lucas Oliveira, Lihua Duan, Aaditya Ramdas, and Rina Foygel Barber. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(153):1–51, 2024.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

Guillaume Principato, Gilles Stoltz, Yvenn Amara-Ouali, Yannig Goude, Bachir Hamrouche, and Jean-Michel Poggi. Conformal prediction for hierarchical data. *Transactions on Machine Learning Research*, 2026.

Emmanuel Rio. *Asymptotic theory of weakly dependent random processes*, volume 80. Springer, 2017.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3543–3553, 2019.

Yusuf Sale and Aaditya Ramdas. Online selective conformal inference: Errors and solutions. *Transactions on Machine Learning Research*, 2025.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Osvaldo Simeone. *Machine learning for engineers*. Cambridge University Press, 2022.

Osvaldo Simeone, Sangwoo Park, and Matteo Zecchin. Conformal calibration: Ensuring the reliability of black-box AI in wireless systems. *arXiv preprint arXiv:2504.09310*, 2025.

Yongyi Su, Xun Xu, Tianrui Li, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering regularized self-training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5524–5540, 2024.

Morgan B Talbot, Rushikesh Zawar, Rohil Badkundri, Mengmi Zhang, and Gabriel Kreiman. Tuned compositional feature replays for efficient stream learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3300–3314, 2025.

Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive inference with feature conformal prediction. *arXiv preprint arXiv:2210.00173*, 2022.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2530–2540, 2019.

Daohong Tu and Kay Giesecke. Online conformal prediction for non-exchangeable panel data. *arXiv preprint arXiv:2605.17705*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (3):1464–1483, 2025.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569, 2021.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pp. 38707–38727, 2023.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1129–1141, 2020.

Wenmian Yang, Lizhi Cheng, Mohamed Ragab, Min Wu, Sinno Jialin Pan, and Zhenghua Chen. A virtual-label-based hierarchical domain adaptation method for time-series classification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11456–11465, 2025.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866, 2022.

Matteo Zecchin and Osvaldo Simeone. Localized adaptive risk control. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8165–8192, 2024.

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Forking uncertainties: Reliable prediction and model predictive control with sequence models via conformal risk control. *IEEE Journal on Selected Areas in Information Theory*, 5:44–61, 2024.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.

A Formal Description and Proof of Theorem 2

Our analysis builds on Theorem 6 in Teng et al. (2022) and Theorem 1 in Chen et al. (2024), with the offline-to-online translation made rigorous via a concentration argument on the data stream. We start by presenting key definitions for AOCP and AFOCP.

At time t, we use the most recent L observed data pairs $\{(X_\tau, Y_\tau)\}_{\tau=t-L}^{t-1}$ to compute NC scores, together with non-negative normalized weights $\{w_t^\tau\}_{\tau=1}^{L+1}$ for weighted quantile calculation.

AOCP. Let $M_{t-L:t-1} = \{M_\tau\}_{\tau=t-L}^{t-1}$ denote the individual lengths in the output space for the latest L data pairs, where $M_\tau = 2s(X_\tau, Y_\tau)$ and $s(X_\tau, Y_\tau)$ is the NC score defined in (1). Thus, the prediction interval length is $Q_{1-\alpha_t}(M_{t-L:t-1})$, where $Q_{1-\alpha_t}(M_{t-L:t-1})$ is the $(1-\alpha_t)$-quantile of the weighted empirical distribution $\sum_{\tau=1}^{L} w_t^\tau \delta_{M_{t-\tau}} + w_t^{L+1}\delta_{+\infty}$, where δ_b is the point mass as in (2).

AFOCP. Let $M^{\mathrm{f}}_{t-L:t-1} = \left\{M^{\mathrm{f}}_\tau\right\}_{\tau=t-L}^{t-1}$ denote the individual lengths in the feature space for the latest L data pairs, where $M^{\mathrm{f}}_\tau = 2s^{\mathrm{f}}(X_\tau, Y_\tau)$ and $s^{\mathrm{f}}(X_\tau, Y_\tau)$ is the feature-based NC score defined in (10). To characterize the prediction length in the output space, we define $\mathcal{H}(M, X)$ as the individual output-space length associated with input X given feature-space length M. Specifically, $\mathcal{H}(M, X)$ represents the length of the set $\{g(U) : \|U - f(X)\| \le M/2\}$, which maps the diameter M, centered at $f(X)$, through the prediction head $g(\cdot)$ and returns the maximal spread of the resulting outputs. Accordingly, the prediction interval length of AFOCP is $\mathcal{H}\big(Q_{1-\alpha_t}(M^{\mathrm{f}}_{t-L:t-1}), X_t\big)$, where $Q_{1-\alpha_t}(M^{\mathrm{f}}_{t-L:t-1})$ calculates the $(1-\alpha_t)$-quantile of the weighted empirical distribution $\sum_{\tau=1}^{L} w_t^\tau \delta_{M^{\mathrm{f}}_{t-\tau}} + w_t^{L+1}\delta_{+\infty}$. Without abuse of notations, operating $\mathcal{H}$ on a dataset means operating $\mathcal{H}$ on each data point in the set, e.g., $\mathcal{H}(M^{\mathrm{f}}_{t-L:t-1}, X_{t-L:t-1}) = \left\{\mathcal{H}(M^{\mathrm{f}}_\tau, X_\tau)\right\}_{\tau=t-L}^{t-1}$.

We assume the data stream $\{(X_t, Y_t)\}_{t\ge1}$ is stationary and α-mixing with summable mixing rate $\sum_{k=1}^{\infty}\alpha(k) < \infty$, a standard weak-dependence condition satisfied by common time-series models such as AR, ARMA, and stationary Markov processes (Bradley, 2005; Rio, 2017). Let $\mathcal{P}$ denote the joint distribution of $L+1$ consecutive observations under this regime, and let P^L denote the marginal distribution of L consecutive observations. We write $(D, X_0) \sim \mathcal{P}$ to denote a window D of L data pairs together with the subsequent test point X_0, and $D \sim P^L$ to denote D alone. We use M_D, M^{f}_D, X_D to denote the analogs of $M_{t-L:t-1}$, $M^{\mathrm{f}}_{t-L:t-1}$, $X_{t-L:t-1}$ defined on D. Following Teng et al. (2022) and Chen et al. (2024), we then present the formal description of Theorem 2.

Theorem 3. *Assume the inequality that there exist constants $\beta > 0$ and $R > 0$ satisfying*

$$|\mathcal{H}(M, X) - \mathcal{H}(M', X)| \le R\,|M - M'|^\beta \tag{34}$$

for all X. Additionally, assume that there exist constants $\epsilon > 0$, $C > 0$ such that for all $\alpha \in [0, 1]$, the following regularity conditions hold:

1) ***Length Preservation:*** *The output-space quantiles induced by the feature-space construction are not significantly larger than the corresponding quantiles based directly on the output-space lengths, namely,*

$$\mathbb{E}_{D\sim P^L}\left[Q_{1-\alpha}(\mathcal{H}(M^{\mathrm{f}}_D, X_D))\right] < \mathbb{E}_{D\sim P^L}\left[Q_{1-\alpha}(M_D)\right] + \epsilon. \tag{35}$$

2) ***Expansion:*** *The operator $\mathcal{H}(M, X)$ expands the differences between individual lengths and their quantiles, namely,*

$$\begin{aligned} &R\cdot\mathbb{E}_{D\sim P^L}\left[\mathbb{M}\left[\,\left|Q_{1-\alpha}(M^{\mathrm{f}}_D) - M^{\mathrm{f}}_D\right|^\beta\,\right]\right] \\ &\quad< \mathbb{E}_{D\sim P^L}\left[\mathbb{M}\left[Q_{1-\alpha}(\mathcal{H}(M^{\mathrm{f}}_D, X_D)) - \mathcal{H}(M^{\mathrm{f}}_D, X_D)\right]\right] - \epsilon - 2\max\{R, 1\}\left(\frac{C}{\sqrt{L}}\right)^{\min\{\beta,1\}}, \end{aligned} \tag{36}$$

where $\mathbb{M}[\cdot]$ denotes the mean of a set.

3) ***Quantile Stability:*** *The quantile of the individual length is stable between a typical window D and a realized window W of L data pairs from the stream, namely,*

$$\mathbb{E}_{D\sim P^L}\left|Q_{1-\alpha}(M_D^{\mathrm{f}}) - Q_{1-\alpha}(M_W^{\mathrm{f}})\right| \le \frac{C}{\sqrt{L}}, \tag{37a}$$

$$\mathbb{E}_{D\sim P^L}\left|Q_{1-\alpha}(M_D) - Q_{1-\alpha}(M_W)\right| \le \frac{C}{\sqrt{L}}. \tag{37b}$$

Finally, assume that $|\Gamma_t^*(X_t)|$ is uniformly bounded for $* \in \{\text{OCP, FOCP, AOCP, AFOCP}\}$.

Then there exists a non-negative correction $\Delta(T) = O(T^{-1/2})$ such that, for all $T \in \mathbb{N}$,

$$\frac{1}{T}\sum_{t=1}^{T}\left|\Gamma_t^{\text{AFOCP/FOCP}}(X_t)\right| \le \frac{1}{T}\sum_{t=1}^{T}\left|\Gamma_t^{\text{AOCP/OCP}}(X_t)\right| + \Delta(T), \tag{38}$$

with $\Delta(T) \to 0$ as $T \to \infty$.

Proof: The proof proceeds in two steps. We first derive a population-level inequality from the regularity conditions (35)–(37a). We then apply a concentration argument based on the mixing assumption to extend this population result to the time-averaged inequality (38).

Step 1. Following the algebraic argument of Theorem 6 in Teng et al. (2022) and its weighted extension Theorem 1 in Chen et al. (2024), we aim to establish the population-level inequality

$$\mathbb{E}_{(D,X_0)\sim\mathcal{P}}\left[\mathcal{H}(Q_{1-\alpha}(M_D^{\mathrm{f}}), X_0)\right] \le \mathbb{E}_{D\sim P^L}\left[Q_{1-\alpha}(M_D)\right]. \tag{39}$$

First, we rearrange the expansion condition (36) using the identity $\mathbb{M}[Q_{1-\alpha}(\cdot) - \cdot] = Q_{1-\alpha}(\cdot) - \mathbb{M}[\cdot]$, which gives

$$\begin{aligned}&\mathbb{E}_{D\sim P^L}\left[\mathbb{M}[\mathcal{H}(M_D^{\mathrm{f}}, X_D)]\right] + R\cdot\mathbb{E}_{D\sim P^L}\left[\mathbb{M}\left[|Q_{1-\alpha}(M_D^{\mathrm{f}}) - M_D^{\mathrm{f}}|^\beta\right]\right]\\ &\quad< \mathbb{E}_{D\sim P^L}\left[Q_{1-\alpha}(\mathcal{H}(M_D^{\mathrm{f}}, X_D))\right] - \epsilon - 2\max\{R,1\}\left(\frac{C}{\sqrt{L}}\right)^{\min\{\beta,1\}}.\end{aligned} \tag{40}$$

Next, we apply the Hölder condition (34) with $M = Q_{1-\alpha}(M_D^{\mathrm{f}})$ and $M' = M_D^{\mathrm{f}}$, and take the mean $\mathbb{M}$ and expectation $\mathbb{E}_{D\sim P^L}$ on both sides, which yields

$$\begin{aligned}\mathbb{E}_{D\sim P^L}\left[\mathbb{M}[\mathcal{H}(Q_{1-\alpha}(M_D^{\mathrm{f}}), X_D)]\right] &\le \mathbb{E}_{D\sim P^L}\left[\mathbb{M}[\mathcal{H}(M_D^{\mathrm{f}}, X_D)]\right]\\ &\quad + R\cdot\mathbb{E}_{D\sim P^L}\left[\mathbb{M}\left[|Q_{1-\alpha}(M_D^{\mathrm{f}}) - M_D^{\mathrm{f}}|^\beta\right]\right].\end{aligned} \tag{41}$$

Combining (40) and (41) eliminates the common terms and gives

$$\mathbb{E}_{D\sim P^L}\left[\mathbb{M}[\mathcal{H}(Q_{1-\alpha}(M_D^{\mathrm{f}}), X_D)]\right] < \mathbb{E}_{D\sim P^L}\left[Q_{1-\alpha}(\mathcal{H}(M_D^{\mathrm{f}}, X_D))\right] - \epsilon - 2\max\{R,1\}\left(\frac{C}{\sqrt{L}}\right)^{\min\{\beta,1\}}. \tag{42}$$

Applying the length preservation condition (35) to (42) cancels the ϵ term and gives

$$\mathbb{E}_{D\sim P^L}\left[\mathbb{M}[\mathcal{H}(Q_{1-\alpha}(M_D^{\mathrm{f}}), X_D)]\right] < \mathbb{E}_{D\sim P^L}\left[Q_{1-\alpha}(M_D)\right] - 2\max\{R,1\}\left(\frac{C}{\sqrt{L}}\right)^{\min\{\beta,1\}}. \tag{43}$$

Finally, by combining the quantile stability condition (37a), the Hölder condition (34), and stationarity of the data stream, we have

$$\mathbb{E}_{(D,X_0)\sim\mathcal{P}}\left[\mathcal{H}(Q_{1-\alpha}(M_D^{\mathrm{f}}), X_0)\right] \le \mathbb{E}_{D\sim P^L}\left[\mathbb{M}[\mathcal{H}(Q_{1-\alpha}(M_D^{\mathrm{f}}), X_D)]\right] + R\left(\frac{C}{\sqrt{L}}\right)^{\beta} + \frac{C}{\sqrt{L}}. \tag{44}$$

Substituting (43) into (44) yields

$$\begin{aligned}\mathbb{E}_{(D,X_0)\sim\mathcal{P}}\big[\mathcal{H}(Q_{1-\alpha}(M_D^{\mathrm{f}}),X_0)\big] &< \mathbb{E}_{D\sim P^L}\big[Q_{1-\alpha}(M_D)\big] - 2\max\{R,1\}\left(\frac{C}{\sqrt{L}}\right)^{\min\{\beta,1\}} + R\left(\frac{C}{\sqrt{L}}\right)^{\beta} + \frac{C}{\sqrt{L}} \\ &\leq \mathbb{E}_{D\sim P^L}\big[Q_{1-\alpha}(M_D)\big]. \end{aligned} \tag{45}$$

Step 2. We introduce the bounded functionals

$$f_1(W,X) = \mathcal{H}(Q_{1-\alpha}(M_W^{\mathrm{f}}),X), \tag{46a}$$

$$f_2(W) = Q_{1-\alpha}(M_W), \tag{46b}$$

so that the time-averaged interval lengths of AFOCP and AOCP can be written as

$$\frac{1}{T}\sum_{t=1}^{T}\left|\Gamma_t^{\mathrm{AFOCP}}(X_t)\right| = \frac{1}{T}\sum_{t=1}^{T} f_1(W_t,X_t), \tag{47a}$$

$$\frac{1}{T}\sum_{t=1}^{T}\left|\Gamma_t^{\mathrm{AOCP}}(X_t)\right| = \frac{1}{T}\sum_{t=1}^{T} f_2(W_t), \tag{47b}$$

where $W_t = \{(X_\tau,Y_\tau)\}_{\tau=t-L}^{t-1}$. The joint sequence $\{(W_t,X_t)\}_{t\geq 1}$ inherits the α-mixing property from the data stream, so by standard concentration inequalities for bounded functionals of α-mixing sequences (Bradley, 2005; Rio, 2017), there exists a constant $C'>0$ such that for $i=1,2$,

$$\left|\frac{1}{T}\sum_{t=1}^{T} f_i(W_t,X_t) - \mathbb{E}_{(D,X_0)\sim\mathcal{P}}[f_i(D,X_0)]\right| \leq \frac{C'}{\sqrt{T}}. \tag{48}$$

That is, the time average $\sum_{t=1}^{T} f_i(W_t,X_t)/T$ converges to the population expectation $\mathbb{E}_{(D,X_0)\sim\mathcal{P}}[f_i(D,X_0)]$ at the standard $O(T^{-1/2})$ rate.

Combining (47), (48), and the population-level inequality (39) yields

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left|\Gamma_t^{\mathrm{AFOCP}}(X_t)\right| - \frac{1}{T}\sum_{t=1}^{T}\left|\Gamma_t^{\mathrm{AOCP}}(X_t)\right| &\leq \mathbb{E}_{(D,X_0)\sim\mathcal{P}}[f_1(D,X_0)] - \mathbb{E}_{D\sim P^L}[f_2(D)] + \frac{2C'}{\sqrt{T}} \\ &\leq \frac{2C'}{\sqrt{T}}. \end{aligned} \tag{49}$$

Setting

$$\Delta(T) = \frac{2C'}{\sqrt{T}} = O(T^{-1/2}) \tag{50}$$

yields the time-averaged inequality (38), with $\Delta(T)\to 0$ as $T\to\infty$. Taking $\{w_t^\tau\}_{\tau=1}^{L+1}$ uniform reduces AOCP and AFOCP to OCP and FOCP, and the same argument applies.

B Effect of the Training Loss

Table 2 compares the two losses for each method, with $M=1$ above the line and $M=5$ below. At $M=1$ the two losses are close, and the squared error gives the shorter interval on most datasets. At $M=5$ the pinball loss is shorter on all datasets for AOCP and on three of five for AFOCP, at comparable or closer-to-target coverage. A plausible reason is that the pinball loss targets the deployed quantile directly, which the higher-capacity $M=5$ model can exploit, while the lower-capacity single-head model benefits from the smoother squared error. The differences are modest, and AFOCP at $M=5$ runs counter to this trend on synthetic and electricity. We use the squared error at $M=1$ and the pinball loss at $M=5$.

Table 2: Interval length under the squared error (23) and the pinball loss (24) for the attention-based methods, with $L = 100$, $D = 50$, $\alpha = 0.1$, over five seeds. For each method and dataset the shorter length is in bold. The default $M = 1$ methods and the $M = 5$ methods are separated by the middle line.

	Synthetic		Electricity		Air quality		Bike sharing		Wind	
Method	sq	pb	sq	pb	sq	pb	sq	pb	sq	pb
AOCP	5.969	5.968	**1.651**	1.659	3.318	**3.140**	**2.729**	2.775	**3.910**	4.380
AFOCP	2.130	2.130	0.169	0.169	0.443	**0.424**	**0.283**	0.288	**0.418**	0.464
AOCP (M=5)	7.742	**7.342**	1.989	**1.971**	2.919	**2.820**	2.813	**2.678**	4.395	**4.347**
AFOCP (M=5)	**2.008**	2.125	**0.153**	0.154	0.554	**0.541**	0.330	**0.314**	0.406	**0.400**

Table 3: Time-averaged coverage and prediction interval length on the five datasets, with $L = 100$, $D = 50$, $\alpha = 0.1$. Each entry reports mean ± standard deviation over five random seeds, evaluated at the final time step T. The smallest interval length on each dataset is shown in bold.

Dataset	Method	Coverage (%)	Interval length
Synthetic	OCP	90.67 ± 0.80	8.28 ± 1.04
	AOCP	90.54 ± 0.74	5.97 ± 0.76
	FOCP	89.69 ± 0.63	2.75 ± 0.72
	AFOCP	89.79 ± 0.45	$\mathbf{2.13} \pm 0.64$
Air quality	OCP	90.54 ± 0.28	4.20 ± 0.39
	AOCP	89.93 ± 0.24	3.32 ± 0.17
	FOCP	90.67 ± 0.28	0.55 ± 0.09
	AFOCP	89.87 ± 0.28	$\mathbf{0.44} \pm 0.10$
Electricity	OCP	91.34 ± 0.60	2.02 ± 0.30
	AOCP	90.67 ± 0.60	1.65 ± 0.25
	FOCP	91.07 ± 0.74	0.21 ± 0.05
	AFOCP	90.54 ± 0.65	$\mathbf{0.17} \pm 0.04$
Bike-sharing	OCP	98.15 ± 0.93	3.02 ± 0.24
	AOCP	91.30 ± 1.40	2.73 ± 0.25
	FOCP	96.85 ± 1.06	0.31 ± 0.04
	AFOCP	89.63 ± 1.78	$\mathbf{0.28} \pm 0.03$
Wind	OCP	93.10 ± 0.40	4.81 ± 0.15
	AOCP	90.17 ± 0.40	3.91 ± 0.11
	FOCP	92.21 ± 2.02	0.50 ± 0.09
	AFOCP	90.73 ± 1.48	$\mathbf{0.42} \pm 0.08$

C Per-Seed Variability of Long-term Metrics

This section supplements Figure 2 by reporting the final time-averaged coverage and time-averaged prediction interval length, together with their per-seed standard deviations over five random seeds.

Table 3 shows that all four methods attain the target coverage level $1 - \alpha = 90\%$ on every dataset, with standard deviations below 2 percentage points on most datasets. AFOCP yields the smallest prediction interval length on every dataset, with feature-space calibration providing the dominant reduction relative to output-space methods and attention-based weighting yielding a further refinement. Moreover, the mean ± standard deviation ranges of FOCP and AFOCP do not overlap with those of OCP and AOCP on any dataset, demonstrating that the efficiency advantage of feature-space calibration is statistically robust under random initialization and consistent across the heterogeneous benchmarks considered.

D Local Behavior under Rolling Windows

Long-term metrics can conceal local deviations from the target coverage, particularly around regime changes. Let

$$c_t = \frac{1}{W} \sum_{\tau=t-W+1}^{t} \mathbb{1}\{Y_\tau \in \Gamma_\tau(X_\tau)\}$$

denote the rolling coverage at time t, with window size $W = 50$. This section visualizes c_t and the corresponding rolling interval length on the five datasets in Figure 5.

Figure 5 (left) shows the rolling coverage of the four methods on the five datasets. All four methods stay close to the target $1 - \alpha = 0.9$ throughout the trajectory. The attention-based methods AOCP and AFOCP fluctuate less around the target than their uniform-weight counterparts, reflecting the ability of attention-based weighting to adaptively reweight historical observations under local distribution changes. The effect is most clearly visible on the synthetic, bike-sharing, and wind benchmarks. On the air-quality and electricity benchmarks, the four methods are closer together, although AOCP and AFOCP still appear to track the target slightly more closely.

Figure 5 (right) shows the rolling interval length on the same datasets. On every dataset, OCP yields the largest rolling interval length, followed by AOCP, then FOCP, with AFOCP yielding the smallest. A substantial gap appears between AOCP and FOCP, reflecting the transition from output-space to feature-space calibration. The rolling length fluctuates more than the cumulative time-averaged length of Figure 2, as expected, but this ordering is consistent across the entire trajectory. AFOCP therefore attains the smallest interval length throughout, demonstrating that its efficiency advantage is sustained at the local time scale.

E Coverage Behavior around Regime Changes

This section examines the local coverage behavior around regime changes, complementing the trajectory-wide view of Section D. We focus on the synthetic benchmark, whose regime changes are explicitly designed and more pronounced than those in the real-world datasets.

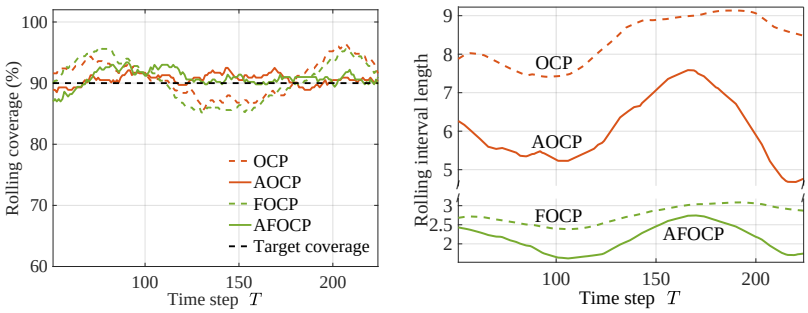
For each regime change in the synthetic data, we extract the per-step coverage indicator $\mathbb{1}\{Y_t \in \Gamma_t(X_t)\}$ within a window of ± 15 time steps centered at the change point. We then average these aligned windows across all regime changes and the five random seeds, and smooth the resulting curve by a moving average of size five.

At the change point (offset 0), all four methods exhibit a coverage drop below the target. The uniform-weight baselines OCP and FOCP drop more deeply, to approximately 70% and 76%. The attention-based methods AOCP and AFOCP drop more shallowly, to approximately 84% and 82%, and remain closer to the target throughout the recovery period. This complements the rolling-window observation in Section D, confirming that attention-based weighting reduces the transient coverage drop following a regime change.

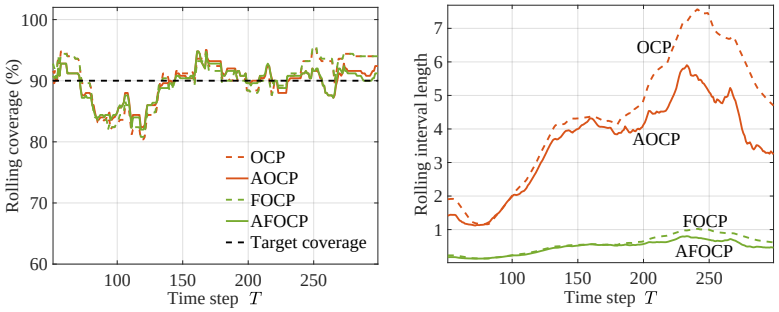
F Comparison with Offline Feature-Space CP

To isolate the contribution of online recalibration, we compare AFOCP against the feature-space attention-based offline CP method of Chen et al. (2024), which shares the feature-space NC score construction with AFOCP but pre-trains the attention module on a fixed calibration split and does not update α_t online. The two methods therefore differ primarily in whether $\{w_t^\tau\}_{\tau=1}^{L+1}$ and α_t are updated online. Table 4 reports empirical coverage and average interval length on the five datasets from Section 4, averaged over five seeds at target $\alpha = 0.1$.

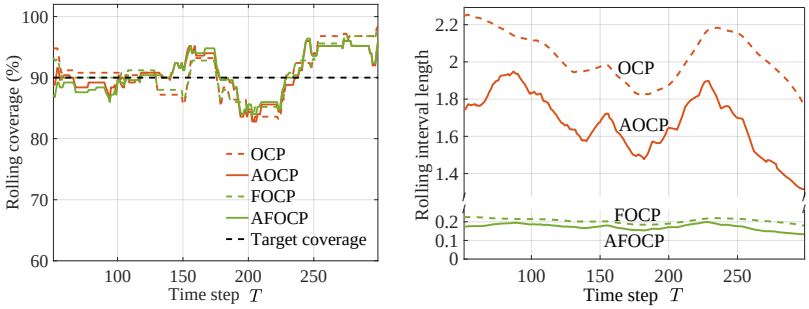
AFOCP holds coverage within $\pm 1\%$ of the target on all five datasets, while the offline baseline deviates by up to 3.15%. The interval-length comparison falls into three categories. First, on synthetic and bike sharing the baseline produces shorter intervals than AFOCP but under-covers by 2.41% to 2.83%, reflecting the standard coverage-length trade-off in CP. Second, on electricity AFOCP achieves both better-targeted coverage and shorter intervals than the baseline (0.169 vs 0.208), and on wind the two methods produce nearly identical



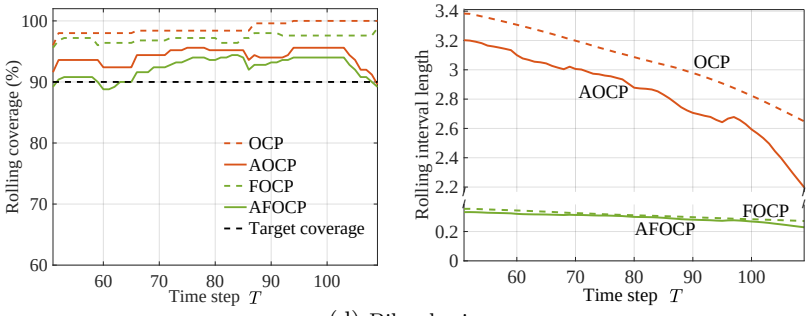
(a) Synthetic



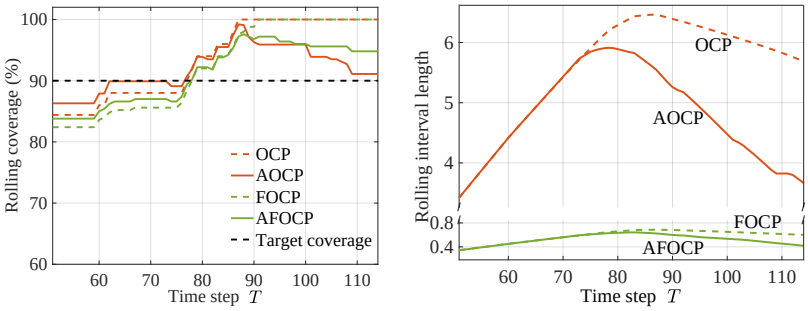
(b) Air quality



(c) Electricity



(d) Bike-sharing



(e) Wind

Figure 5: Rolling-window coverage (left) and rolling-window prediction interval length (right) versus time step T across various datasets, with rolling window $W = 50$, calibration window length $L = 100$, feature dimension $D = 50$, and target miscoverage rate $\alpha = 0.1$.

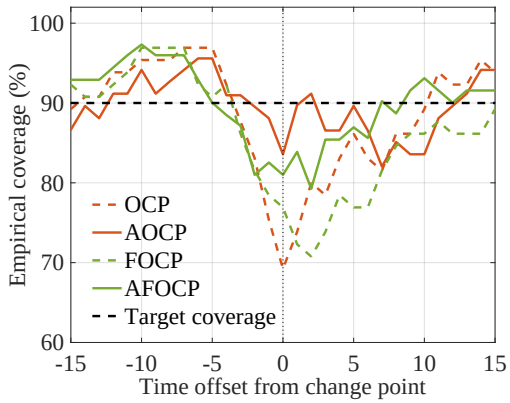


Figure 6: Coverage behavior of OCP, AOCP, FOCP, and AFOCP around regime changes on the synthetic benchmark, with calibration window length $L = 100$, feature dimension $D = 50$, and target miscoverage rate $\alpha = 0.1$. The horizontal axis is the time offset relative to a regime change at offset 0, indicated by the vertical dotted line. The dashed black line marks the target coverage $1 - \alpha = 0.9$. Curves report the empirical coverage averaged over all regime-change windows and five random seeds, smoothed by a moving average of size five.

Table 4: Empirical coverage (%) and average interval length at target coverage 90%, averaged over five seeds. The offline baseline of Chen et al. (2024) deviates from the target by up to 3.15%; AFOCP holds coverage within ±1% on all five datasets.

	Offline (Chen et al., 2024)		AFOCP (this work)	
Dataset	Coverage	Length	Coverage	Length
Synthetic	87.17	1.626	89.79	2.130
Electricity	93.15	0.208	90.54	0.169
Air quality	92.28	0.319	89.87	0.443
Bike sharing	87.59	0.250	89.63	0.283
Wind	89.03	0.424	90.73	0.418

lengths (0.418 vs 0.424). Third, on air quality the baseline produces shorter intervals than AFOCP despite over-covering by 2.28%; this is the only dataset where AFOCP's improved coverage tracking does not also translate to a length advantage, reflecting the trade-off between online adaptivity and length efficiency on datasets where a fixed calibration is already well-aligned with the test distribution. Since the two methods share the feature-space NC score, the differences in coverage tracking are attributable to the online update of $\{w_t^\tau\}_{\tau=1}^{L+1}$ and α_t.