

Article

Structure-Guided Dual-Timescale Learning for Heterogeneous Short-Horizon Time-Series Forecasting

Tian Peng ¹, Junrong Ye ¹, Min Qi ¹, Yuqiang Bai ¹ and Honghao Wu ^{2,*}¹ Department of Academic Research, Beijing Police College, Beijing 102202, China; pt_2012@126.com (T.P.)² Department of Intelligent Policing Technology, Beijing Police College, Beijing 102202, China

* Correspondence: wuhonghao@bjpc.edu.cn

Abstract

Many real-world forecasting tasks involve heterogeneous temporal sources, where high-frequency dynamic observations coexist with low-frequency semantic or contextual information. Existing methods often treat these inputs through simple broadcasting or direct concatenation, which weakens cross-scale dependency modeling and makes the prediction process vulnerable to unstable contextual correlations. To address this issue, this paper proposes a structure-guided dual-timescale learning framework for heterogeneous short-horizon time-series forecasting. A fast-scale temporal encoder is used to extract fine-grained dynamic patterns, while a slow-scale semantic encoder is introduced to characterize slowly varying contextual information. On this basis, a structure-guided semantic refinement module is designed to suppress unstable semantic components before fusion, and a structure-aware cross-scale attention mechanism is developed to adaptively align fast dynamics with the most relevant slow-varying context. In addition, a coarse-to-fine prediction structure is employed to separate background tendency modeling from short-term fluctuation correction, and intervention-based consistency regularization is incorporated to improve robustness under changing conditions. Feature contribution analysis further confirms the importance of selecting effective contextual variables to form a compact slow-timescale representation. Experiments on a real-world ultra-short-term aggregate load forecasting task demonstrate that the proposed framework achieves superior accuracy and robustness, indicating its potential as a general solution for forecasting problems with heterogeneous temporal resolutions.

Keywords: heterogeneous temporal forecasting; dual-timescale learning; structure-guided forecasting; cross-scale attention; coarse-to-fine prediction; intervention-based regularization



Academic Editors: Shuangqi Li,
Alexis Pengfei Zhao, Yichen Shen and
Sergio Spanò

Received: 21 April 2026

Revised: 23 June 2026

Accepted: 26 June 2026

Published: 1 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Short-horizon time-series forecasting plays a critical role in many practical systems, including power system operation, industrial process management, transportation scheduling, and demand-side coordination. In these applications, an effective forecasting model should capture short-term temporal dependence from dense historical observations while also exploiting slowly varying contextual factors, such as weather conditions, calendar effects, operating modes, and semantic descriptors. Among these tasks, short-term load forecasting has been widely studied because accurate prediction directly supports dispatch, reserve allocation, demand response, and secure system operation [1–8]. Recent reviews further show that exogenous and contextual variables are often essential for improving forecasting accuracy, although their effective use remains challenging [3,8–10].

Early forecasting studies mainly relied on statistical and econometric models, which were interpretable and effective when temporal patterns were relatively stable [2,11]. As forecasting tasks became more nonlinear and data-rich, deep learning methods gradually became dominant. Recurrent neural networks, especially LSTM and GRU, were widely adopted for sequential dependency modeling [12]. Later, convolution-based sequence models such as TCNs and dilated autoregressive architectures improved receptive-field coverage and computational efficiency [13,14]. In parallel, probabilistic forecasting methods, including DeepAR, MQ-RNN, and deep state space models, extended sequence learning to uncertainty-aware multi-step prediction [15–17].

More recently, global forecasting architectures have further advanced the field. N-BEATS and N-HiTS showed that residual and hierarchical designs can deliver strong forecasting accuracy with good scalability [18,19]. Transformer-based methods, including TFT, Informer, Autoformer, and FEDformer, further strengthened long-range dependency modeling and multi-source feature fusion [20–23]. At the same time, studies such as DLinear, TimesNet, and TiDE suggest that forecasting performance depends not only on model complexity, but also on whether the architecture matches the temporal structure of the task [24–26]. Despite this progress, three issues remain insufficiently addressed: heterogeneous temporal resolutions, effective contextual fusion, and robustness under distribution shifts.

A central difficulty is that many real-world forecasting tasks involve heterogeneous temporal resolutions. The target series is often observed at a fast timescale, whereas contextual or exogenous information is only available at a slower rate. Existing methods typically address this mismatch by direct broadcasting, naive concatenation, or undifferentiated multivariate encoding. Although simple, these strategies tend to blur the distinct roles of fast-varying dynamics and slow-varying contextual drivers, weaken cross-scale dependency modeling, and increase sensitivity to unstable contextual correlations [9,20,27]. Recent studies on exogenous-variable forecasting show that external information is beneficial only when its temporal role and structural heterogeneity are explicitly modeled [20,27]. In addition, formal causal inference studies indicate that purely correlation-driven learning can become fragile under changing environments and distribution shifts [28].

Recent studies further suggest that forecasting performance depends not only on the use of advanced neural architectures but also on whether the model structure is consistent with the temporal and contextual characteristics of the task. PatchTST shows that representing time series as local temporal patches can improve the utilization of historical information in Transformer-based forecasting [29]. Non-stationary Transformers further indicate that distributional changes in real-world time series can weaken conventional attention-based forecasting if non-stationarity is not explicitly considered [30]. In renewable energy forecasting, Chen et al. proposed a wind power ultra-short-term ensemble forecasting framework considering wind speed correction and scenario classification, demonstrating that exogenous correction and regime-aware modeling are useful for improving short-horizon forecasting under variable operating conditions [31]. These studies support the view that short-horizon forecasting should not only model temporal dependencies but also consider exogenous effects, operating regimes, and distributional variability. However, the structural alignment between dense fast-timescale target dynamics and slowly varying contextual semantics remains insufficiently explored, especially when contextual correlations are unstable under changing operating conditions.

Motivated by these observations, this paper studies short-horizon forecasting from a general algorithmic perspective rather than as a task-specific solution. We propose a structure-guided dual-timescale learning framework for heterogeneous short-horizon time-series forecasting. Specifically, a fast-scale temporal encoder is used to capture fine-grained

target dynamics, while a slow-scale semantic encoder models low-frequency contextual information. A structure-guided semantic refinement module is introduced to suppress unstable contextual components before fusion, and a structure-aware cross-scale attention mechanism is designed to align fast dynamic states with the most relevant slow-varying context. In addition, a coarse-to-fine prediction structure is adopted to separate background tendency modeling from local fluctuation correction, and intervention-based consistency regularization is incorporated to improve robustness under changing operating conditions.

As shown in Figure 1, although the proposed framework is formulated for general heterogeneous short-horizon forecasting tasks, this paper uses industrial load forecasting as a representative validation scenario. This application naturally contains dense target observations and slowly varying contextual information, making it well-suited for evaluating multiscale forecasting models. The main contributions of this paper are summarized as follows:

- (1) A structure-guided dual-timescale forecasting framework is proposed for short-horizon time-series prediction with heterogeneous temporal resolutions.
- (2) A structure-guided semantic refinement and cross-scale attention mechanism is developed to improve slow-context utilization and reduce irrelevant contextual interference.
- (3) A coarse-to-fine prediction structure is introduced to separate background tendency modeling from local fluctuation correction.
- (4) Counterfactual consistency regularization is incorporated to improve forecasting robustness under changing operating conditions.

The remainder of this paper is organized as follows. Section 2 details the proposed methodology. Section 3 presents the experimental validation on a real-world short-horizon load forecasting task, including overall comparison, ablation analysis, and robustness evaluation. Section 4 concludes the paper.

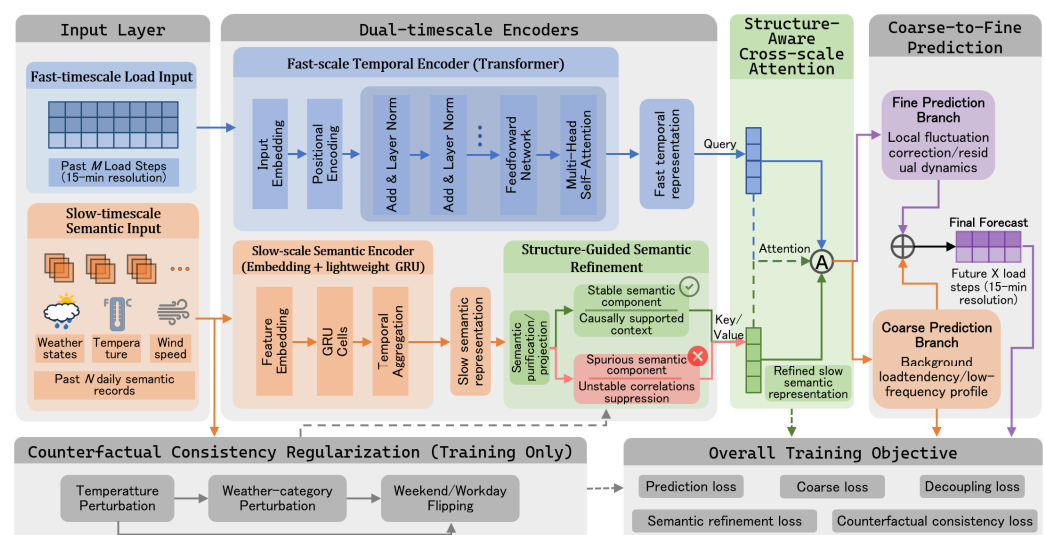


Figure 1. Structure-guided dual-timescale forecasting framework.

2. Methodology

2.1. Problem Formulation

In the considered forecasting setting, the target sequence is observed at a fast timescale, whereas auxiliary contextual variables are available at a slower timescale. This temporal mismatch makes the task fundamentally different from conventional single-timescale forecasting. Accordingly, the model should not only capture rapid dynamic evolution from recent observations, but also incorporate slowly varying context in a structurally appropri-

ate manner. To keep the formulation general, the following definitions are presented for heterogeneous short-horizon forecasting tasks; the load forecasting case study introduced later serves only as one representative application.

Let $x_t \in \mathbb{R}$ denote the target observation at fast-time step t . For a historical window of length L , the fast-timescale input sequence is defined as

$$X_t^f = \{x_{t-L+1}, x_{t-L+2}, \dots, x_t\} \quad (1)$$

Let $z_d \in \mathbb{R}^m$ denote the slow-timescale contextual vector at slow-time index d , including semantic, meteorological, calendar-related, or other exogenous operating descriptors. For a historical window of K slow-time steps, the slow-timescale contextual sequence is defined as

$$Z_d^s = \{z_{d-K+1}, z_{d-K+2}, \dots, z_d\} \quad (2)$$

To connect the two timescales, a mapping $\pi(t)$ is introduced, where $\pi(t) = d$ indicates that fast-time step t corresponds to slow-time index d . The forecasting objective is to learn a mapping

$$\hat{Y}_{t+1:t+H} = f_{\Theta}(X_t^f, Z_{\pi(t)}^s) \quad (3)$$

where H is the prediction horizon, and Θ denotes all trainable parameters.

Different from ordinary correlation-driven forecasting schemes, the learned mapping should satisfy three requirements simultaneously: (1) it should extract short-term and longer-range temporal dependencies from the fast-timescale history; (2) it should encode slowly varying context without resorting to crude temporal duplication; and (3) it should exploit contextual variables in a manner that is resistant to spurious correlations and robust under changing operating conditions.

2.2. Dual-Timescale Representation Learning

This subsection presents the dual-timescale representation learning strategy that forms the basis of the proposed framework. The key idea is to encode fast-varying observations and slow-varying contextual information through two dedicated branches, so that each source can be modeled according to its own temporal characteristics before cross-scale interaction is introduced.

2.2.1. Fast-Scale Temporal Encoder

The fast-timescale sequence contains the dominant information for short-horizon forecasting because it directly reflects recent operating states, local fluctuations, ramps, and short-term persistence. To simultaneously capture local temporal variation and long-range dependency, a Transformer-based temporal encoder is adopted for the fast branch.

The fast-scale representation is written as

$$H_t^f = \text{Enc}_f(X_t^f) \in \mathbb{R}^{L \times d_f} \quad (4)$$

where d_f is the latent dimension of the fast branch. Let

$$H_t^f = [h_{t-L+1}^f, h_{t-L+2}^f, \dots, h_t^f] \quad (5)$$

where each $h_t^f \in \mathbb{R}^{d_f}$ is the hidden representation of one 15 min step.

Within the Transformer encoder, self-attention is defined as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

where Q , K , and V denote the query, key, and value matrices, respectively, and d_k is the key dimension. Through this mechanism, the fast branch captures fine-grained temporal dependence among historical observations and provides the principal dynamic representation for downstream forecasting.

2.2.2. Slow-Scale Temporal Encoder

Unlike the fast-timescale sequence, slow-timescale contextual variables evolve gradually and mainly describe the background operating regime rather than instantaneous fluctuations. Therefore, the slow branch is designed to encode low-frequency semantic information in a compact yet informative form, providing contextual support for downstream forecasting rather than competing with the fast branch in modeling short-term dynamics.

The slow-scale semantic encoder is defined as

$$H_d^s = \text{Enc}_s(Z_d^s) \in \mathbb{R}^{K \times d_s} \quad (7)$$

where d_s is the latent semantic dimension. Denote

$$H_d^s = [h_{d-K+1}^s, h_{d-K+2}^s, \dots, h_d^s] \quad (8)$$

where $h_l^s \in \mathbb{R}^{d_s}$ is the encoded representation of one day-level semantic input.

For each day-level semantic vector, a latent embedding is first obtained by

$$s_d = \sigma(W_s z_d + b_s) \quad (9)$$

where W_s and b_s are learnable parameters and $\sigma(\cdot)$ is a nonlinear activation function.

To preserve the temporal persistence of slow-varying context, the historical semantic sequence can be further aggregated as

$$\bar{s}_d = \sum_{\tau=d-K+1}^d \beta_\tau s_\tau \quad (10)$$

where β_τ is a learnable normalized importance coefficient satisfying $\sum_{\tau=d-K+1}^d \beta_\tau = 1$. Compared with directly copying day-level variables into all intraday slots, this design explicitly treats semantic information as a slow contextual process and is therefore more consistent with its temporal meaning.

2.3. Structure-Guided Semantic Refinement

Although daily semantic variables are informative, they may also contain components that appear predictive only under specific historical environments rather than under stable operating mechanisms. Directly injecting such information into the forecasting model may therefore introduce contextual bias. To alleviate this issue, a causality-guided refinement step is introduced before cross-scale fusion.

Conceptually, the learned slow semantic representation can be decomposed as

$$h_d^s = h_d^{s,*} + r_d \quad (11)$$

where $h_d^{s,*}$ denotes the mechanism-consistent component and r_d denotes the undesired component associated with spurious or weakly justified correlations.

A domain-informed dependency graph over semantic factors is defined based on prior domain knowledge, and its adjacency matrix is denoted by A_c . A projection operator is constructed as

$$P_c = A_c^\top (A_c A_c^\top + \delta I)^{-1} A_c \quad (12)$$

where $\delta > 0$ is a small regularization coefficient for numerical stability. The refined slow semantic representation is then given by

$$H_d^{s,c} = P_c H_d^s \quad (13)$$

This operation suppresses semantic components that are weakly aligned with the assumed dependency structure while preserving those that are more consistent with domain-supported dependencies. In this way, the slow branch is refined before it interacts with the fast branch, which helps prevent irrelevant context from dominating the fusion process.

$$\mathcal{L}_{\text{causal}} = \sum_d \|H_d^s - H_d^{s,c}\|_2^2 \quad (14)$$

In essence, this module acts as a structured denoising mechanism for slow-timescale context and reduces the vulnerability of semantic fusion to dataset-specific bias.

2.4. Structure-Aware Cross-Scale Attention Fusion

After obtaining the fast temporal representation H_t^f and the refined slow semantic representation $H_d^{s,c}$, the next step is to couple the two branches. Rather than using direct concatenation, the proposed framework introduces a cross-scale attention mechanism so that the fast branch can selectively retrieve the slow-varying contextual information that is most relevant to the current prediction state.

Specifically, the fast branch acts as the query, and the refined slow branch acts as the key-value memory:

$$Q_t = H_t^f W_Q, K_d = H_d^{s,c} W_K, V_d = H_d^{s,c} W_V \quad (15)$$

where W_Q , W_K , and W_V are trainable projection matrices.

The cross-scale attention matrix is then computed as

$$A_{t,d} = \text{Softmax} \left(\frac{Q_t K_d^\top}{\sqrt{d_k}} + B_{\pi(t),d} \right) \quad (16)$$

where $B_{\pi(t),d}$ is a time-mapping bias that encodes the alignment relationship between the current prediction context at the fast timescale and the candidate contextual representations at the slow timescale.

The attended slow context is obtained by

$$C_t = A_{t,d} V_d \quad (17)$$

Finally, the cross-scale fusion representation is defined as

$$\tilde{H}_t = \text{FFN}([H_t^f \| C_t]) \quad (18)$$

where $[\cdot \| \cdot]$ denotes concatenation and $\text{FFN}(\cdot)$ is a feed-forward projection layer.

The physical meaning is that the fast branch determines which contextual information is currently needed, while the slow branch provides semantically refined and structurally filtered background memory. This design is more suitable than naive broadcast-and-

concatenate alignment because slow-timescale variables should guide the prediction as background context instead of being treated as pseudo high-frequency observations.

2.5. Coarse-to-Fine Forecast Generation

A key challenge in heterogeneous-timescale forecasting is that slow contextual variables and fast observations affect the target at different functional levels. Slow-timescale context mainly determines the background operating regime, whereas recent fast-timescale observations govern local fluctuations, turning points, and residual corrections. To explicitly reflect this difference, a coarse-to-fine prediction head is designed.

2.5.1. Coarse Prediction Branch

The coarse branch predicts a background trajectory from the refined slow context and its cross-scale interaction summary:

$$\hat{Y}_{t+1:t+H}^{\text{coarse}} = f_{\text{coarse}}(G_t) \quad (19)$$

where G_t is a global context representation obtained from $H_d^{s,c}$ and C_t , for example, through average pooling or attention pooling:

$$G_t = \text{Pool}(H_d^{s,c}, C_t) \quad (20)$$

This branch is intended to capture the coarse shape of the forthcoming target profile, including overall level, background rise/fall tendency, and dominant operating regime.

2.5.2. Fine Prediction Branch

The fine branch predicts short-term residual adjustments from the fused fast-scale representation:

$$\hat{Y}_{t+1:t+H}^{\text{fine}} = f_{\text{fine}}(\tilde{H}_t) \quad (21)$$

This branch focuses on local details such as short-term ramps, oscillations, and abrupt fluctuations that are primarily governed by recent fast-timescale evolution.

2.5.3. Final Output

The final forecast is obtained by additive reconstruction:

$$\hat{Y}_{t+1:t+H} = \hat{Y}_{t+1:t+H}^{\text{coarse}} + \hat{Y}_{t+1:t+H}^{\text{fine}} \quad (22)$$

This design avoids forcing a single output head to model both regime-level background variation and high-frequency detail simultaneously, thereby improving both interpretability and modeling efficiency.

2.6. Intervention-Based Consistency Regularization

Contextual relationships observed in the training data may not remain valid under holidays, extreme weather, emergency operations, or atypical demand patterns. To improve robustness under such changing conditions, an intervention-based regularization strategy is introduced in the slow-timescale semantic branch.

Given a factual slow semantic input Z_d^s , an intervention-based semantic input $(Z_d^s)^{cf}$ is constructed by intervening on selected variables while keeping other conditions unchanged. Typical interventions include replacing a workday label with a holiday label, perturbing weather-category variables within a physically meaningful range, or modifying event indicators according to causal assumptions.

The intervention-based prediction is written as

$$\hat{Y}_{t+1:t+H}^{cf} = f_{\Theta}(X_t^f, (Z_d^s)^{cf}) \quad (23)$$

Rather than forcing the intervention-based prediction to be identical to the factual prediction, the objective is to penalize unreasonable response patterns under controlled interventions. In other words, the model is encouraged to react to contextual changes in a manner that is directionally plausible and structurally consistent with prior causal assumptions.

$$\mathcal{L}_{cf} = \left\| \hat{Y}_{t+1:t+H}^{cf} - \hat{Y}_{t+1:t+H}^{ref} \right\|_2^2 \quad (24)$$

where $\hat{Y}_{t+1:t+H}^{ref}$ is a reference response constructed according to the intervention rule.

To further encourage behavior-level consistency, a directional constraint can also be imposed for specific interventions:

$$\mathcal{L}_{dir} = \sum_{h=1}^H l\left(\text{sign}\left(\hat{y}_{t+h}^{cf} - \hat{y}_{t+h}\right), \text{sign}(\Delta_{t+h}^{causal})\right) \quad (25)$$

where Δ_{t+h}^{causal} denotes the expected response direction and $l(\cdot, \cdot)$ is a mismatch penalty. This term is optional but useful when domain prior knowledge is available.

2.7. Overall Objective Function

The total training loss combines predictive accuracy, structure-guided semantic refinement, and intervention robustness:

$$\mathcal{L} = \mathcal{L}_{pred} + \lambda_1 \mathcal{L}_{causal} + \lambda_2 \mathcal{L}_{cf} + \lambda_3 \mathcal{L}_{dir} + \lambda_4 \mathcal{L}_{coarse} \quad (26)$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are balancing coefficients.

Here, the primary forecasting loss is defined as

$$\mathcal{L}_{pred} = \frac{1}{H} \sum_{h=1}^H \left\| \hat{y}_{t+h} - y_{t+h} \right\|_2^2 \quad (27)$$

and the auxiliary supervision term for the coarse branch is defined as

$$\mathcal{L}_{coarse} = \frac{1}{H} \sum_{h=1}^H \left\| \hat{y}_{t+h}^{coarse} - y_{t+h}^{coarse} \right\|_2^2 \quad (28)$$

If no explicit coarse target is constructed, $\mathcal{L}_{coarse}$ can be omitted, and the coarse branch can be optimized indirectly through the overall prediction loss.

2.8. Training Procedure

The overall training procedure of the proposed model can be summarized as follows.

- Construct a fast-timescale observation window X_t^f and a slow-timescale contextual window $Z_{\pi(t)}^s$.
- Encode the fast-timescale sequence through the fast-scale encoder to obtain H_t^f .
- Encode the slow-timescale contextual sequence through the slow-scale encoder to obtain H_d^s .
- Refine the slow semantic representation through the structure-guided projection operator to obtain $H_d^{s,c}$.

- Apply structure-aware cross-scale attention, where H_t^f serves as query and $H_d^{s,c}$ serves as key-value memory, to obtain the contextual representation C_t .
- Fuse H_t^f and C_t to obtain $\tilde{H}_t$.
- Generate the coarse forecast and fine forecast, and combine them to produce the final prediction.
- Construct intervention-based semantic inputs, repeat the forward pass, and compute the intervention-based consistency loss.
- Optimize the total objective in Equation (26) through backpropagation.

Compared with direct daily-feature broadcasting or simple feature concatenation, the proposed framework offers four main advantages. First, it explicitly respects the heterogeneous temporal resolutions of the two data sources rather than forcing them into an artificial single-timescale representation. Second, the structure-guided semantic refinement step reduces the risk that the slow branch amplifies unstable correlations. Third, cross-scale attention enables high-frequency dynamics to adaptively retrieve the most relevant slow-varying context. Fourth, the coarse-to-fine output structure separates regime-level background modeling from local fluctuation correction, improving both interpretability and predictive efficiency.

3. Experimental Validation on a Real-World Load Forecasting Task

3.1. Dataset Description

To validate the effectiveness of the proposed forecasting framework, experiments are conducted on a large-scale industrial load dataset from the power system domain. This dataset mainly contains electricity consumption records of large industrial users and is suitable for load forecasting analysis. It provides detailed load observations from different industrial sectors and thus offers a valuable benchmark for studying short-horizon forecasting methods under realistic industrial electricity demand patterns. The dataset is publicly available at <https://gitcode.com/Universal-Tool/c8337> (accessed on 15 May 2026) (repository name: Universal-Tool/c8337).

The dataset covers three main aspects. First, it includes multiple types of industrial loads, corresponding to different industrial categories. Second, it provides time-series records of electricity consumption, which describe the temporal variation in industrial power demand over different time periods. Third, it contains load measurements at multiple temporal granularities, including hourly, daily, and monthly electricity consumption data. Such a data organization makes the dataset well-suited for evaluating forecasting models from both local dynamic and broader temporal trend perspectives.

In this paper, the industrial load dataset is used as the experimental benchmark to assess the predictive accuracy and generalization capability of the proposed method. Since the dataset contains representative temporal demand variations across industrial sectors, it provides an appropriate real-world testbed for validating the effectiveness of short-horizon time-series forecasting algorithms in power system applications.

3.2. Experimental Settings

To verify the effectiveness of the proposed framework on a representative real-world task, experiments are conducted on a feeder-level aggregate load forecasting dataset spanning from 1 January 2018 to 31 August 2021. The load measurements are recorded at a 15-min resolution, yielding 128,156 target observations in total. In parallel, daily meteorological and semantic information are collected over the same period, resulting in 1339 day-level context records. Therefore, this validation task naturally involves heterogeneous temporal resolutions, where the load sequence serves as the fast-timescale target signal and the meteorological-semantic variables serve as slow-timescale contextual inputs.

In the main setting, the fast-timescale input consists of the most recent 96 load steps. Since the load measurements are recorded at a 15 min resolution, this input window corresponds to the past 24 h. The slow-timescale input consists of the most recent seven daily meteorological-semantic records. The model is trained to predict the next 16 load steps. Under the 15 min sampling resolution, the 16-step prediction horizon corresponds to 4 h rather than 1 h. Therefore, the forecasting task in this paper is a 16-step, 4 h ahead short-horizon forecasting problem. The dataset is divided into training, validation, and test subsets in chronological order with a ratio of 70%/15%/15%, so as to avoid information leakage and better reflect practical forecasting scenarios.

The meteorological-semantic variables include weather status, temperature information, and wind-related descriptions. Specifically, the maximum temperature, minimum temperature, average temperature, and calendar-related indicators are treated as numerical slow-timescale features, while weather conditions and wind descriptions are encoded as categorical semantic features. For the load sequence, in addition to the historical aggregate load values, temporal indicators including hour-of-day, day-of-week, month-of-year, and weekend flag are also incorporated through sinusoidal and cosine transformations to improve periodicity modeling.

Four evaluation metrics, namely mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2), are adopted to evaluate forecasting performance from complementary perspectives. MAE measures absolute prediction accuracy, RMSE emphasizes large forecasting errors, MAPE reflects relative percentage error, and R^2 evaluates the goodness of fit between the predicted and actual load profiles:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (29)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (30)$$

$$\text{MAPE} = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (31)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (\bar{y} - y_i)^2}, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i \quad (32)$$

Here, y_i and $\hat{y}_i$ denote the ground-truth and predicted load values, respectively; $\bar{y}$ denotes the mean value of the ground-truth load values, and N is the number of forecasted points.

The proposed model is implemented in PyTorch (version 2.5.1), and the hyperparameter settings are summarized in Table 1. Unless otherwise specified, the baseline and the proposed model are trained and evaluated under the same protocol to ensure a fair comparison.

Table 1. Core hyperparameter settings of the proposed model.

Hyperparameters	Value
Load history length	96
Weather-semantic history length	7
Prediction horizon	16 steps (4 h)
Batch size	64
Number of epochs	20

Table 1. *Cont.*

Hyperparameters	Value
Initial learning rate	1×10^{-3}
Fast encoder hidden dimension	64
Slow encoder hidden dimension	64
Fusion dimension	64
Number of attention heads	4
Dropout rate	0.1
Early stopping patience	8

3.3. Feature Contribution Analysis of Meteorological-Semantic Variables

Before constructing the final slow-timescale input set, a leave-one-out feature contribution analysis is conducted to identify the meteorological-semantic variables that provide effective contextual information. Specifically, each slow-timescale feature is removed from the semantic branch one at a time, and the corresponding change in the test MAPE is recorded as ΔMAPE . A positive ΔMAPE means that removing the feature degrades forecasting performance and therefore indicates a beneficial contribution. In contrast, a negative ΔMAPE suggests that the feature introduces redundant or even detrimental information under the current forecasting setting.

The complete ranking results are reported in Table 2. Daytime Wind Direction (DWD), Nighttime Wind Direction (NWD), and Nighttime Wind Level (NWL) provide the strongest positive contributions, followed by Maximum Temperature (MaxT) and Weekend Indicator (WI). In contrast, Weather during the Day (WD), Weather during the Night (WN), Average Temperature (AT), and Minimum Temperature (MinT) exhibit negative contributions, indicating that these variables introduce contextual redundancy or noise for the adopted short-horizon forecasting task. Based on this analysis, the subsequent dual-timescale models retain only the positively contributing variables to construct a more compact and effective slow-timescale representation.

Table 2. Feature contribution ranking of meteorological-semantic variables.

Feature	Feature Type	MAPE	ΔMAPE	Contribution
DWD	Categorical	2.5301	0.1420	Positive
NWD	Categorical	2.5183	0.1303	Positive
NWL	Categorical	2.4509	0.0628	Positive
MaxT	Numerical	2.4156	0.0275	Positive
WI	Numerical	2.3952	0.0072	Positive
DWL	Categorical	2.3893	0.0012	Positive
WD	Categorical	2.3360	−0.0520	Negative
WN	Categorical	2.3347	−0.0533	Negative
AT	Numerical	2.3167	−0.0713	Negative
MinT	Numerical	2.3121	−0.0759	Negative

3.4. Overall Performance Comparison

To evaluate the overall effectiveness of the proposed framework, we compare it with both single-timescale and dual-timescale baseline models. The single-timescale baselines (denoted by “F-”) include F-LSTM, F-GRU, F-TCN, and F-Transformer, which use only fast-timescale load observations. The dual-timescale baselines (denoted by “D-”) incorporate both fast-timescale load data and slow-timescale meteorological-semantic information. Relative to these baselines, the proposed method further introduces structure-guided semantic refinement, cross-scale attention fusion, coarse-to-fine prediction, and intervention-based consistency regularization.

As shown in Table 3 and Figure 2, the proposed method achieves the best overall performance among all compared models. It attains an MAE of 3828.11 kW, an RMSE of 7942.71 kW, and a MAPE of 2.93%. Compared with the best single-timescale baseline, F-Transformer, the proposed model improves MAE, RMSE, and MAPE by 6.44%, 5.62%, and 6.98%, respectively. Compared with the best dual-timescale baseline, D-Transformer, the improvements reach 14.56%, 4.36%, and 9.01%, respectively. These results indicate that performance gains do not come merely from adding slow-timescale inputs; instead, they arise from more effective refinement and utilization of contextual information.

Table 3. Overall performance comparison results.

Method	MAE (kW)	RMSE (kW)	MAPE (%)	R ² (%)	Test Time (s)
F-LSTM	4826.02	8408.10	3.39	94.48	9.27
F-GRU	4207.37	8291.08	3.12	94.63	10.22
F-TCN	6109.56	11532.63	3.91	89.61	9.28
F-Transformer	4091.54	8416.18	3.15	94.47	9.34
D-LSTM	5614.10	9087.35	3.74	93.55	15.11
D-GRU	4511.17	8287.63	3.27	94.89	15.01
D-TCN	5153.45	8715.43	3.49	94.07	15.61
D-Transformer	4480.47	8304.36	3.22	94.35	14.58
Ours	3828.11	7942.71	2.93	95.59	14.63

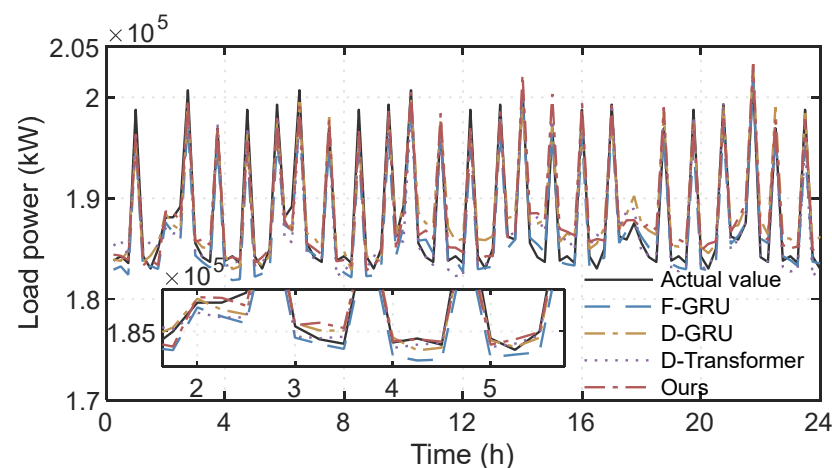


Figure 2. Comparison of overall test results (one day at random).

More importantly, the comparison highlights the novelty of the proposed structure-guided dual-timescale design. The proposed method differs from ordinary single-timescale models by explicitly separating fast load dynamics from slow contextual semantics. It also differs from conventional dual-timescale baselines by introducing structure-guided semantic refinement and structure-aware cross-scale attention before prediction. Therefore, the observed improvement is not simply caused by using more input variables, but by modeling heterogeneous temporal sources according to their different structural roles.

The test time of the proposed model (14.63 s) is close to that of the dual-timescale baselines, indicating that the accuracy improvement is achieved without a significant computational burden. This result suggests that the proposed framework can balance forecasting performance and inference efficiency on the adopted validation task.

3.5. Training-Stage Performance and Computational Cost

To further evaluate the fitting behavior and computational cost during the training phase, representative models are additionally evaluated on the training set. The selected methods include F-GRU, F-Transformer, D-Transformer, and the proposed model. These

models respectively represent recurrent single-timescale modeling, attention-based single-timescale modeling, ordinary dual-timescale attention-based modeling, and the proposed structure-guided dual-timescale framework.

Table 4 reports the training-stage prediction performance and training time of the representative models. Compared with the single-timescale baselines, the two dual-timescale models achieve lower training errors, indicating that slow-timescale contextual information is useful not only in the test stage but also during the model fitting process. Among the representative models, the proposed model achieves the best Train RMSE, Train MAPE, and Train R^2 , although D-Transformer obtains a slightly lower Train MAE. This suggests that the proposed framework can effectively fit the training data while maintaining good agreement with the actual load profile.

Table 4. Training-stage performance and computational cost of representative models.

Method	MAE (kW)	RMSE (kW)	MAPE (%)	R^2 (%)	Train Time (s)
F-GRU	4310.22	9148.31	2.76	0.9664	1134.51
F-Transformer	4157.05	9139.30	2.68	0.9665	1415.44
D-Transformer	3663.89	8106.87	2.35	0.9736	1645.98
Ours	3710.90	8000.93	2.34	0.9743	1711.20

In terms of computational cost, the proposed model requires a longer training time than the single-timescale baselines because it includes additional structure-guided semantic refinement, cross-scale attention fusion, coarse-to-fine prediction, and consistency regularization modules. However, its training time is only slightly higher than that of D-Transformer, indicating that the additional structural components introduce acceptable computational overhead. Therefore, the proposed model provides a reasonable balance between training-stage fitting ability and computational cost.

3.6. Statistical Significance Analysis

Although the overall comparison results show that the proposed model achieves lower forecasting errors than the representative baselines, it is still necessary to examine whether these differences are statistically significant. Therefore, a window-level paired statistical analysis is further conducted. Since the prediction horizon of the forecasting task is 16 load steps, 16 consecutive 15 min forecasting points are grouped as one prediction window. For each prediction window, MAE, RMSE, and MAPE are calculated for the proposed model and the representative baseline models. Then, paired statistical tests are performed over the same prediction windows.

Three representative baselines are selected for statistical comparison, including F-GRU, F-Transformer, and D-Transformer. F-GRU represents a strong recurrent single-timescale model, F-Transformer represents a strong attention-based single-timescale model, and D-Transformer represents a representative dual-timescale baseline. Since all models are evaluated on the same test samples, the window-level errors of the proposed model and each baseline naturally form paired observations. The null hypothesis is that there is no significant difference between the errors of the proposed model and those of the corresponding baseline. The paired t-test is used to evaluate the significance of the performance difference.

As shown in Table 5, the proposed model achieves lower average window-level MAE, RMSE, and MAPE than all representative baselines. The differences are statistically significant at the 0.001 level for all reported metrics. Specifically, compared with F-GRU, the proposed model reduces the window-level MAE, RMSE, and MAPE by 349.25 kW, 374.72 kW, and 0.1696 percentage points, respectively. Compared with F-Transformer, the corresponding reductions are 233.42 kW, 335.00 kW, and 0.0896 percentage points. The improvement is more pronounced when compared with D-Transformer, where the

reductions reach 782.36 kW, 788.65 kW, and 0.3424 percentage points, respectively. These results indicate that the performance gain of the proposed model is not caused by isolated test samples, but is consistently observed across prediction windows.

Table 5. Window-level statistical significance analysis between the proposed model and representative baselines.

Comp.	Metric	Ours	Base.	Diff.	<i>p</i>	Sig.
Ours-F-GRU	MAE	3858.12 ± 4606.36	4207.37 ± 5252.87	−349.25	<0.001	***
Ours-F-GRU	RMSE	4707.05 ± 6071.98	5081.78 ± 6551.32	−374.72	<0.001	***
Ours-F-GRU	MAPE (%)	2.95 ± 32.49	3.12 ± 32.47	−0.17	<0.001	***
Ours-F-Trans.	MAE	3858.12 ± 4606.36	4091.54 ± 5205.81	−233.42	<0.001	***
Ours-F-Trans.	RMSE	4707.05 ± 6071.98	5042.06 ± 6738.85	−335.00	<0.001	***
Ours-F-Trans.	MAPE (%)	2.95 ± 32.49	3.04 ± 32.46	−0.09	<0.001	***
Ours-D-Trans.	MAE	3858.12 ± 4606.36	4640.47 ± 5211.62	−782.36	<0.001	***
Ours-D-Trans.	RMSE	4707.05 ± 6071.98	5495.71 ± 6490.26	−788.65	<0.001	***
Ours-D-Trans.	MAPE (%)	2.95 ± 32.49	3.30 ± 32.05	−0.34	<0.001	***

Note: The paired tests are conducted using window-level grouped forecasting errors computed over the same test samples. Each prediction window contains 16 consecutive 15 min forecasting points. Diff. denotes ours minus baseline; negative values indicate lower forecasting errors of the proposed model. *** denotes statistical significance at the 0.001 level.

3.7. Ablation Study

We conducted an ablation study to evaluate the contribution of each component in the proposed model. The variants tested include fast-only, using only fast-timescale inputs; baseline dual-timescale, which combines fast and slow inputs without additional enhancements; w/o Causal Refinement (w/o CR), w/o Cross-Scale Attention (w/o CSA), w/o Coarse-to-Fine (w/o CTF), and w/o Counterfactual Consistency (w/o CC), where key components are removed to assess their impact.

The results in Table 6 and Figure 3 show a clear performance hierarchy. Fast-only performs the worst, with an MAE of 4630.51 kW and a MAPE of 3.33%, indicating that recent high-frequency observations alone are insufficient for capturing the full forecasting context. Baseline DT improves upon fast-only (MAE: 3979.62 kW; MAPE: 3.02%), confirming the value of introducing slow-timescale contextual information. However, removing any proposed component from the full model leads to performance degradation, which demonstrates that the gain is produced by the coordinated design of the framework rather than by a single isolated module.

Table 6. Ablation tests performance comparison results.

Method	MAE (kW)	RMSE (kW)	MAPE (%)	R ²	Test Time (s)
Fast-only	4630.51	8604.74	3.33	94.21	13.58
Baseline DT	3979.62	8179.18	3.02	94.77	14.08
w/o CR	4263.60	8253.17	3.13	94.68	13.90
w/o CSA	3897.89	7743.40	2.94	95.31	16.69
w/o CTF	4302.11	8507.78	3.15	94.35	14.53
w/o CC	3857.50	7595.55	2.96	95.39	15.70
Ours	3681.69	7636.52	2.87	95.61	13.83

Among the ablated variants, removing structure-guided refinement or coarse-to-fine prediction causes relatively larger performance drops, suggesting that suppressing unstable semantic information and separating background tendency from local fluctuation are both important to the final prediction quality. Removing cross-scale attention or intervention-based consistency also degrades performance, indicating that adap-

tive fusion and robustness-oriented regularization further enhance the usefulness of slow-timescale context.

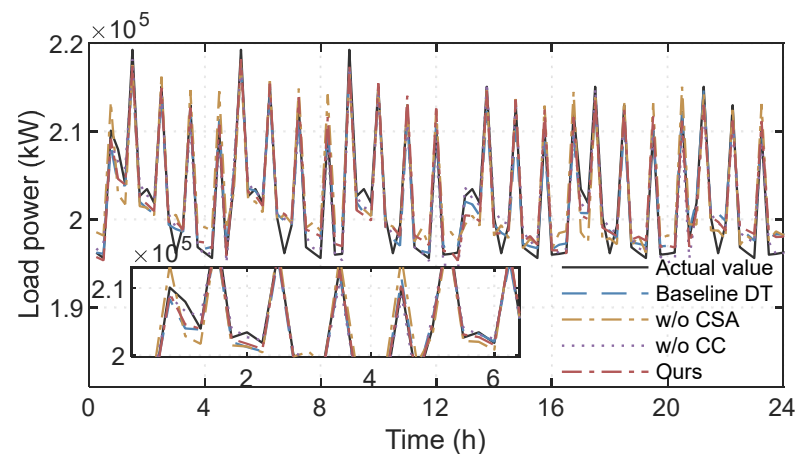


Figure 3. Comparison of ablation test results (one day at random).

The full model achieves the best overall results, with an MAE of 3681.69 kW, an RMSE of 7636.52 kW, and a MAPE of 2.87%. These results verify that each component contributes to the final framework and that their integration yields the strongest predictive accuracy and robustness.

The ablation results further explain the source of the model novelty. Removing structure-guided refinement weakens the model's ability to suppress unstable contextual components, while removing cross-scale attention reduces its ability to adaptively align fast dynamics with slow semantic information. In addition, removing the coarse-to-fine prediction structure weakens the separation between background tendency modeling and local fluctuation correction. These observations demonstrate that the proposed framework improves forecasting performance through coordinated structural modeling rather than through a single isolated module or a simple increase in model complexity.

3.8. Robustness Evaluation Under Different Operating Scenarios

To further assess robustness under practically relevant uncertainties, four testing scenarios are constructed according to the explicit implementation in the code: summer, winter, network interference, and weather anomaly. All models are trained using the same training set, while scenario-specific selection or perturbation is introduced only during evaluation. Therefore, the reported results reflect robustness and generalization under a unified training condition rather than scenario-dependent retraining.

The first two scenarios correspond to natural seasonal distribution shifts. Specifically, the summer scenario contains test samples whose timestamps fall in June, July, and August, whereas the winter scenario contains test samples from December, January, and February. No artificial perturbation is introduced in these two cases. Instead, they evaluate whether the learned forecasting models can maintain satisfactory predictive accuracy when load characteristics vary with season, temperature conditions, and weather-driven demand behavior.

The third scenario, denoted as network interference, is designed to emulate fast-timescale data corruption. Let X^f denote the fast-timescale input sequence. The perturbed input is generated as

$$\tilde{X}^f = (X^f + \epsilon) \odot M \quad (33)$$

where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ is additive Gaussian noise with $\sigma_n = 0.10$ is a Bernoulli mask whose entries are sampled as $M_{i,j} \sim \text{Bernoulli}(1 - p_d)$ with $p_d = 0.05$. This design simulates communication disturbance, sensor noise, and random packet loss in recent load observations. Since short-horizon forecasting strongly relies on high-frequency temporal inputs, this scenario directly examines model robustness against corruption of the fast-timescale branch.

The fourth scenario, referred to as a weather anomaly, targets uncertainty in the slow-timescale meteorological-semantic input. For numerical temperature-related variables, the perturbed input is generated as

$$\tilde{\mathbf{X}}_{[\dots,0]}^{s,\text{num}} = \mathbf{X}_{[\dots,0]}^{s,\text{num}} + \delta_T \quad (34)$$

where

$$\delta_T = \text{clip}(\eta, -c_T, c_T), \eta \sim \mathcal{N}(0, \sigma_T^2) \quad (35)$$

with $\sigma_T = 2.5$ and $c_T = 6.0$. For each categorical weather descriptor, random semantic replacement is further applied with probability $p_c = 0.35$, i.e.,

$$\tilde{x}_{i,j}^{s,\text{cat}} = \begin{cases} r_{i,j}, & z_{i,j} = 1 \\ x_{i,j}^{s,\text{cat}}, & z_{i,j} = 0 \end{cases} \quad (36)$$

where $z_{i,j} \sim \text{Bernoulli}(p_c)$ and $r_{i,j}$ is a randomly sampled valid category index. This scenario mimics weather forecast bias, abrupt meteorological shifts, and semantic-context distortion. Compared with the network interference scenario, which perturbs recent observations, the weather anomaly scenario evaluates the sensitivity of the forecasting models to unreliable slow-timescale contextual information.

Table 7 and Figure 4 report the MAE, RMSE, and MAPE results under the four testing scenarios. Several observations can be drawn. In the summer scenario, the proposed method achieves the best results on all three metrics, with an MAE of 3310.56, an RMSE of 6692.71, and a MAPE of 3.42%. Relative to F-Transformer, the proposed framework improves MAE, RMSE, and MAPE by 13.52%, 3.95%, and 7.07%, respectively. The larger gains over F-GRU further suggest that dual-timescale modeling and semantic refinement are especially beneficial when load–weather coupling becomes stronger during high-temperature periods.

Table 7. Performance metrics comparison under different operating scenarios.

Scenario	Metric	Method			
		F-Transformer	F-GRU	D-Transformer	Ours
Summer scenario	MAE (kW)	3827.88	4495.38	4442.83	3310.56
	RMSE (kW)	6967.79	7570.66	7511.61	6692.71
	MAPE (%)	3.68	4.01	3.90	3.42
Winter scenario	MAE (kW)	5570.05	4829.24	6693.56	4744.04
	RMSE (kW)	7925.60	6730.32	8772.47	7039.31
	MAPE (%)	3.79	3.54	4.64	3.20
Network Interference scenario	MAE (kW)	6972.59	7478.08	6867.09	6277.50
	RMSE (kW)	10,785.02	11,287.96	10534.32	10,259.97
	MAPE (%)	4.35	4.62	4.35	4.10
Weather anomaly scenario	MAE (kW)	4367.25	4874.87	5013.90	3739.96
	RMSE (kW)	7924.26	8262.61	8536.19	7639.13
	MAPE (%)	3.20	3.43	3.47	2.88

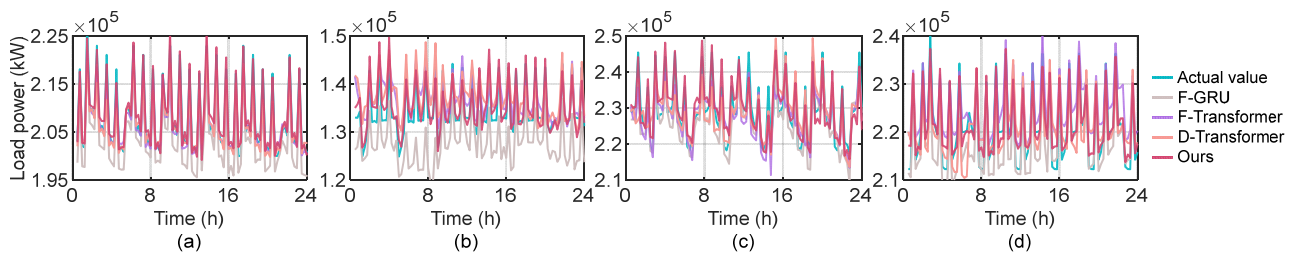


Figure 4. Comparison of robustness test results under different operating scenarios.

In the winter scenario, the proposed method again achieves the lowest MAE and MAPE, reaching 4744.04 and 3.20%, respectively, while F-GRU yields a slightly lower RMSE of 6730.32 compared with 7039.31 for the proposed model. This indicates that the proposed framework provides the best overall pointwise prediction accuracy, although the GRU baseline remains competitive in suppressing several larger errors under winter conditions. More importantly, the proposed method still substantially outperforms D-Transformer, reducing MAE, RMSE, and MAPE by 29.12%, 19.75%, and 31.03%, respectively. This result shows that simply introducing dual-timescale inputs is insufficient unless the slow-timescale information is effectively refined and fused.

In the network interference scenario, all models experience visible performance degradation, confirming that corruption of recent observations severely affects short-horizon forecasting. Nevertheless, the proposed model remains the best-performing method across all three metrics, with an MAE of 6277.50, an RMSE of 10,259.97, and a MAPE of 4.10%. Compared with the strongest baseline, D-Transformer, the proposed model reduces MAE and RMSE by 8.58% and 2.60%, respectively, while lowering MAPE from 4.35% to 4.10%. These results indicate that the proposed architecture is not only accurate under clean inputs but also more resilient to high-frequency measurement disturbance.

In the weather anomaly scenario, the superiority of the proposed method becomes even more evident. It achieves the best MAE, RMSE, and MAPE values of 3739.96, 7639.13, and 2.88%, respectively. Relative to F-Transformer, the proposed model improves MAE, RMSE, and MAPE by 14.37%, 3.60%, and 10.00%, respectively. Compared with D-Transformer, the corresponding improvements increase to 25.41%, 10.51%, and 17.00%. These results verify that the proposed structure-guided semantic refinement and dual-timescale fusion mechanism can more effectively filter unreliable meteorological-semantic information and preserve predictive stability under contextual anomalies.

Overall, the results across all four scenarios consistently show that the proposed method delivers the best or near-best performance under seasonal distribution shifts, fast-timescale observation corruption, and slow-timescale semantic perturbation. This consistency indicates that the gain of the proposed framework does not rely on a specific operating condition, but stems from its stronger ability to jointly exploit short-term dynamics and slowly varying context in a robust and structured manner.

3.9. Analysis of Difficult-to-Predict Samples

To further analyze the data points that are relatively difficult for the proposed model to predict, an error-distribution analysis is conducted based on the temporal characteristics of the test samples. Since the forecasting task is formulated as a multi-step prediction problem with a prediction horizon of 16 load steps, the exported prediction results contain flattened multi-step outputs. In other words, the same target timestamp may appear in multiple overlapping prediction windows. Directly treating all flattened prediction rows as independent time points would over-count repeated target samples and may lead to a biased error-distribution analysis.

To address this issue, the flattened prediction results are first reconstructed according to the prediction origin and the forecast horizon step. Each predicted value is mapped back to its corresponding target timestamp using the original 15 min load timestamp sequence. For a target timestamp that is predicted by multiple overlapping windows, the absolute errors from different prediction windows are averaged to obtain a unique forecasting error for that timestamp. In this way, a corrected non-duplicated error series is obtained and used for subsequent temporal error analysis.

Based on the corrected error series, an hour–month MAE heatmap is constructed for the proposed model. Specifically, the unique absolute errors are grouped according to the month and the hour of day of their corresponding target timestamps, and the mean absolute error is calculated for each month–hour cell. This heatmap provides a visual way to identify the temporal regions in which the forecasting task becomes more difficult.

As shown in Figure 5, the forecasting errors are not uniformly distributed over time. Relatively larger errors are concentrated in several specific month–hour regions, especially during parts of February, April, May, June, and August. These high-error regions indicate that some periods are more difficult to forecast than others. Such difficulty is likely related to stronger load variability, peak-load transitions, weather-sensitive operating conditions, or abrupt changes in the daily load profile. In contrast, several other time periods show relatively lower MAE values, suggesting that the proposed model can capture regular intraday patterns more effectively when the load evolution is more stable.

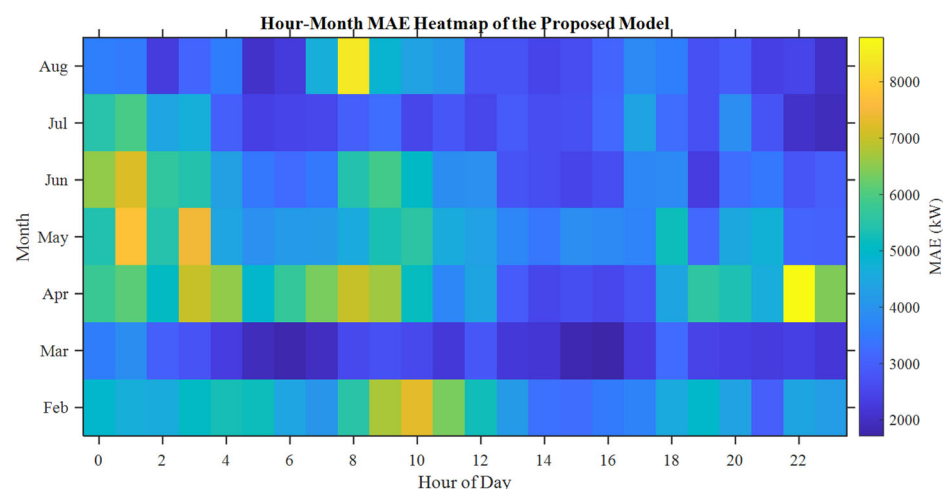


Figure 5. Hour–month MAE heatmap of the proposed model after removing duplicated multi-step prediction targets. Blank cells indicate month–hour combinations without test samples.

It should also be noted that the test set does not cover all calendar months. Therefore, blank cells in the heatmap indicate months or month–hour combinations without test samples rather than zero forecasting error. This analysis further confirms that the model’s prediction difficulty is temporally heterogeneous. It also suggests that future work may improve the model by introducing uncertainty-aware forecasting, adaptive loss weighting for high-error periods, and enhanced modeling of abrupt load changes.

3.10. Sensitivity Analysis

To further examine the influence of key hyperparameters, a one-factor-at-a-time sensitivity analysis is conducted under the same 16-step forecasting setting as the main experiments. Three factors are considered: the fast-timescale history length L , the slow-timescale semantic window length K , and the regularization scale RegScale . The default setting is $L = 96$, $K = 7$, and $\text{RegScale} = 1.0$. In each group, only one factor is changed while the remaining

factors are fixed at their default values. Therefore, the analysis focuses on the individual influence of each factor rather than performing a full grid-search optimization.

As shown in Table 8, the model performance is affected by the fast-timescale history length. When L increases from 48 to 144, the MAE decreases from 3792.84 kW to 3433.33 kW, and the MAPE decreases from 2.91% to 2.73%. This indicates that, for the 16-step forecasting task, a longer fast-timescale history can provide more useful information for capturing daily load patterns and short-term trend evolution. Although $L = 96$ corresponds to a physically meaningful 24 h input window, the sensitivity results suggest that extending the history window to 36 h may further improve multi-step forecasting accuracy.

Table 8. One-factor-at-a-time sensitivity analysis of the proposed model under the 16-step forecasting setting.

Factor	Level	MAE	RMSE	MAPE (%)
L	48	3792.84	7715.00	2.91
L	96	3966.90	7928.49	2.98
L	144	3433.33	7336.49	2.73
K	3	4258.46	8521.47	3.12
K	7	3966.90	7928.49	2.98
K	14	3644.33	7501.03	2.85
RegScale	0.5	3796.66	7679.02	2.91
RegScale	1.0	3966.90	7928.49	2.98
RegScale	2.0	3985.55	7792.00	3.01

The slow-timescale semantic window also has a clear influence on performance. The setting $K = 3$ leads to the largest errors, while $K = 14$ achieves the best results among the tested levels. This suggests that a short semantic history may be insufficient for characterizing slowly varying meteorological and calendar-related background conditions. For a 4 h forecasting horizon, a longer semantic context can provide more stable background information and improve the utilization of slow-timescale variables.

For the regularization scale, reducing RegScale from 1.0 to 0.5 improves MAE, RMSE, and MAPE, whereas increasing it to 2.0 does not consistently improve all metrics. This result indicates that structure-related regularization is useful, but overly strong regularization may constrain the model and slightly weaken point forecasting accuracy. Overall, the sensitivity analysis shows that the proposed framework remains effective under different parameter settings, while longer historical and semantic windows can further enhance performance for the 16-step forecasting task.

It should be noted that this analysis is intended to evaluate parameter sensitivity rather than to replace the default configuration used in the main comparison experiments. The default setting, $L = 96$ and $K = 7$, corresponds to a 24 h fast-timescale history and a one-week slow-timescale semantic context, which provides a clear and interpretable experimental configuration. The sensitivity results further indicate that additional performance gains may be obtained by extending the temporal context, which will be considered in future work.

4. Conclusions

This paper presented a structure-guided dual-timescale learning framework for heterogeneous temporal forecasting. The proposed model explicitly separates fast-timescale dynamic modeling from slow-timescale contextual representation, and further improves cross-scale information utilization through structure-guided semantic refinement, cross-scale attention fusion, coarse-to-fine prediction, and intervention-based consistency regularization. Unlike narrowly task-specific formulations, the proposed framework is designed at

the methodological level and is applicable to forecasting problems in which high-frequency target dynamics coexist with slowly varying exogenous context.

A real-world ultra-short-term aggregate load forecasting task was adopted as a representative validation case. The experimental results demonstrated that the proposed method consistently outperformed competitive baselines in terms of accuracy, robustness, and efficiency. The robustness analysis further showed that the framework maintains strong predictive performance under seasonal distribution shifts, fast-timescale observation corruption, and slow-timescale contextual anomalies. These results verify that the proposed approach is not merely effective for the selected load dataset, but also promising as a general solution for heterogeneous-timescale forecasting problems. Therefore, the proposed framework provides practical value for short-horizon forecasting applications that require both accurate prediction and robust utilization of slowly varying contextual information, such as load forecasting, renewable energy forecasting, traffic prediction, and industrial process monitoring.

5. Discussion

In this study, we proposed a structure-guided dual-timescale learning framework for heterogeneous short-horizon time-series forecasting, and the experimental results on a real-world industrial load forecasting task demonstrate consistent improvements over both single- and dual-timescale baselines. The performance gains are primarily attributed to the synergistic integration of structure-guided semantic refinement, which suppresses spurious or unstable contextual components as evidenced by the feature contribution analysis, and cross-scale attention that enables the fast branch to dynamically query the most relevant slow-varying context rather than relying on naive concatenation. The ablation study confirms that removing any key component—especially structure-guided refinement or coarse-to-fine prediction—substantially degrades accuracy, indicating that the framework's strength lies in the structural coordination among its modules. Furthermore, the model exhibits remarkable robustness under four challenging scenarios (seasonal shifts, network interference, and weather anomalies), largely due to the decoupled coarse-to-fine prediction head that isolates background tendency from local fluctuations, and the intervention-based consistency regularization that enforces causally plausible responses to semantic changes. Compared with existing methods such as Temporal Fusion Transformers or decomposition-based architectures, our approach explicitly respects the structural asymmetry between fast and slow timescales and provides a principled way to filter uninformative contextual variables.

From a practical perspective, the proposed framework is valuable for short-horizon operational forecasting tasks in which high-frequency target observations coexist with low-frequency contextual information. In power system applications, such forecasts can support dispatch scheduling, reserve planning, demand response, and short-term energy management. The robustness results under seasonal shifts, network interference, and weather anomalies further indicate that the proposed model can maintain stable forecasting performance when data quality or operating conditions change. This property is particularly important for real-world decision-making scenarios, where forecasting models are expected to remain reliable under imperfect measurements and uncertain contextual inputs.

Nevertheless, several limitations should be acknowledged. First, the structure-guided semantic refinement module currently relies on a domain-informed dependency structure, and automatic structure discovery from data remains a direction for future work. Second, the fast–slow alignment assumes a fixed mapping between high-frequency target observations and low-frequency contextual records, which may need to be extended for irregularly sampled or asynchronously updated data. Third, although the proposed

framework is designed for general heterogeneous-timescale forecasting, the empirical validation is conducted on a representative load forecasting task, and further tests on other domains such as renewable energy, traffic, and industrial processes would help verify its broader applicability. Finally, this study focuses on deterministic point forecasting, while probabilistic forecasting and uncertainty quantification will be considered in future work. Overall, this study offers a robust and interpretable solution for heterogeneous short-horizon forecasting problems.

Author Contributions: Investigation, T.P. and J.Y.; writing—original draft preparation, T.P. and J.Y.; formal analysis, M.Q. and Y.B.; visualization, M.Q. and Y.B.; conceptualization, H.W.; supervision, H.W.; writing—review and editing, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the R&D Program of the Beijing Municipal Education Commission (Grant No. KM202414019003). The APC was funded by the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All experimental datasets analysed in this study are openly accessible via the public Baidu AI Studio repository (repository name: Universal-Tool/c8337) at <https://aistudio.baidu.com/datasetdetail/140138> (accessed on 15 May 2026).

Acknowledgments: The authors would like to thank the anonymous reviewers and the editor for their helpful comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nkongolo, M. Using ARIMA to predict the growth in the subscriber data usage. *Eng* **2023**, *4*, 92–120. [CrossRef]
2. Soares, L.J.; Medeiros, M.C. Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data. *Int. J. Forecast.* **2008**, *24*, 630–644. [CrossRef]
3. Hong, T.; Fan, S. Probabilistic electric load forecasting: A tutorial review. *Int. J. Forecast.* **2016**, *32*, 914–938. [CrossRef]
4. Kuster, C.; Rezgui, Y.; Mourshed, M. Electrical load forecasting models: A critical systematic review. *Sustain. Cities Soc.* **2017**, *35*, 257–270. [CrossRef]
5. Hafeez, F.; Arfeen, Z.A.; Masud, M.I.; Azam, M.K.; Al-Shammari, S.; Aman, M.; Hamid, M.; Haq, M.I.U. Accurate Solar Radiation Forecasting Using Spectral Feature Engineering and Bayesian Optimization. *Eng* **2026**, *7*, 77. [CrossRef]
6. Minaev, Y.M.; Filimonova, O.Y.; Minaeva, Y.I. Forecasting of fuzzy time series based on the concept of the nearest fuzzy sets and tensor models of time series. *Cybern. Syst. Anal.* **2023**, *59*, 165–176. [CrossRef]
7. Cortez, J.C.; Lopez, J.C.; Ullon, H.R.; Giesbrecht, M.; Rider, M.J. Day-ahead photovoltaic power forecasting using deep learning with an autoencoder-based correction strategy. *J. Control Autom. Electr. Syst.* **2024**, *35*, 662–676. [CrossRef]
8. Li, Q.; Zhang, X.; Wang, S.; Wei, J. DPANet: Dual Pyramid Attention Network for Multivariate Time Series Forecasting. *arXiv* **2025**, arXiv:2509.14868. [CrossRef]
9. Christen, R.; Mazzola, L.; Denzler, A.; Portmann, E. Exogenous data for load forecasting: A review. In Proceedings of the 12th International Joint Conference on Computational Intelligence (IJCCI 2020), Budapest, Hungary, 2–4 November 2020; pp. 489–500. [CrossRef]
10. Shataashvili, A.D.; Didmanidze, I.S.; Kakhiani, G.A.; Fomina, T.A. A method of preliminary forecasting of time series of financial data. *Cybern. Syst. Anal.* **2020**, *56*, 296–302. [CrossRef]
11. Lima, J.M.; Guetter, A.K.; Freitas, S.R.; Panetta, J.; de Mattos, J.G. A meteorological-statistic model for short-term wind power forecasting. *J. Control Autom. Electr. Syst.* **2017**, *28*, 679–691. [CrossRef]
12. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [CrossRef]
13. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271. [CrossRef]

14. Van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499. [[CrossRef](#)]
15. Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191. [[CrossRef](#)]
16. Wen, R.; Torkkola, K.; Narayanaswamy, B.; Madeka, D. A multi-horizon quantile recurrent forecaster. *arXiv* **2017**, arXiv:1711.11053. [[CrossRef](#)]
17. Rangapuram, S.S.; Seeger, M.; Gasthaus, J.; Stella, L.; Wang, Y.; Januschowski, T. Deep state space models for time series forecasting. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018; pp. 7796–7805. Available online: <https://papers.nips.cc/paper/2018/hash/5cf68969fb67aa6082363a6d4e6468e2-Abstract.html> (accessed on 12 March 2026).
18. Oreshkin, B.N.; Carпов, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26 April–1 May 2020. [[CrossRef](#)]
19. Challu, C.; Olivares, K.G.; Oreshkin, B.N.; Ramirez, F.G.; Canseco, M.M.; Dubrawski, A. N-HiTS: Neural hierarchical interpolation for time series forecasting. In Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 6989–6997. [[CrossRef](#)]
20. Lim, B.; Arik, S.O.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [[CrossRef](#)]
21. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient Transformer for long sequence time-series forecasting. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 11106–11115. [[CrossRef](#)]
22. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with auto-correlation for long-term series forecasting. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Virtual Event, 6–14 December 2021; pp. 22419–22430. Available online: <https://dl.acm.org/doi/10.5555/3540261.3541978> (accessed on 15 April 2026).
23. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. FEDformer: Frequency enhanced decomposed Transformer for long-term series forecasting. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 27268–27286. Available online: <https://proceedings.mlr.press/v162/zhou22g.html> (accessed on 20 April 2026).
24. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are Transformers effective for time series forecasting? In Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 14 February 2023; pp. 11121–11128. [[CrossRef](#)]
25. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-variation modeling for general time series analysis. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. [[CrossRef](#)]
26. Das, A.; Kong, W.; Leach, A.; Mathur, S.; Sen, R.; Yu, R. Long-term Forecasting with TiDE: Time-series Dense Encoder. *arXiv* **2024**, arXiv:2304.08424. [[CrossRef](#)]
27. Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; Long, M. TimeXer: Empowering Transformers for time series forecasting with exogenous variables. In Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; pp. 469–498. [[CrossRef](#)]
28. Runge, J.; Gerhardus, A.; Varando, G.; Eyring, V.; Camps-Valls, G. Causal inference for time series. *Nat. Rev. Earth Environ.* **2023**, *4*, 487–505. [[CrossRef](#)]
29. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with Transformers. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. [[CrossRef](#)]
30. Liu, Y.; Wu, H.; Wang, J.; Long, M. Non-stationary Transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2022; Volume 35, pp. 9881–9893. Available online: <https://arxiv.org/html/2205.14415v4> (accessed on 5 May 2026).
31. Chen, C.; Chai, L.; Wang, Q. A wind power ultra-short-term ensemble forecast framework considering wind speed correction and scenario classification. *Energy* **2025**, *336*, 138276. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.