NEIN-TSF: NOT EVERY IRREGULARITY IS NOISE IN TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern time-series forecasting models capture increasingly fine-grained historical dynamics, yet not every observed pattern should be extrapolated. Irregular changes may be transient disturbances or genuine regime shifts, while seemingly stable patterns may lose predictive relevance beyond the input window. We propose NEIN-TSF, a unified framework that regulates historical dynamics through forecasting consequence, historical support, and correction utility. Forecast-Damage Ranking Pretraining (FDRP) learns relative teacher-derived forecast damage from matched local counterfactuals, encouraging the backbone to capture predictive consequence rather than input-space deviation. Supported Predictive Evidence (SPE) combines historical retrieval, periodicity, local trends, and persistence to produce bounded, support-conditioned residual corrections. Horizon-Adaptive Evidence Utility (HAEU) calibrates correction strength across prediction positions, while Chronological Forward Utility Selection (CFUS) selects model states on chronologically later validation data. Experiments on fifteen real-world benchmarks across multiple horizons show consistent MSE and MAE improvements over strong recent baselines. Mechanism analyses further show stronger alignment between FDRP-predicted and teacher-derived damage than raw input deviation, positive support–benefit associations, and horizon- and time-dependent correction utility.

1 INTRODUCTION

Time-series forecasting is central to applications such as energy, transportation, weather, and industrial monitoring. Recent Transformer, linear/MLP, frequency-domain, decomposition, and multi-scale models have substantially improved temporal representation learning (Zeng et al., 2023; Nie et al., 2022; Wu et al., 2022; Liu et al., 2023a; Wang et al., 2024). Yet stronger representation capacity does not imply that every historical pattern should be extrapolated. An irregular change may be transient or mark the onset of a new regime, while an apparently stable pattern may lose its predictive relevance beyond the input window. Historical appearance alone is therefore insufficient to determine predictive value. This motivates a fundamental question: **Should forecasting models predict everything they observe?** We argue that forecasting should assess the consequences of historical changes and regulate their influence on future predictions, rather than merely modeling their observed appearance (Wang et al., 2026d; Lu et al., 2025).

We first examine this premise through an event-level counterfactual analysis on hourly Electricity (ECL). For a matched terminal irregularity e, we attenuate the irregular segment toward a seasonal historical reference and define

$$C_e(H) = \mathcal{L}_H\big(f(X_e^{\mathrm{att}}), Y\big) - \mathcal{L}_H(f(X_e), Y)\,,$$

where $C_e(H) > 0$ means that attenuation increases forecast error and hence removes information useful to the evaluated forecaster. Figure 1 reveals that appearance-matched events can evolve differently, similar irregularity magnitudes can induce opposite forecasting consequences, and the consequence of the same event can weaken or change sign as the prediction horizon grows. Thus, neither appearance nor deviation magnitude alone reliably determines what should be propagated into the future.

The distinction is also intuitive in practice. A sudden rise in electricity demand may be a temporary event or the beginning of a sustained demand level, while a traffic surge may reflect either an isolated

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

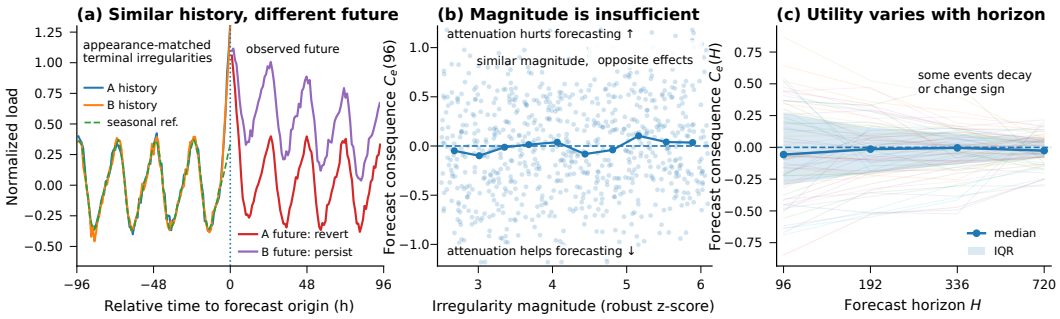


Figure 1: **Empirical motivation: historical irregularity is not equivalent to predictive value.** (a) Appearance-matched events can lead to different future evolutions. (b) Similar irregularity magnitudes can have opposite forecast consequences $C_e(96)$. (c) Event consequence can decay or change sign with forecasting horizon. Events are constructed using historical information only; details are provided in Appendix A.

incident or the onset of a recurrent peak (Wang et al., 2026c;a). More importantly, recognizing that a historical change matters does not by itself specify a reliable correction: a plausible continuation may lack sufficient historical support, and even a useful correction may lose value at distant horizons or under later temporal regimes (Liang et al., 2025). Related work on robust and non-stationary forecasting has emphasized separating informative variation from harmful disturbance (Cheng et al., 2024; Liu et al., 2022; 2024c). We therefore ask three connected questions: *Does a historical change matter for forecasting? Is a corresponding correction supported by history? How does its utility vary across forecasting distance and chronological time?*

To answer these questions, we propose **NEIN-TSF** (**N**ot **E**very **I**rregularity is **N**oise in **T**ime-**S**eries **F**orecasting), which regulates historical dynamics through *forecasting consequence*, *historical support*, and *correction utility*. **Forecast-Damage Ranking Pretraining (FDRP)** learns relative teacher-derived forecast damage from matched local counterfactuals, encouraging the backbone to capture predictive consequence rather than raw deviation. **Supported Predictive Evidence (SPE)** combines retrieval, periodicity, local trend, and persistence to produce bounded, support-conditioned residual refinement of the frozen base forecast. **Horizon-Adaptive Evidence Utility (HAEU)** then calibrates correction strength across prediction positions, while **Chronological Forward Utility Selection (CFUS)** selects model states using strictly later validation data. Together, these components determine *what matters*, *what is historically supported*, and *how strongly supported evidence should influence the forecast*.

In summary, our contributions are:

- We introduce consequence-aware forecasting through FDRP, which learns relative forecast damage from matched counterfactuals rather than using input-space irregularity magnitude as a proxy for predictive relevance.
- We develop historically supported and utility-controlled refinement: SPE produces bounded evidence-conditioned corrections, while HAEU and CFUS regulate their utility across forecasting distance and chronological time.
- We evaluate NEIN-TSF on fifteen real-world benchmarks across multiple horizons, obtaining broad MSE/MAE improvements over strong recent baselines. Mechanism and transferability analyses further examine consequence learning, support–benefit relationships, horizon-dependent utility, and backbone transferability.

2 Related Work

Pretraining and non-stationary dynamics. RevIN (Kim et al., 2021) and Non-stationary Transformers (Liu et al., 2022) address distribution shifts, while SimMTM (Dong et al., 2023), Koopa (Liu et al., 2023b), MOMENT (Goswami et al., 2024), and Timer (Liu et al., 2024d) learn temporal representations through reconstruction, dynamical modeling, or generative pretraining. FDRP instead

learns relative forecasting consequences through teacher-derived damage ranking of matched counterfactuals rather than input-space deviation alone. **Historical structure and retrieval-augmented forecasting.** FEDformer (Zhou et al., 2022), TimesNet (Wu et al., 2022), Crossformer (Zhang & Yan, 2023), and Pathformer (Chen et al., 2024) exploit structural or multiscale temporal dependencies, while RATD (Liu et al., 2024b) and RAFT (Han et al., 2025) augment forecasting with retrieved historical references or continuations. SPE differs by assessing complementary historical evidence and using it to constrain bounded residual refinement, rather than assuming that a plausible historical continuation is necessarily beneficial. **Horizon adaptation and temporal stability.** Scaleformer (Shabani et al., 2022) and TimeMixer (Wang et al., 2024) exploit multiscale information, whereas OneNet (Wen et al., 2023), FOIL (Liu et al., 2024a), and TAFAS (Kim et al., 2025) address temporal distribution shifts through adaptation or invariant learning. HAEU instead calibrates correction strength across prediction positions, while CFUS uses chronologically ordered forward validation for model-state selection.[1]

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

Given a multivariate time series with C variables, we use a historical window of length L to predict the next H time steps. For a batch of size B, the input and target are $X \in \mathbb{R}^{B\times L\times C}$ and $Y \in \mathbb{R}^{B\times H\times C}$, where b, t, h, c index the sample, historical position, prediction position, and variable, respectively. We compute the input-window statistics

$$\mu_{b,c} = \frac{1}{L}\sum_{t=1}^{L} X_{b,t,c}, \qquad \sigma_{b,c} = \sqrt{\frac{1}{L}\sum_{t=1}^{L}(X_{b,t,c} - \mu_{b,c})^2 + \epsilon}, \qquad \widetilde{X}_{b,t,c} = \frac{X_{b,t,c} - \mu_{b,c}}{\sigma_{b,c}}, \quad (1)$$

with $\epsilon > 0$. These statistics are computed exclusively from the observed input window and are broadcast when necessary. Unless otherwise stated, forecasts and candidate continuations remain in the original data scale.

NEIN-TSF regulates historical dynamics through three complementary stages: *forecast-consequence learning*, *historically supported refinement*, and *correction-utility control*. As illustrated in Figure 2, FDRP learns a consequence-aware forecasting backbone from matched counterfactuals and teacher-derived forecast damage; SPE then constructs complementary historical evidence and produces a bounded, support-conditioned residual correction. HAEU calibrates the correction strength across prediction positions, while CFUS selects model states using a strictly later chronological validation period. Importantly, the teacher, auxiliary damage supervision, and CFUS selection procedure are used only during offline learning or calibration and do not introduce additional inference branches.

The resulting forecast can be written compactly as

$$\widehat{Y} = \widehat{Y}^0 + \alpha^\star \odot \Delta Y, \qquad \alpha^\star \in [0, \alpha_{\max}]^H, \quad (2)$$

where $\alpha^\star$ is shared across samples and variables at each prediction position. Thus, FDRP determines a consequence-aware base representation, SPE determines whether and how the base forecast should be refined from historical evidence, and HAEU together with CFUS controls the utility of that refinement across forecasting distance and chronological time. CFUS affects model selection only and incurs no additional inference branch.

3.2 FORECAST-CONSEQUENCE-AWARE PRETRAINING

Matched counterfactual construction. FDRP learns forecast damage rather than relying on predefined normal/abnormal labels or raw deviation magnitude. Let R denote a local contextual reference and $M \in \{0,1\}^{B\times L\times C}$ select a local perturbation region. For perturbation type $q \in \{\text{jump}, \text{noise}, \text{erase}, \text{drift}\}$, we construct matched mild and severe counterfactuals

$$X^a = \mathcal{T}_q(X; \lambda^a) = X + M \odot P_q(\lambda^a), \qquad a \in \{m, s\}, \quad 0 < \lambda^m < \lambda^s, \quad (3)$$

[1] An expanded discussion of related work is provided in Appendix B.

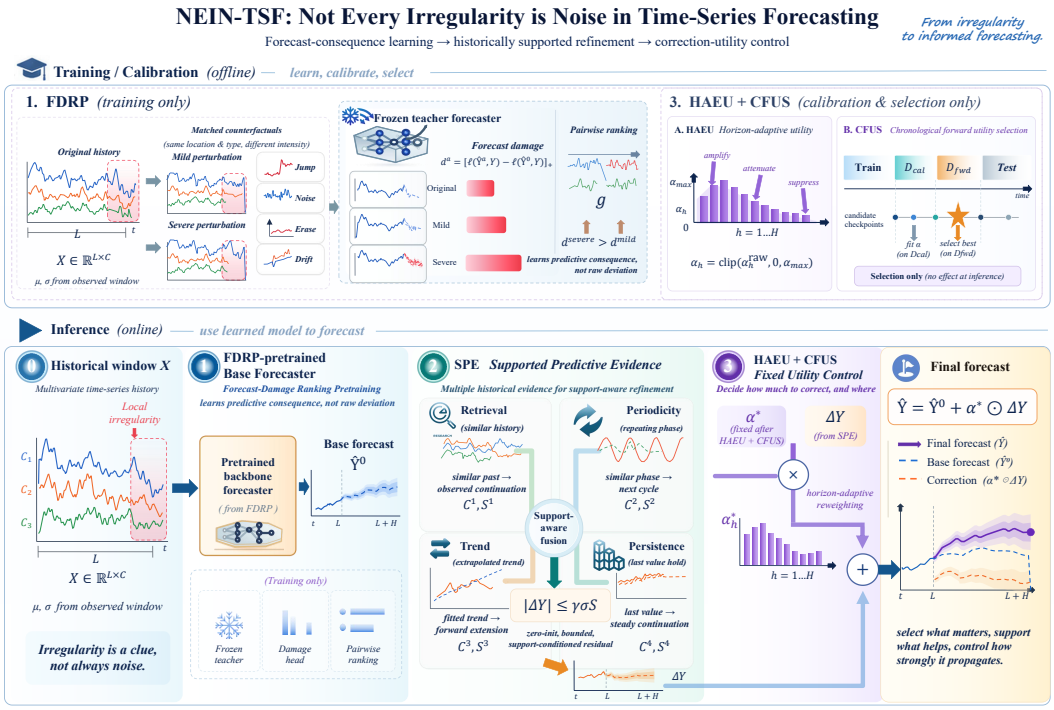


Figure 2: **Overview of NEIN-TSF.** The upper part summarizes offline training, calibration, and model selection: FDRP learns forecast consequences from matched counterfactuals, HAEU calibrates horizon-wise correction strength, and CFUS performs chronological forward selection. The lower part shows online forecasting: the FDRP-pretrained backbone produces the base forecast $\widehat{Y}^0$, SPE constructs historically supported evidence and predicts a bounded residual ΔY, and the fixed utility coefficients $\alpha^\star$ regulate its contribution to the final forecast. Training-only supervision and validation labels are not used at inference.

where the pair shares its temporal location, affected variables, width, perturbation type, and stochastic realization; only the perturbation intensity differs. For erasure, the perturbation interpolates toward R with $\lambda^s \leq 1$. The complete perturbation operators and contextual features are given in Appendix C.1. Matching controls construction factors but does not assume that the stronger perturbation must cause larger forecasting damage.

Teacher-derived forecast damage. We first train a forecaster $f_{\bar{\theta}}$ on unperturbed training data and then freeze it, so that the supervision target does not move during FDRP. Let $X^0 = X$ and $\widehat{Y}^a = f_{\bar{\theta}}(X^a)$. For each sample and prediction position, define

$$\ell_{b,h}(\widehat{Y}, Y) = \frac{1}{C}\sum_{c=1}^{C}(\widehat{Y}_{b,h,c} - Y_{b,h,c})^2. \tag{4}$$

The teacher-derived damage is

$$d^a_{b,h} = \left[\ell_{b,h}(\widehat{Y}^a, Y) - \ell_{b,h}(\widehat{Y}^0, Y)\right]_+, \qquad a \in \{m, s\}, \tag{5}$$

where $[z]_+ = \max(z, 0)$. Thus, $d^a_{b,h}$ measures nonnegative excess teacher error and should be interpreted as a surrogate of predictive sensitivity rather than a causal effect or an intrinsic label of irregularity. The clipping is intentional: FDRP learns relative sensitivity to forecast degradation, whereas the direction of a potentially beneficial correction is deferred to the historical alternatives constructed by SPE.

Table 1: **Historical evidence bank.** Each source captures a complementary historical cue and produces a candidate continuation together with a support score. Support reflects historical agreement rather than a calibrated probability of future correctness.

Source	Historical cue	Candidate continuation	Support	High support indicates
Retrieval	local recurrence	observed matched continuation	$\exp(-d_b^{\mathrm{ret}}/\tau_r)$	close historical match
Periodicity	recurrent phase structure	repeated selected phases	$\mathrm{clip}(A_b(p_b^\star), 0, 1)$	strong periodic agreement
Trend	local directional evolution	least-squares extrapolation	$\exp(-r_{b,c}^{\mathrm{tr}}/\tau_t)$	low trend-fit residual
Persistence	recent level stability	last observation	$\exp(-v_{b,c}^{\mathrm{pers}}/\tau_p)$	low recent variation

Forecast-damage ranking. We retain only counterfactual pairs with a distinguishable damage gap,

$$m_{b,h}^{\mathrm{rank}} = \mathbb{I}\big[\big|d_{b,h}^s - d_{b,h}^m\big| > \delta_d\big], \qquad s_{b,h} = \mathrm{sign}(d_{b,h}^s - d_{b,h}^m), \tag{6}$$

where $\delta_d > 0$ removes nearly tied pairs. The trainable backbone f_θ, initialized from $f_{\bar{\theta}}$, exposes a shared representation to a lightweight auxiliary damage head q_ψ, which predicts $\widehat{d}_{b,h}^a$. Let $N_{\mathrm{rank}} = \sum_{b,h} m_{b,h}^{\mathrm{rank}}$. We optimize

$$\mathcal{L}_{\mathrm{rank}} = \frac{\sum_{b,h} m_{b,h}^{\mathrm{rank}} \log\Big(1 + \exp\Big[-s_{b,h}(\widehat{d}_{b,h}^s - \widehat{d}_{b,h}^m)\Big]\Big)}{N_{\mathrm{rank}} + \epsilon}, \tag{7}$$

and

$$\mathcal{L}_{\mathrm{FDRP}} = \frac{1}{BHC} \sum_{b,h,c} \big(f_\theta(X)_{b,h,c} - Y_{b,h,c}\big)^2 + \lambda_{\mathrm{rank}} \mathcal{L}_{\mathrm{rank}}. \tag{8}$$

The ranking sign is determined by the observed teacher-damage ordering, not by perturbation intensity. The clean forecasting term anchors the backbone to the original task. After pretraining, the auxiliary head q_ψ is discarded and the resulting backbone f_{θ_0} produces the frozen base forecast $\widehat{Y}^0 = f_{\theta_0}(X)$.

3.3 Historically Supported Forecast Refinement

FDRP identifies predictive sensitivity but does not determine how a forecast should be corrected. SPE therefore constructs complementary historical alternatives and uses their support to regulate a residual correction. During this stage, the FDRP backbone f_{θ_0} remains frozen; only the evidence-fusion and correction parameters are learned.

Historical evidence bank. Let $\mathcal{K} = \{\mathrm{ret}, \mathrm{per}, \mathrm{tr}, \mathrm{pers}\}$. Each source $k \in \mathcal{K}$ produces a candidate continuation $C^k \in \mathbb{R}^{B\times H\times C}$ and a support score $S^k \in [0,1]^{B\times H\times C}$. Table 1 summarizes the four sources; complete definitions, retrieval validity conditions, and fallback rules are provided in Appendix C.2. Support measures agreement with a historical rule and is not interpreted as a calibrated probability of future correctness. Retrieval uses valid nonoverlapping historical *matches*; if an observed continuation is unavailable, its support is set to zero and the candidate falls back to the base forecast.

Supported and bounded correction. For each (b, h, c), define normalized candidate deviations and base-forecast dynamics as

$$D_{b,h,c}^k = \frac{C_{b,h,c}^k - \widehat{Y}_{b,h,c}^0}{\sigma_{b,c} + \epsilon}, \qquad U_{b,h,c}^0 = \frac{\widehat{Y}_{b,h,c}^0 - \widehat{Y}_{b,h-1,c}^0}{\sigma_{b,c} + \epsilon}, \tag{9}$$

with $\widehat{Y}_{b,0,c}^0 = X_{b,L,c}$ and normalized forecast position $\eta_h = h/H$. The evidence feature is

$$F_{b,h,c} = \mathrm{Concat}\left(\{D_{b,h,c}^k, S_{b,h,c}^k\}_{k\in\mathcal{K}}, U_{b,h,c}^0, \eta_h\right). \tag{10}$$

A lightweight weighting network a_{ϕ_a} and correction network g_{ϕ_g}, with $\phi = \{\phi_a, \phi_g\}$, produce

$$\pi_{b,h,c} = \mathrm{softmax}(a_{\phi_a}(F_{b,h,c})), \qquad S_{b,h,c} = \sum_{k\in\mathcal{K}} \pi_{b,h,c}^k S_{b,h,c}^k, \tag{11}$$

followed by the residual

$$\Delta Y_{b,h,c} = \gamma \sigma_{b,c} S_{b,h,c} \tanh\big(g_{\phi_g}(F_{b,h,c})\big)\,. \tag{12}$$

Because the evidence weights sum to one and each $S^k \in [0,1]$, we have $0 \le S_{b,h,c} \le 1$ and

$$|\Delta Y_{b,h,c}| \le \gamma \sigma_{b,c} S_{b,h,c}. \tag{13}$$

The output layer of g_{ϕ_g} is zero-initialized, so SPE starts from the frozen base forecast. The intermediate prediction is $\widehat{Y}^{\mathrm{SPE}} = \widehat{Y}^0 + \Delta Y$. The residual is evidence-conditioned rather than a convex combination of candidates, and the bound limits correction amplitude without guaranteeing pointwise improvement.

Benefit-aware conservative learning. Candidate benefit is used only as training supervision. Define normalized candidate and base errors

$$\bar{e}^k_{b,h,c} = \left(\frac{C^k_{b,h,c} - Y_{b,h,c}}{\sigma_{b,c} + \epsilon}\right)^2, \qquad \bar{e}^0_{b,h,c} = \left(\frac{\widehat{Y}^0_{b,h,c} - Y_{b,h,c}}{\sigma_{b,c} + \epsilon}\right)^2, \tag{14}$$

and the training-only benefit mask

$$m_{b,h,c} = \mathbb{I}\left[\min_{k \in \mathcal{K}} \bar{e}^k_{b,h,c} + \delta_{\mathrm{ben}} < \bar{e}^0_{b,h,c}\right], \qquad \delta_{\mathrm{ben}} \ge 0. \tag{15}$$

For compactness, let $i = (b,h,c)$, $N = BHC$, $e_i = ((\widehat{Y}^{\mathrm{SPE}}_i - Y_i)/(\sigma_i + \epsilon))^2$, and define the masked mean $\langle z \rangle_w = \sum_i w_i z_i / (\sum_i w_i + \epsilon)$. The SPE objective is

$$\mathcal{L}_{\mathrm{SPE}} = \mathcal{L}_{\mathrm{fit}} + \lambda_{\mathrm{nh}} \mathcal{L}_{\mathrm{nh}} + \lambda_{\mathrm{con}} \mathcal{L}_{\mathrm{con}}, \tag{16a}$$

$$\mathcal{L}_{\mathrm{fit}} = \left\langle (\widehat{Y}^{\mathrm{SPE}} - Y)^2 \right\rangle_m, \tag{16b}$$

$$\mathcal{L}_{\mathrm{nh}} = \left\langle [e - \bar{e}^0 - \delta_{\mathrm{nh}}]_+ \right\rangle_{1-m}, \tag{16c}$$

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{N} \sum_i \frac{|\Delta Y_i|}{\sigma_i + \epsilon}. \tag{16d}$$

The fit term remains in the original scale to match the forecasting objective, whereas the no-harm and conservative terms are normalized so that their thresholds and penalties are comparable across variables with different scales. Empty masked groups contribute zero. The hinge discourages excessive error increases outside the benefit mask but provides no pointwise no-harm guarantee. Because the base forecaster is frozen, the penalty cannot be reduced by degrading the reference forecast. Further training details are provided in Appendix C.3.

3.4 Utility-Controlled Evidence Extrapolation

Historical support does not imply uniform correction utility across prediction positions or temporal regimes. HAEU calibrates residual strength over the forecasting horizon, while CFUS selects model states according to utility on strictly later validation data.

Horizon-Adaptive Evidence Utility. For a fixed SPE checkpoint, let $\mathcal{D}_{\mathrm{cal}}$ denote an earlier calibration subset and define $R^Y_{b,h,c} = Y_{b,h,c} - \widehat{Y}^0_{b,h,c}$. HAEU fits one scalar per prediction position by ridge-regularized least squares:

$$\alpha^{\mathrm{raw}}_h = \frac{\sum_{b \in \mathcal{D}_{\mathrm{cal}}} \sum_c \Delta Y_{b,h,c} R^Y_{b,h,c}}{\sum_{b \in \mathcal{D}_{\mathrm{cal}}} \sum_c (\Delta Y_{b,h,c})^2 + \lambda_\alpha}, \qquad \alpha_h = \mathrm{clip}\left(\alpha^{\mathrm{raw}}_h, 0, \alpha_{\max}\right). \tag{17}$$

Here $\lambda_\alpha \ge 0$ is the ridge coefficient and $\alpha_{\max} > 0$ bounds amplification; our configuration uses $\alpha_{\max} = 1.2$. Each α_h is shared across samples and variables. Because $\alpha_h \ge 0$, HAEU cannot reverse the learned residual, but it may amplify, attenuate, or suppress it. No monotonic horizon profile is imposed. Calibration uses labels from $\mathcal{D}_{\mathrm{cal}}$ only; the derivation and zero-denominator convention are given in Appendix C.4.

Chronological Forward Utility Selection. We divide validation chronologically into $\mathcal{D}_{\mathrm{cal}}$ and a strictly later $\mathcal{D}_{\mathrm{fwd}}$, with disjoint target timestamps. All candidate SPE checkpoints are obtained from training data only; neither validation subset is used to update ϕ. Let n index candidate checkpoints $\Theta_n = (\theta_0, \phi_n)$. For each n, we fit $\alpha^{(n)}$ on $\mathcal{D}_{\mathrm{cal}}$, freeze it, and select

$$n^\star = \arg\min_n \frac{1}{|\mathcal{D}_{\mathrm{fwd}}|HC} \sum_{b \in \mathcal{D}_{\mathrm{fwd}}} \sum_{h,c} \left(\widehat{Y}^0_{b,h,c} + \alpha^{(n)}_h \Delta Y^{(n)}_{b,h,c} - Y_{b,h,c} \right)^2 . \tag{18}$$

Since the base forecaster is fixed, minimizing corrected MSE on $\mathcal{D}_{\mathrm{fwd}}$ is equivalent to maximizing aggregate MSE reduction relative to the base. We retain $\Theta^\star = \Theta_{n^\star}$ and $\alpha^\star = \alpha^{(n^\star)}$; the latter is not refitted on $\mathcal{D}_{\mathrm{fwd}}$. Thus, the chronological roles are separated as $\mathcal{D}_{\mathrm{train}} \to \phi_n$, $\mathcal{D}_{\mathrm{cal}} \to \alpha^{(n)}$, $\mathcal{D}_{\mathrm{fwd}} \to n^\star$, while the test set is reserved for final evaluation. CFUS changes model selection rather than the optimizer or inference architecture and does not guarantee performance under arbitrary future distribution shifts.

Final inference. For a new history, we compute the frozen base forecast, construct SPE evidence using only available observations, and apply Eq. (2) with the selected checkpoint and fixed calibration coefficients. The frozen teacher, auxiliary damage head, training-only benefit mask, and validation labels are not used at inference. Consequently, the complete calibrated correction satisfies

$$\left| \widehat{Y}_{b,h,c} - \widehat{Y}^0_{b,h,c} \right| \le \alpha_{\max} \gamma \sigma_{b,c} S_{b,h,c}. \tag{19}$$

No future target is accessed when producing a test-time forecast.

4 Experiments

We evaluate NEIN-TSF in terms of forecasting accuracy, mechanism behavior, component contribution, and backbone transferability. Additional benchmarks, full mechanism diagnostics, fine-grained ablations, hyperparameter sensitivity, and random-seed stability are reported in the appendix.

4.1 Experimental Setup

Datasets and metrics. Our main evaluation covers seven widely used long-term forecasting benchmarks: ETTh1, ETTh2, ETTm1, ETTm2, Electricity (ECL), Traffic, and Weather, with $H \in \{96, 192, 336, 720\}$. We further evaluate NEIN-TSF on AQShunyi, AQWan, CzeLan, and Temp under the same long-term horizons, and on PEMS03/04/07/08 with $H \in \{12, 24, 48, 96\}$, yielding fifteen benchmarks in total. We report MSE and MAE throughout; lower is better. Additional results are provided in the appendix.

Baselines and implementation. We compare NEIN-TSF with Time-o1 (Wang et al., 2026b), OLinear (Yue et al., 2026), TimeMixer++ (Wang et al., 2025), Fredformer (Piao et al., 2024), iTransformer (Liu et al., 2023a), PatchTST (Nie et al., 2022), and TimesNet (Wu et al., 2022). For reimplemented methods, we use official code and recommended configurations whenever available, with validation-only hyperparameter selection; otherwise, we retain published results under comparable settings. NEIN-TSF-Tuned uses non-exhaustive, validation-guided tuning of standard training and backbone parameters. Unless stated otherwise, results are averaged over three independent random seeds. Full implementation and tuning details are provided in Appendix D.2.

Chronological protocol. All datasets are split chronologically, and the test set is never used for training, calibration, or model selection. For NEIN-TSF, the validation period is further divided into an earlier calibration set $\mathcal{D}_{\mathrm{cal}}$ and a strictly later forward-validation set $\mathcal{D}_{\mathrm{fwd}}$. HAEU coefficients are fitted only on $\mathcal{D}_{\mathrm{cal}}$, whereas CFUS uses $\mathcal{D}_{\mathrm{fwd}}$ solely for checkpoint selection, without refitting the coefficients. Window-wise statistics (μ, σ) are computed only from the observed input, as defined in Section 3.1.

4.2 Overall Forecasting Performance

Table 2 reports forecasting performance averaged over $H \in \{96, 192, 336, 720\}$. Under the shared default configuration, NEIN-TSF outperforms all compared baselines in 11 of the 14 dataset–metric

Table 2: **Long-term forecasting performance** averaged over $H \in \{96, 192, 336, 720\}$. Each entry reports MSE/MAE; lower is better. Best and second-best available results are shown in **bold** and underline, respectively. $\star$ denotes the tuned NEIN-TSF configuration.

Dataset	NEIN-TSF	NEIN-TSF-Tuned*	Time-o1	OLinear	TimeMixer++	Fredformer	iTransformer	PatchTST	TimesNet
ETTh1	0.396/0.415	**0.384/0.407**	0.431/0.429	0.424/0.424	0.419/0.432	0.435/0.426	0.454/0.447	0.469/0.454	0.458/0.450
ETTh2	0.337/0.383	**0.327/0.375**	0.359/0.388	0.367/0.388	0.339/0.380	0.365/0.393	0.383/0.407	0.384/0.405	0.414/0.427
ETTm1	0.360/0.375	**0.348/0.366**	0.379/0.393	0.374/0.377	0.369/0.378	0.384/0.395	0.407/0.410	0.387/0.400	0.400/0.406
ETTm2	0.250/0.308	**0.240/0.300**	0.272/0.317	0.270/0.313	0.269/0.320	0.279/0.324	0.288/0.332	0.281/0.326	0.291/0.333
Weather	0.233/0.256	**0.224/0.249**	0.241/0.280	0.237/0.260	0.226/0.262	0.246/0.272	0.258/0.278	0.259/0.281	0.259/0.287
Electricity	0.156/0.245	**0.150/0.239**	0.170/0.260	0.159/0.248	0.165/0.253	0.176/0.269	0.178/0.270	0.208/0.295	0.192/0.295
Traffic	0.391/0.270	**0.385**/0.263	0.419/0.280	0.451/**0.247**	0.416/0.264	0.433/0.291	0.428/0.282	0.531/0.343	0.620/0.336

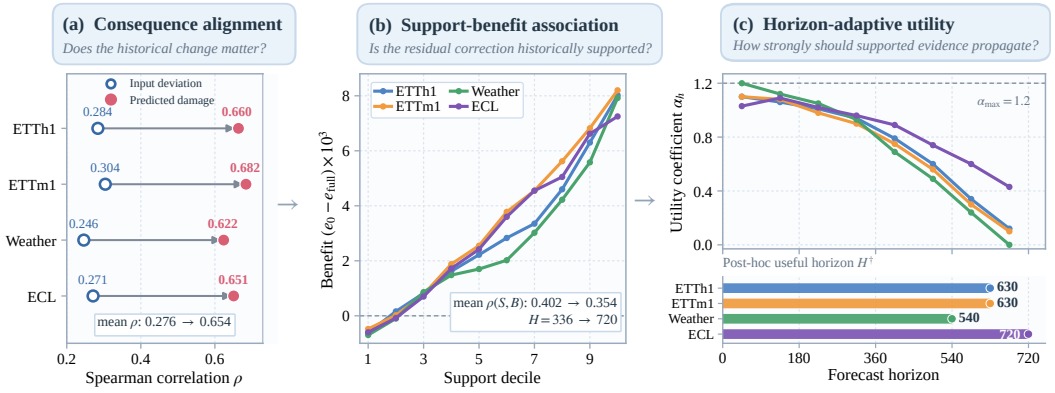


Figure 3: **Unified mechanism audit of NEIN-TSF.** (a) FDRP-predicted damage is more aligned with teacher-derived damage than raw input deviation. (b) SPE correction benefit generally increases with historical support. (c) HAEU learns horizon-dependent correction utility; $H^\dagger$ is a descriptive post-hoc persistence boundary. Full diagnostics are provided in Appendix G.

comparisons. Dataset-specific validation tuning further improves the results, achieving the best overall result in 13 of 14 comparisons and the second-best Traffic MAE. The gains span both ETT sampling regimes, Weather, Electricity, and Traffic, suggesting that the improvement is not confined to a particular dataset family or metric.

4.3 Mechanism Analysis: Consequence, Support, and Utility

We jointly audit the three decisions underlying NEIN-TSF: which historical changes matter for forecasting, when historical evidence supports residual refinement, and how correction utility varies across forecasting distance and chronological time.

Across ETTh1, ETTm1, Weather, and ECL, the mean Spearman correlation with teacher-derived damage increases from 0.276 for input deviation to 0.654 for FDRP prediction (Figure 3a). SPE support is likewise informative about correction utility: the mean support–benefit correlation remains positive from $H = 336$ to $H = 720$ ($0.402 \rightarrow 0.354$; Figure 3b). Finally, HAEU learns non-uniform horizon-wise coefficients, with descriptive $H^\dagger$ values of 630, 630, 540, and 720 for ETTh1, ETTm1, Weather, and ECL, respectively (Figure 3c). Full consequence-gap, harm-rate, and chronological analyses are deferred to Appendix G.

4.4 Ablation and Additional Analyses

Contribution of individual stages. Table 3 shows a consistent progressive effect. FDRP and SPE each improve the vanilla backbone, their combination improves all four benchmarks further, and adding either HAEU or CFUS yields additional reductions. Full NEIN-TSF performs best throughout. Averaged over the four datasets, it reduces MSE and MAE by 5.48% and 4.03%, respectively, relative to the vanilla backbone. Fine-grained component studies are reported in Appendix H.

Table 3: **Core ablation of NEIN-TSF.** Results are averaged over $H \in \{96, 192, 336, 720\}$. Dataset entries report MSE/MAE. The last two columns report the mean of per-dataset relative reductions with respect to the vanilla backbone.

Variant	Components				Representative Benchmarks				Avg. Rel. Gain (%)	
	F	S	H	C	ETTh1	ETTm1	Weather	ECL	MSE	MAE
Vanilla backbone	–	–	–	–	0.416/0.429	0.374/0.388	0.247/0.269	0.169/0.257	0.00	0.00
Vanilla + FDRP	✓	–	–	–	0.407/0.422	0.368/0.382	0.241/0.264	0.164/0.252	2.29	1.75
Vanilla + SPE	–	✓	–	–	0.411/0.425	0.371/0.385	0.243/0.263	0.165/0.251	1.50	1.57
FDRP + SPE	✓	✓	–	–	0.402/0.418	0.364/0.378	0.237/0.259	0.160/0.248	3.85	3.09
FDRP + SPE + HAEU	✓	✓	✓	–	0.399/0.417	0.362/0.377	0.235/0.258	0.158/0.247	4.67	3.40
FDRP + SPE + CFUS	✓	✓	–	✓	0.400/0.416	0.363/0.376	0.236/0.257	0.159/0.246	4.29	3.72
Full NEIN-TSF	✓	✓	✓	✓	**0.396/0.415**	**0.360/0.375**	**0.233/0.256**	**0.156/0.245**	**5.48**	**4.03**

Table 4: **Backbone transferability of NEIN-TSF.** Each cell reports Base → +NEIN-TSF in MSE/MAE, averaged over $H \in \{96, 192, 336, 720\}$; lower is better.

Dataset	iTransformer	SAMformer	Dataset	iTransformer	SAMformer
ETTh1	0.454/0.447 → **0.435/0.427**	0.410/0.424 → **0.397/0.412**	ETTm2	0.288/0.332 → **0.275/0.316**	0.269/0.327 → **0.258/0.317**
ETTh2	0.383/0.407 → **0.373/0.394**	0.344/0.392 → **0.333/0.383**	Electricity	0.178/0.270 → **0.167/0.258**	0.181/0.275 → **0.162/0.258**
ETTm1	0.407/0.410 → **0.383/0.383**	0.373/0.388 → **0.355/0.375**	Weather	0.258/0.278 → **0.248/0.265**	0.261/0.293 → **0.229/0.265**

Mean relative MSE/MAE reduction: iTransformer **4.54%/4.70%**; SAMformer **6.34%/4.55%**.

4.5 Transferability Across Forecasting Backbones

We further examine whether NEIN-TSF is tied to a particular forecasting architecture by integrating it with two distinct backbones, iTransformer (Liu et al., 2023a) and SAMformer (Ilbert et al., 2024), while keeping their original configurations unchanged. As shown in Table 4, NEIN-TSF improves the horizon-averaged MSE and MAE on all six datasets for both backbones. Across datasets, the mean relative MSE/MAE reductions reach $4.54\%/4.70\%$ for iTransformer and $6.34\%/4.55\%$ for SAMformer. These consistent gains suggest that the proposed consequence learning, supported refinement, and utility control are transferable across different forecasting architectures rather than being specific to one backbone. Full horizon-wise results are provided in Appendix F.

5 Limitations and Future Work

NEIN-TSF has several limitations. FDRP relies on teacher-derived forecast damage as a surrogate for predictive consequence, SPE uses a predefined historical evidence bank that may not capture all predictive structures, and HAEU/CFUS depend on chronologically held-out validation data whose utility may degrade under unseen distribution shifts. Future work may explore learnable evidence construction, uncertainty-aware correction, and adaptive utility estimation under evolving temporal regimes.

6 Conclusion

We presented NEIN-TSF, a forecasting framework that regulates historical dynamics through forecasting consequence, historical support, and correction utility. FDRP learns consequence-aware representations from forecast-damage ranking, SPE performs bounded refinement using complementary historical evidence, and HAEU with CFUS regulates correction utility across forecasting distance and chronological time. Experiments on diverse real-world benchmarks show broad improvements over strong baselines, while mechanism analyses support the intended relationships between forecast consequence, historical support, and correction utility. Overall, NEIN-TSF suggests that reliable forecasting should consider not only what occurred in history, but also whether and how strongly it remains useful for the future.

AI USE STATEMENT

Generative AI tools were used solely to assist with manuscript translation and language polishing. All research ideas, methodological design, experiments, data analysis, figures, tables, and reported results were produced and verified manually by the authors. The authors reviewed all AI-assisted text and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the complete formulation of NEIN-TSF in Section 3 and detailed dataset, implementation, and hyperparameter settings in Appendix D. Additional benchmark results, mechanism diagnostics, backbone transfer experiments, ablation studies, hyperparameter sensitivity, and random-seed analyses are also reported in the appendix. These details are provided to facilitate independent reproduction of the reported results.

REFERENCES

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *ICLR 2024: The Twelfth International Conference on Learning Representations.*, 2024.

Hao Cheng, Qingsong Wen, Yang Liu, and Liang Sun. Robusttsf: Towards theory and design of robust time series forecasting with anomalies. In *International Conference on Learning Representations*, volume 2024, pp. 5787–5813, 2024.

Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36:29996–30025, 2023.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2505.04163*, 2025.

Romain Ilbert, Ambroise Odonnat, Vasilii Feofanov, Aladin Virmaux, Giuseppe Paolo, Themis Palpanas, and Ievgen Redko. Samformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention. In *ICML 2024-Proceedings of the 41st International Conference on Machine Learning*, 2024.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 17868–17876, 2025.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Mengfan Liang, Shixiang Jia, Yepeng Liu, Xiaofeng Zhang, Hua Wang, and Yujuan Sun. Tritracknet: A dual-channel time series forecasting model with multi-path interaction and perturbation optimization. *Neurocomputing*, pp. 132519, 2025.

Haoxin Liu, Harshavardhan Kamarthi, Lingkai Kong, Zhiyuan Zhao, Chao Zhang, and B Aditya Prakash. Time-series forecasting for out-of-distribution generalization using invariant learning. *arXiv preprint arXiv:2406.09130*, 2024a.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. *Advances in Neural Information Processing Systems*, 37:2766–2786, 2024b.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting. 2024c.

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

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2023a.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. *Advances in neural information processing systems*, 36:12271–12290, 2023b.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*, 2024d.

Jinghao Lu, Fan Zhang, Xiaofeng Zhang, Yujuan Sun, and Hua Wang. Mcnr: Multiscale feature-based latent data component extraction linear regression model. *Expert Systems with Applications*, 292:128634, 2025.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Xihao Piao, Zheng Chen, Taichi Murayama, Yasuko Matsubara, and Yasushi Sakurai. Fredformer: Frequency debiased transformer for time series forecasting. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2400–2410, 2024.

Amin Shabani, Amir Abdi, Lili Meng, and Tristan Sylvain. Scaleformer: Iterative multi-scale refining transformers for time series forecasting. *arXiv preprint arXiv:2206.04038*, 2022.

Caixia Wang, Fan Zhang, Xiaofeng Zhang, and Hua Wang. Dtfnet: A dual-modal time-frequency fusion network for non-stationary time series modeling. *Knowledge-Based Systems*, pp. 116022, 2026a.

Hao Wang, Licheng Pan, Zhichao Chen, Xu Chen, Qingyang Dai, Lei Wang, Haoxuan Li, and Zhouchen Lin. Time-o1: Time-series forecasting needs transformed label alignment. *Advances in Neural Information Processing Systems*, 38:8660–8690, 2026b.

Hua Wang, Jinghao Lu, and Fan Zhang. Eeo-tfv: Escape-explore optimizer for web-scale time-series forecasting and vision analysis. In *Proceedings of the ACM Web Conference 2026*, pp. 7271–7282, 2026c.

Hua Wang, Jinghao Lu, and Fan Zhang. Idealtsf: Can non-ideal data contribute to enhancing the performance of time series forecasting models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 26224–26232, 2026d.

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

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, volume 2025, pp. 16980–17016, 2025.

Qingsong Wen, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan, et al. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. *Advances in Neural Information Processing Systems*, 36:69949–69980, 2023.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2022.

Wenzhen Yue, Yong Liu, Hao Wang, Haoxuan Li, Xianghua Ying, Ruohao Guo, Bowei Xing, and Ji Shi. Olinear: A linear model for time series forecasting in orthogonally transformed domain. *Advances in Neural Information Processing Systems*, 38:60573–60632, 2026.

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

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

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

A Empirical Motivation Details

The motivation experiment in Figure 1 examines whether the magnitude of a historical irregularity is sufficient to determine its forecasting value. We conduct the analysis on the hourly Electricity (ECL) dataset with an input length of $L = 96$ and focus on local irregularities occurring within the final six hours before the forecast origin.

Event identification and matching. Candidate events are detected from historical observations using a robust deviation score relative to the same hour in the preceding seven days. To reduce confounding from periodic position and local geometry, events are matched using historical-only attributes, including irregularity magnitude, duration, direction, pre-event trend, and hour of day. No future target is used for event detection or matching.

Counterfactual attenuation. For each matched event e, the irregular interval is attenuated toward its same-hour seasonal historical reference while the remainder of the input window is unchanged. Let X_e and X_e^{att} denote the original and attenuated histories. Using a forecasting model independent of the proposed correction mechanism, we define the event-level forecast consequence at horizon H as

$$C_e(H) = \mathcal{L}_H\big(f(X_e^{\mathrm{att}}), Y\big) - \mathcal{L}_H(f(X_e), Y)\,. \tag{20}$$

Thus, $C_e(H) > 0$ indicates that attenuation increases forecast error under the evaluated forecaster, whereas $C_e(H) < 0$ indicates that attenuation improves the forecast. This quantity is used only as an empirical diagnostic and is never an inference-time signal in NEIN-TSF.

B Additional Related Work

Pretraining and non-stationary dynamics. RevIN (Kim et al., 2021) mitigates distribution shifts through reversible instance normalization, while Non-stationary Transformers (Liu et al., 2022) combine series stationarization with de-stationary attention. SimMTM (Dong et al., 2023), Koopa (Liu et al., 2023b), MOMENT (Goswami et al., 2024), and Timer (Liu et al., 2024d) learn transferable temporal representations through reconstruction, dynamical modeling, large-scale pre-training, or generative prediction. FDRP is complementary to these approaches: it explicitly learns the relative forecasting consequences of matched local counterfactuals using teacher-derived forecast damage, rather than relying on input-space deviation alone.

Table 5: **Configuration of the empirical motivation analysis.** Only settings fixed in the current experimental protocol are reported.

Item	Setting
Dataset	Electricity (ECL), hourly
Input length	$L = 96$
Event location	final 6 hours before forecast origin
Seasonal reference	same hour in previous 7 days
Detection statistic	robust z-score
Matching attributes	magnitude, duration, direction, trend, hour of day
Evaluated horizons	$H \in \{96, 192, 336, 720\}$

Historical structure and retrieval-augmented forecasting. FEDformer (Zhou et al., 2022), TimesNet (Wu et al., 2022), Crossformer (Zhang & Yan, 2023), and Pathformer (Chen et al., 2024) exploit decomposition, frequency structure, cross-dimensional interactions, or multiscale temporal pathways. Retrieval-based methods further use historical precedents: RATD (Liu et al., 2024b) retrieves relevant series for diffusion-based forecasting, while RAFT (Han et al., 2025) augments prediction with similar historical patterns and their subsequent trajectories. SPE differs in that it evaluates complementary historical evidence and uses that support to constrain bounded residual refinement, rather than assuming that a plausible historical continuation is necessarily beneficial.

Horizon adaptation and temporal stability. Scaleformer (Shabani et al., 2022) and TimeMixer (Wang et al., 2024) exploit multiscale information, whereas OneNet (Wen et al., 2023), FOIL (Liu et al., 2024a), and TAFAS (Kim et al., 2025) address temporal distribution shifts through online adaptation, invariant learning, and test-time adaptation, respectively. HAEU instead calibrates correction strength across prediction positions, while CFUS uses chronologically ordered forward validation for model-state selection, jointly regulating correction utility across forecasting distance and chronological time.

C Additional Methodology Details

This section provides the complete construction and training definitions corresponding to Section 3. All quantities retain the same meanings and scales as in the main text.

C.1 Counterfactual Construction and Damage Supervision

Local contextual reference. For $r \geq 1$, define

$$\mathcal{N}_t = \{j : 1 \leq j \leq L,\ 0 < |j - t| \leq r\}, \qquad R_{b,t,c} = \frac{1}{|\mathcal{N}_t|} \sum_{j \in \mathcal{N}_t} X_{b,j,c}. \tag{21}$$

Thus, $\mathbf{R} \in \mathbb{R}^{B \times L \times C}$ is computed entirely from the observed history. The neighborhood excludes the center position and contains only valid historical indices.

Perturbation operators. The mask $\mathbf{M}$ selects the location, variable, and width of a local perturbation. The complete operators used in Eq. (3) are

$$\mathbf{P}_q(\lambda) = \begin{cases} \lambda \boldsymbol{\sigma} \xi, & q = \text{jump}, \\ \lambda \boldsymbol{\sigma} \odot \boldsymbol{\zeta}, & q = \text{noise}, \\ \lambda(\mathbf{R} - \mathbf{X}), & q = \text{erase}, \\ \lambda \boldsymbol{\sigma} \xi \boldsymbol{\omega}, & q = \text{drift}. \end{cases} \tag{22}$$

Here $\xi \in \{-1, +1\}$ denotes the shared direction, $\boldsymbol{\zeta}$ is a zero-mean random realization, and $\boldsymbol{\omega}$ increases monotonically from zero to one over the selected interval. Window statistics are computed before perturbation. Mild and severe variants share $\mathbf{M}$, q, ξ, $\boldsymbol{\zeta}$, and $\boldsymbol{\omega}$ when applicable, differing only in intensity. For erasure, $0 < \lambda^{\text{m}} < \lambda^{\text{s}} \leq 1$; the other perturbation strengths are positive. Matching controls construction factors but does not assume that larger perturbation intensity must induce larger teacher damage.

Context-residual representation. For $a \in \{\mathrm{m}, \mathrm{s}\}$, we construct

$$\mathbf{Z}^a = \mathrm{Concat}\Big(\frac{\mathbf{X}^a - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \frac{\mathbf{R} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \frac{\mathbf{X}^a - \mathbf{R}}{\boldsymbol{\sigma}}, \frac{|\mathbf{X}^a - \mathbf{R}|}{\boldsymbol{\sigma}}, \frac{\Delta_t \mathbf{X}^a - \Delta_t \mathbf{R}}{\boldsymbol{\sigma}}\Big). \tag{23}$$

These terms encode the counterfactual observation, local reference, signed residual, residual magnitude, and mismatch in temporal changes. Δ_t denotes a first-order temporal-difference feature; the difference features are aligned to the common temporal length before concatenation.

Let $\mathbf{H}_\theta^a$ be the shared representation exposed by f_θ for $\mathbf{X}^a$. The auxiliary head is

$$\widehat{\mathbf{d}}^a = q_\psi(\mathbf{H}_\theta^a, \mathbf{Z}^a) \in \mathbb{R}^{B \times H}. \tag{24}$$

The head is trained jointly with the backbone using Eq. (8) and removed after pretraining. The frozen teacher is used only to generate supervision. Since the damage in Eq. (5) is clipped at zero, improvements relative to the clean teacher do not receive negative damage labels. If $N_{\mathrm{rank}} = 0$, the ranking contribution is defined to be zero.

C.2 Complete Historical Evidence Bank

The four candidates and support scores are computed from the observed history and frozen base forecast only. All temperature parameters are positive. A high support score indicates agreement with the corresponding historical rule rather than a calibrated probability of future correctness.

Historical retrieval. The query is the last ℓ_r normalized historical observations. A candidate context ending at j must be separated from the query by the prescribed gap g_r. We select

$$\begin{aligned} j_b^\star &= \arg\min_{j \in \mathcal{J}_b} \frac{1}{\ell_r C} \left\| \widetilde{\mathbf{X}}_{b, L-\ell_r+1:L, :} - \widetilde{\mathbf{X}}_{b, j-\ell_r+1:j, :} \right\|_F^2, \\ \mathcal{J}_b &= \{ j : \ell_r \le j \le L - \ell_r - g_r \}. \end{aligned} \tag{25}$$

Let d_b^{ret} denote the minimum distance. Whenever the required observed continuation is available, i.e., $j_b^\star + h \le L$, define

$$C_{b,h,c}^{\mathrm{ret}} = X_{b, j_b^\star + h, c}, \qquad S_{b,h,c}^{\mathrm{ret}} = \exp(-d_b^{\mathrm{ret}} / \tau_r). \tag{26}$$

If $\mathcal{J}_b$ is empty or the continuation is unavailable, we set $C_{b,h,c}^{\mathrm{ret}} = \widehat{Y}_{b,h,c}^0$ and $S_{b,h,c}^{\mathrm{ret}} = 0$. The matched historical context is separated from the query under the stated index restriction, and its continuation is always drawn from already observed history.

Periodic evidence. For a nonempty candidate set $\mathcal{P} \subset \{1, \dots, L-1\}$, define

$$\begin{aligned} A_b(p) &= \frac{1}{C(L-p)} \sum_{c=1}^{C} \sum_{t=p+1}^{L} \widetilde{X}_{b,t,c} \widetilde{X}_{b,t-p,c}, \\ p_b^\star &= \arg\max_{p \in \mathcal{P}} A_b(p). \end{aligned} \tag{27}$$

The candidate repeats the selected phase:

$$\begin{aligned} C_{b,h,c}^{\mathrm{per}} &= X_{b, L-p_b^\star+1+\mathrm{mod}(h-1, p_b^\star), c}, \\ S_{b,h,c}^{\mathrm{per}} &= \mathrm{clip}(A_b(p_b^\star), 0, 1). \end{aligned} \tag{28}$$

The lagged-product statistic uses full-window standardized values and is not, in general, a Pearson correlation bounded by one. The explicit clipping is therefore part of the support definition. A large value indicates historical phase agreement under this statistic rather than guaranteed future recurrence.

Local-trend evidence. For the last $\ell_t \geq 2$ observations of each variable, let

$$\bar{\imath} = \frac{\ell_t + 1}{2}, \qquad \bar{X}_{b,c} = \frac{1}{\ell_t} \sum_{i=1}^{\ell_t} X_{b,L-\ell_t+i,c}. \tag{29}$$

The least-squares slope and intercept are

$$\begin{aligned} \beta_{b,c} &= \frac{\sum_{i=1}^{\ell_t} (i - \bar{\imath})(X_{b,L-\ell_t+i,c} - \bar{X}_{b,c})}{\sum_{i=1}^{\ell_t} (i - \bar{\imath})^2}, \\ a_{b,c} &= \bar{X}_{b,c} - \beta_{b,c}\bar{\imath}. \end{aligned} \tag{30}$$

The normalized fitting error is

$$r^{\mathrm{tr}}_{b,c} = \frac{1}{\ell_t} \sum_{i=1}^{\ell_t} \left(\frac{X_{b,L-\ell_t+i,c} - a_{b,c} - \beta_{b,c} i}{\sigma_{b,c} + \epsilon} \right)^2. \tag{31}$$

The trend candidate and support are

$$C^{\mathrm{tr}}_{b,h,c} = a_{b,c} + \beta_{b,c}(\ell_t + h), \qquad S^{\mathrm{tr}}_{b,h,c} = \exp(-r^{\mathrm{tr}}_{b,c}/\tau_t). \tag{32}$$

Historical fit quality alone does not guarantee continuation of a linear trend; large extrapolations are subsequently constrained by the bounded SPE residual.

Persistence evidence. Persistence repeats the last observation:

$$C^{\mathrm{pers}}_{b,h,c} = X_{b,L,c}. \tag{33}$$

For $\ell_p \geq 2$, recent normalized variation and support are

$$\begin{aligned} v^{\mathrm{pers}}_{b,c} &= \frac{1}{\ell_p - 1} \sum_{i=2}^{\ell_p} \frac{|X_{b,L-\ell_p+i,c} - X_{b,L-\ell_p+i-1,c}|}{\sigma_{b,c} + \epsilon}, \\ S^{\mathrm{pers}}_{b,h,c} &= \exp(-v^{\mathrm{pers}}_{b,c}/\tau_p). \end{aligned} \tag{34}$$

Smaller recent variation yields stronger support for persistence without implying that future observations must remain constant.

C.3 Evidence Features and Conservative Training

For each source k, define the normalized deviation from the frozen base:

$$D^k_{b,h,c} = \frac{C^k_{b,h,c} - \widehat{Y}^0_{b,h,c}}{\sigma_{b,c} + \epsilon}. \tag{35}$$

Let $\widehat{Y}^0_{b,0,c} = X_{b,L,c}$. The normalized base-forecast motion and prediction position are

$$U^0_{b,h,c} = \frac{\widehat{Y}^0_{b,h,c} - \widehat{Y}^0_{b,h-1,c}}{\sigma_{b,c} + \epsilon}, \qquad \eta_h = h/H. \tag{36}$$

Together with the four candidate deviations and support scores, these quantities form the evidence feature in Eq. (10). The weighting network produces

$$\boldsymbol{\pi}_{b,h,c} = \mathrm{softmax}(a_{\phi_a}(\mathbf{F}_{b,h,c})), \qquad S_{b,h,c} = \sum_{k \in \mathcal{K}} \pi^k_{b,h,c} S^k_{b,h,c}. \tag{37}$$

Because the weights are nonnegative and sum to one, aggregate support lies in $[0, 1]$. The correction network receives the same features and produces the bounded residual in Eq. (12).

Expanded benefit-aware objective. The target-dependent benefit mask is constructed during training only and is not an inference-time gate. The expanded loss terms are

$$\mathcal{L}_{\text{fit}} = \frac{\sum_{b,h,c} m_{b,h,c}(\widehat{Y}^{\text{SPE}}_{b,h,c} - Y_{b,h,c})^2}{\sum_{b,h,c} m_{b,h,c} + \epsilon}, \tag{38}$$

$$\begin{aligned}\mathcal{L}_{\text{nh}} = &\frac{1}{\sum_{b,h,c}(1 - m_{b,h,c}) + \epsilon} \sum_{b,h,c} (1 - m_{b,h,c}) \\ &\times \left[\left(\frac{\widehat{Y}^{\text{SPE}}_{b,h,c} - Y_{b,h,c}}{\sigma_{b,c} + \epsilon} \right)^2 - \bar{e}^0_{b,h,c} - \delta_{\text{nh}} \right]_+ ,\end{aligned} \tag{39}$$

and

$$\mathcal{L}_{\text{con}} = \frac{1}{BHC} \sum_{b,h,c} \frac{|\Delta Y_{b,h,c}|}{\sigma_{b,c} + \epsilon}. \tag{40}$$

These are the components summarized by Eq. (16a). The original-scale fit term and normalized penalties serve different roles. The thresholds δ_{ben} and δ_{nh} respectively define a strict candidate improvement criterion and permit a small error increase outside that group. Empty masked groups contribute zero. The soft no-harm term does not imply a pointwise guarantee, and freezing the base forecaster prevents the penalty from being reduced by degrading the reference prediction.

C.4 Calibration, Chronological Selection, and Inference

Horizon-wise ridge calibration. For a fixed SPE checkpoint, define $R^Y_{b,h,c} = Y_{b,h,c} - \widehat{Y}^0_{b,h,c}$. The coefficient at prediction position h minimizes

$$\min_{\alpha \in \mathbb{R}} \sum_{b \in \mathcal{D}_{\text{cal}}} \sum_{c=1}^{C} (R^Y_{b,h,c} - \alpha \Delta Y_{b,h,c})^2 + \lambda_\alpha \alpha^2. \tag{41}$$

Differentiating gives

$$\left(\sum_{b,c} (\Delta Y_{b,h,c})^2 + \lambda_\alpha \right) \alpha = \sum_{b,c} \Delta Y_{b,h,c} R^Y_{b,h,c}, \tag{42}$$

where the sums are over calibration samples and variables. This yields the closed form in Eq. (17). Since the objective is a scalar convex quadratic, clipping the unconstrained minimizer to $[0, \alpha_{\max}]$ yields the box-constrained minimizer whenever the denominator is positive. If the denominator is zero, the coefficient is set to zero. We use $\alpha_{\max} = 1.2$ and $\lambda_\alpha = 10^{-4}$.

The calibrated prediction is

$$\widehat{Y}_{b,h,c} = \widehat{Y}^0_{b,h,c} + \alpha_h \Delta Y_{b,h,c}. \tag{43}$$

The nonnegative coefficient can suppress, attenuate, retain, or moderately amplify the SPE residual but cannot reverse its direction. No monotonicity is imposed across prediction positions.

Chronological validation. Let $\text{Time}_Y(\mathcal{D})$ denote the target timestamps of a partition. CFUS requires

$$\max \text{Time}_Y(\mathcal{D}_{\text{cal}}) < \min \text{Time}_Y(\mathcal{D}_{\text{fwd}}). \tag{44}$$

Training targets precede validation targets, and final test targets remain held out. Historical context may overlap across neighboring forecast windows only when it consists of observations available at the corresponding forecast origin; target leakage across partitions is not permitted.

All candidate SPE checkpoints are obtained from training data only. For each candidate $\Theta_n = (\theta_0, \phi_n)$, HAEU coefficients are fitted on $\mathcal{D}_{\text{cal}}$ and held fixed when evaluating corrected MSE on $\mathcal{D}_{\text{fwd}}$. CFUS selects $n^\star$ using Eq. (18) and retains $\Theta^\star = \Theta_{n^\star}$ together with $\boldsymbol{\alpha}^\star = \boldsymbol{\alpha}^{(n^\star)}$. The forward validation targets are used only for checkpoint selection; the selected calibration coefficients are not refitted on them.

Table 6: **Summary of the forecasting datasets.** Dim. denotes the number of variables and #Steps the number of raw temporal observations. H lists the evaluated prediction horizons.

Task	Dataset	Scenario	Dim.	#Steps	Freq.	Prediction horizons H	Reported
Long-term	ETTm1	Power transformer	7	69,680	15 min	$\{96, 192, 336, 720\}$	Main
	ETTm2	Power transformer	7	69,680	15 min	$\{96, 192, 336, 720\}$	Main
	ETTh1	Power transformer	7	17,420	1 hour	$\{96, 192, 336, 720\}$	Main
	ETTh2	Power transformer	7	17,420	1 hour	$\{96, 192, 336, 720\}$	Main
	Weather	Meteorology	21	52,696	10 min	$\{96, 192, 336, 720\}$	Main
	Electricity	Electricity	321	26,304	1 hour	$\{96, 192, 336, 720\}$	Main
	Traffic	Road occupancy	862	17,544	1 hour	$\{96, 192, 336, 720\}$	Main
Short-term	PEMS03	Traffic flow	358	26,208	5 min	$\{12, 24, 48, 96\}$	Appx.
	PEMS04	Traffic flow	307	16,992	5 min	$\{12, 24, 48, 96\}$	Appx.
	PEMS07	Traffic flow	883	28,224	5 min	$\{12, 24, 48, 96\}$	Appx.
	PEMS08	Traffic flow	170	17,856	5 min	$\{12, 24, 48, 96\}$	Appx.
Additional	AQShunyi	Air quality	11	35,064	1 hour	$\{96, 192, 336, 720\}$	Appx.
	AQWan	Air quality	11	35,064	1 hour	$\{96, 192, 336, 720\}$	Appx.
	CzeLan	Sap flow	11	19,934	30 min	$\{96, 192, 336, 720\}$	Appx.
	Temp	Temperature	184	11,688	3 hours	$\{96, 192, 336, 720\}$	Appx.

Final inference. For a new input $\mathbf{X}$, the selected model computes $\widehat{\mathbf{Y}}^0 = f_{\theta_0}(\mathbf{X})$, constructs the four evidence sources from available history, and produces $\Delta\mathbf{Y}$ with $\phi_{n^\star}$. The final prediction follows Eq. (2). The teacher, auxiliary damage head, benefit mask, training targets, and validation labels are not required at inference.

D Dataset and Implementation Details

D.1 Datasets

We evaluate NEIN-TSF on fifteen real-world multivariate time-series benchmarks spanning industrial sensing, meteorology, electricity demand, transportation, air quality, ecological monitoring, and temperature forecasting. The seven main long-term benchmarks are ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity (ECL), and Traffic. We further evaluate short-term traffic forecasting on PEMS03/04/07/08 and long-term forecasting on AQShunyi, AQWan, CzeLan, and Temp. All datasets are partitioned chronologically following their standard forecasting protocols.

D.2 Implementation and Hyperparameters

All dataset partitions preserve temporal order, and all model selection and hyperparameter tuning use validation data only. Unless otherwise stated, NEIN-TSF adopts the shared default configuration summarized in Table 7. The standard NEIN-TSF results use this common configuration across datasets, whereas NEIN-TSF-Tuned selects dataset-specific configurations from a fixed, predefined search space using validation data only. For long-term forecasting, we evaluate $H \in \{96, 192, 336, 720\}$; for PEMS03/04/07/08, we use $H \in \{12, 24, 48, 96\}$. HAEU coefficients are calibrated on $\mathcal{D}_{\text{cal}}$, while the strictly later $\mathcal{D}_{\text{fwd}}$ is used only for CFUS checkpoint selection. The two subsets contain 512 forecasting windows each in the default configuration. Unless otherwise stated, reported multi-seed results are averaged over three independent runs with seeds $\{2022, 2023, 2024\}$.

Base backbone. Unless otherwise stated, NEIN-TSF uses a lightweight cross-variable self-attention forecaster as its base backbone. It constructs Q, K, and V from the historical sequence of each variable, applies scaled dot-product attention across variables, adds the attention output back to the input through a residual connection, and uses a linear prediction head to produce the H-step forecast. The attention hidden dimension is set to 16.

Table 7: **Default experimental configuration of NEIN-TSF.** The standard model uses the shared configuration below across datasets; dataset-specific tuned variants are selected using validation data only.

Optimization		Training / Forecasting		Calibration / Utility	
Parameter	**Setting**	**Parameter**	**Setting**	**Parameter**	**Setting**
Learning rate	3×10^{-4}	Batch size	128	Calibration size	512
Weight decay	1×10^{-4}	Feature batch size	128	Selection size	512
Maximum epochs	160	Random seeds	$\{2022, 2023, 2024\}$	HAEU range	$\alpha_h \in [0, 1.2]$
Early-stop patience	45	Attention hidden dim.	16	Ridge coefficient λ_α	10^{-4}
Early stopping	enabled	Long-term horizons	$\{96, 192, 336, 720\}$	$\lvert\mathcal{D}_{\mathrm{cal}}\rvert : \lvert\mathcal{D}_{\mathrm{fwd}}\rvert$	1:1
λ_{rank}	0.10	PEMS horizons	$\{12, 24, 48, 96\}$	Correction scale γ	0.30
λ_{nh}	0.020	Damage-gap threshold δ_d	10^{-5}	Benefit threshold δ_{ben}	0
λ_{con}	0.010	No-harm margin δ_{nh}	0.02		

The method-specific defaults are $\lambda_{\mathrm{rank}} = 0.10$, $\lambda_{\mathrm{nh}} = 0.020$, $\lambda_{\mathrm{con}} = 0.010$, $\gamma = 0.30$, $\delta_d = 10^{-5}$, $\delta_{\mathrm{ben}} = 0$, and $\delta_{\mathrm{nh}} = 0.02$. Accordingly, the $1.0\times$ settings in the hyperparameter-sensitivity analysis correspond to the relevant default values reported above.

Tuning protocol. NEIN-TSF-Tuned performs non-exhaustive, validation-guided hyperparameter tuning for each dataset. The considered parameters are learning rate, maximum training epochs, batch size, hidden dimension, and lookback length. Their candidate ranges are learning rate $\in \{10^{-5}, 3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$, maximum epochs $\in \{50, 80, 120, 160, 200, 240, 300\}$, batch size $\in \{8, 16, 32, 64, 128, 256\}$, hidden dimension $\in \{8, 16, 32, 64, 128, 256, 512\}$, and lookback length $L \in \{48, 96, 192, 336, 512, 720, 1024\}$. Only a subset of candidate configurations is evaluated for each dataset; the full Cartesian product is not exhaustively searched. Configurations are selected based on validation performance, and the test set is never used for hyperparameter selection. The FDRP, SPE, HAEU, and CFUS-specific hyperparameters remain fixed at their shared default values.

Evidence and counterfactual settings. The evidence-construction settings are shared across datasets rather than tuned individually. Retrieval uses a context length of $\ell_r = 32$ and distance scale $\tau_r = 0.75$, while the local-trend window is $\ell_t = 24$. Periodic evidence considers the fixed candidate-lag set $\mathcal{P} = \{24, 48, 96, 168\}$. For FDRP, the mild perturbation intensity λ_m is sampled uniformly from $[0.10, 0.30]$; its severe counterpart follows the matched construction described in Appendix C.1 with a larger perturbation intensity. Other support-normalization constants are fixed globally and are not included in dataset-specific tuning.

Configuration protocol. The shared default configuration limits dataset-specific tuning and provides a consistent comparison setting. HAEU uses $\lambda_\alpha = 10^{-4}$ and constrains α_h to $[0, 1.2]$, allowing a correction to be suppressed, attenuated, retained, or moderately amplified without reversing its direction. The calibration and forward-selection subsets contain 512 windows each and are strictly ordered chronologically, preserving the separation between coefficient fitting and checkpoint selection.

E Additional Real-World Benchmarks

Tables 8 and 9 extend the evaluation to eight additional real-world benchmarks. NEIN-TSF achieves the best MSE and MAE on all four PEMS benchmarks and on all four additional environmental and temperature datasets. Together with the seven main-text benchmarks, these experiments broaden the evaluation to fifteen real-world time-series datasets spanning long- and short-term forecasting settings.

Table 8: **Average short-term traffic forecasting results on PEMS benchmarks.** Results are averaged over $H \in \{12, 24, 48, 96\}$ and reported in MSE/MAE. Best results are shown in **bold** and second-best results are underlined; lower is better.

Dataset	NEIN-TSF	Twins	iTransformer	RLinear	PatchTST	Crossformer	TiDE	TimesNet
PEMS03	**0.103/0.209**	0.109/0.219	0.116/0.226	0.495/0.472	0.180/0.304	0.169/0.272	0.376/0.329	0.159/0.262
PEMS04	**0.106/0.210**	0.111/0.219	0.121/0.232	0.526/0.491	0.195/0.307	0.209/0.317	0.353/0.420	0.130/0.241
PEMS07	**0.091/0.186**	0.094/0.196	0.100/0.204	0.505/0.478	0.211/0.303	0.235/0.315	0.380/0.440	0.125/0.226
PEMS08	**0.126/0.214**	0.133/0.222	0.151/0.234	0.529/0.493	0.280/0.321	0.268/0.307	0.441/0.464	0.193/0.271

Table 9: **Average long-term forecasting results on additional real-world benchmarks.** Results are reported in MSE/MAE. Best results are shown in **bold** and second-best results are underlined; lower is better.

Dataset	NEIN-TSF	TimeBase	SparseTSF	TimeMixer	FITS	iTransformer	DLinear	PatchTST	TimesNet
AQShunyi	**0.673/0.495**	0.675/0.506	0.760/0.546	0.736/0.536	0.763/0.548	0.723/0.520	0.757/0.572	0.691/0.503	0.788/0.559
AQWan	**0.775/0.487**	0.779/0.499	0.826/0.526	0.822/0.521	0.813/0.519	0.843/0.538	0.883/0.560	0.810/0.513	0.857/0.543
CzeLan	**0.217/0.252**	0.225/0.262	0.241/0.292	0.234/0.282	0.250/0.304	0.245/0.297	0.242/0.290	0.229/0.276	0.258/0.312
Temp	**0.175/0.329**	0.187/0.338	0.295/0.413	0.302/0.423	0.315/0.442	0.302/0.423	0.292/0.418	0.291/0.406	0.304/0.427

F Backbone Transferability

We further examine whether NEIN-TSF depends on a particular forecasting architecture. We choose iTransformer (Liu et al., 2023a) and SAMformer (Ilbert et al., 2024) as two complementary Transformer backbones. iTransformer adopts an inverted, variate-centric formulation in which individual variables are treated as tokens to model multivariate dependencies, whereas SAMformer is a shallow and lightweight Transformer that combines channel-wise attention with sharpness-aware optimization. Their different tokenization, model capacity, and optimization strategies therefore provide a useful test of whether the proposed consequence learning, historically supported refinement, and utility control transfer beyond a single backbone design.

Table 10 reports the complete horizon-wise results. For each backbone, its original forecasting configuration is kept unchanged when NEIN-TSF is introduced, so the comparison isolates the effect of the proposed framework.

Across all six datasets, NEIN-TSF improves the horizon-averaged MSE and MAE for both forecasting backbones. Averaged over datasets, the relative MSE/MAE reductions are $4.54\%/4.70\%$ for iTransformer and $6.34\%/4.55\%$ for SAMformer. The gains are observed across the ETT, Electricity, and Weather benchmarks despite the substantial architectural and optimization differences between the two backbones. At individual horizons, a few metrics remain nearly unchanged or slightly favor the base model (e.g., SAMformer on ETTh1 or ETTh2 at $H = 720$), but the horizon-averaged results improve consistently. These results provide evidence that NEIN-TSF is not tied to a particular forecasting backbone and that its consequence-aware learning, support-conditioned refinement, and utility control can complement different forecasting architectures.

G Full Mechanism Diagnostics: Consequence, Support, and Utility

This section provides the full versions of the three mechanism analyses summarized in Figure 3: forecast-consequence learning, support-conditioned refinement, and correction utility across forecasting distance and chronological time.

G.1 Forecast-Consequence Audit of FDRP

Figure 4 compares FDRP-predicted damage with the teacher-derived target. Across ETTh1, ETTm1, Weather, and ECL, the mean Spearman correlation with teacher-derived damage increases from 0.276 for input deviation to 0.654 for predicted damage. On ETTh1, the mild–severe damage gaps additionally achieve $\rho = 0.660$ and 77.8% ranking accuracy.

Table 10: **Transferability of NEIN-TSF across forecasting backbones.** Base and +NEIN-TSF share the same backbone configuration. Results are reported in MSE/MAE; lower is better. Boldface marks the better value within each Base/+NEIN-TSF pair.

Dataset	H	iTransformer		SAMformer		Dataset	H	iTransformer		SAMformer	
		Base	+NEIN	Base	+NEIN			Base	+NEIN	Base	+NEIN
ETTh1	96	0.386/0.405	**0.372/0.389**	0.381/0.402	**0.356/0.386**	ETTh2	96	0.297/0.349	**0.285/0.331**	0.295/0.358	**0.273/0.339**
	192	0.441/0.436	**0.423/0.418**	0.409/0.418	**0.389/0.409**		192	0.380/0.400	**0.374/0.387**	0.340/0.386	**0.328/0.369**
	336	0.487/0.458	**0.466/0.439**	0.423/0.425	**0.413/0.413**		336	0.428/0.432	**0.414/0.423**	0.350/0.395	**0.349/0.391**
	720	0.503/0.491	**0.478/0.460**	**0.427**/0.449	0.428/**0.440**		720	0.427/0.445	**0.419/0.435**	**0.391/0.428**	**0.382**/0.431
	Avg	0.454/0.447	**0.435/0.427**	0.410/0.424	**0.397/0.412**		*Avg*	0.383/0.407	**0.373/0.394**	0.344/0.392	**0.333/0.383**
ETTm1	96	0.334/0.368	**0.310/0.340**	0.329/0.363	**0.301/0.346**	ETTm2	96	0.180/0.264	**0.169/0.249**	0.181/0.274	**0.170/0.266**
	192	0.377/0.391	**0.359/0.367**	0.353/0.378	**0.337/0.361**		192	0.250/0.309	**0.234/0.288**	0.233/0.306	**0.221/0.295**
	336	0.426/0.420	**0.400/0.395**	0.382/0.394	**0.363/0.380**		336	0.311/0.348	**0.302/0.333**	0.285/0.338	**0.276/0.328**
	720	0.491/0.459	**0.462/0.429**	0.429/0.418	**0.419/0.412**		720	0.412/0.407	**0.393/0.393**	0.375/0.390	**0.366/0.378**
	Avg	0.407/0.410	**0.383/0.383**	0.373/0.388	**0.355/0.375**		*Avg*	0.288/0.332	**0.275/0.316**	0.269/0.327	**0.258/0.317**
Electricity	96	0.148/0.240	**0.136/0.224**	0.155/0.252	**0.137/0.236**	Weather	96	0.174/0.214	**0.162/0.200**	0.197/0.249	**0.170/0.216**
	192	0.162/0.253	**0.152/0.245**	0.168/0.263	**0.151/0.252**		192	0.221/0.254	**0.211/0.241**	0.235/0.277	**0.200/0.253**
	336	0.178/0.269	**0.159/0.255**	0.183/0.277	**0.164/0.260**		336	0.278/0.296	**0.272/0.285**	0.276/0.304	**0.240/0.274**
	720	0.225/0.317	**0.218/0.307**	0.219/0.306	**0.196/0.284**		720	0.358/0.347	**0.346/0.332**	0.334/0.342	**0.305/0.318**
	Avg	0.178/0.270	**0.167/0.258**	0.181/0.275	**0.162/0.258**		*Avg*	0.258/0.278	**0.248/0.265**	0.261/0.293	**0.229/0.265**

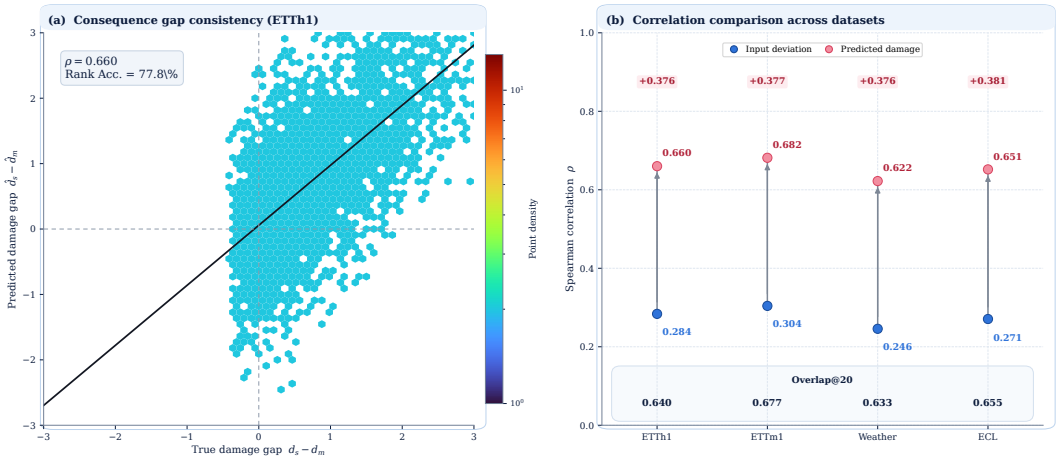


Figure 4: **Full forecast-consequence audit of FDRP.** (a) Teacher-derived versus predicted mild–severe damage gaps on ETTh1. (b) Correlation of true damage with input deviation and FDRP-predicted damage across four datasets; Overlap@20 compares the corresponding top-20% sets.

Table 11: **Dataset-wise FDRP diagnostics.** Pred. ρ is the Spearman correlation between predicted and teacher-derived damage; Input ρ uses raw input deviation. Overlap@20 compares the top-20% sets.

Dataset	Rank Acc.	Pred. ρ	Input ρ	Overlap@20
ETTh1	0.778	0.660	0.284	0.640
ETTm1	0.786	0.682	0.304	0.677
Weather	0.750	0.622	0.246	0.633
ECL	0.763	0.651	0.271	0.655
Mean	0.769	0.654	0.276	0.651

G.2 Support–Utility Diagnostics of SPE

Let $B = e_0 - e_{\text{full}}$ denote the squared-error reduction relative to the frozen base forecast. Figure 5 shows that correction benefit generally increases with support, whereas the harmful-correction rate decreases. The positive support–benefit association persists at $H = 720$, with the mean Spearman

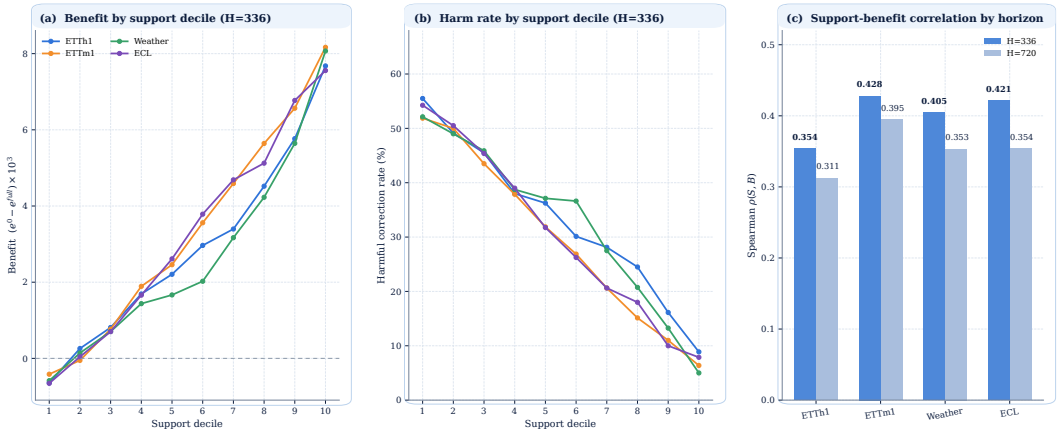


Figure 5: **Full support–utility analysis of SPE.** (a) Correction benefit and (b) harmful-correction rate across support deciles at $H = 336$. (c) Support–benefit Spearman correlation at $H \in \{336, 720\}$.

Table 12: **SPE evidence diagnostics.** Source-level support statistics are reported where available.

Split	Candidate Win	Mean Support	Retrieval	Periodicity	Trend
Train	0.737	0.708	0.352	0.679	0.243
Val.	0.749	0.725	0.341	0.697	0.242
Test	0.729	0.712	0.370	0.687	0.228

Table 13: **Support–benefit Spearman correlation** at long horizons.

Dataset	$H = 336$	$H = 720$
ETTh1	0.354	0.311
ETTm1	0.428	0.395
Weather	0.405	0.353
ECL	0.421	0.354
Mean	0.402	0.354

correlation changing from 0.402 at $H = 336$ to 0.354. Support is therefore informative about correction utility, although it is neither a calibrated probability nor a pointwise guarantee of improvement.

Candidate availability. The benefit mask asks whether at least one historical candidate improves on the frozen base by the prescribed margin. Candidate-win rate therefore measures oracle candidate availability rather than the realized success rate of the learned correction.

G.3 Horizon- and Time-Dependent Correction Utility

Figure 6 provides a post-hoc analysis of how evidence-based correction utility varies across forecasting distance and chronological time. For ETTh1, ETTm1, and Weather, MSE gain weakens at distant horizons and becomes slightly negative in some later blocks, whereas ECL retains positive utility over a longer range. HAEU correspondingly learns horizon-dependent coefficients without imposing monotonic decay.

Defining $H^\dagger$ as the largest horizon boundary for which all displayed chronological blocks retain positive MSE gain gives 630, 630, 540, and 720 for ETTh1, ETTm1, Weather, and ECL, respectively. These values are descriptive post-hoc summaries rather than tuning targets or guarantees. Test-block statistics are used only for this analysis and are never used to fit HAEU coefficients or select CFUS checkpoints.

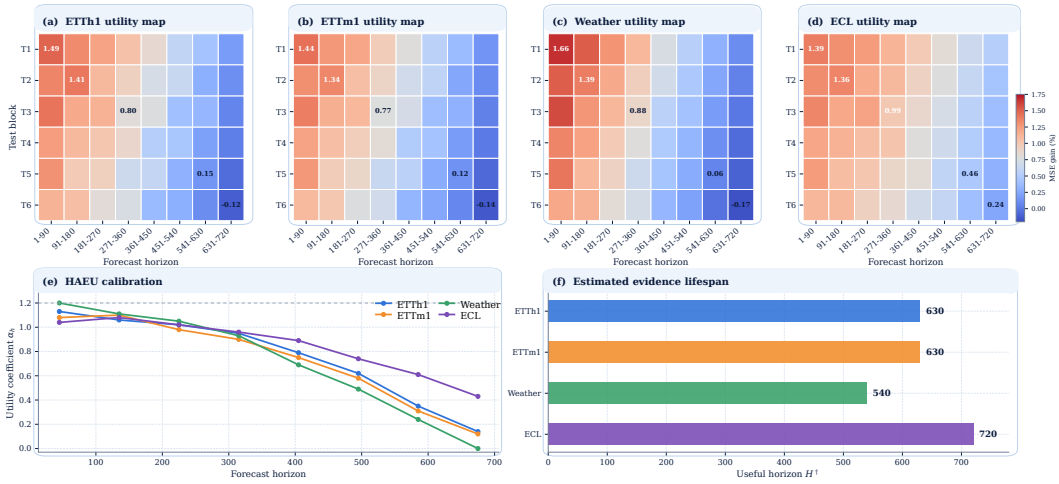


Figure 6: **Horizon- and time-dependent correction utility.** (a–d) Post-hoc MSE-gain maps across horizon bins and chronologically ordered held-out blocks. (e) HAEU coefficients under $0 \leq \alpha_h \leq 1.2$. (f) Descriptive persistence boundary $H^\dagger$. Test statistics are used only for post-hoc analysis.

Table 14: **Core module ablation of NEIN-TSF.** Results are averaged over $H \in \{96, 192, 336, 720\}$. Each entry reports MSE/MAE; lower is better.

Variant	ETTh1	ETTm1	Weather	ECL
Vanilla backbone	0.416/0.429	0.374/0.388	0.247/0.269	0.169/0.257
Vanilla + FDRP	0.407/0.422	0.368/0.382	0.241/0.264	0.164/0.252
Vanilla + SPE	0.411/0.425	0.371/0.385	0.243/0.263	0.165/0.251
FDRP + SPE	0.402/0.418	0.364/0.378	0.237/0.259	0.160/0.248
FDRP + SPE + HAEU	0.399/0.417	0.362/0.377	0.235/0.258	0.158/0.247
FDRP + SPE + CFUS	0.400/0.416	0.363/0.376	0.236/0.257	0.159/0.246
Full NEIN-TSF	**0.396/0.415**	**0.360/0.375**	**0.233/0.256**	**0.156/0.245**

H Ablation Studies

We conduct comprehensive ablations on ETTh1, ETTm1, Weather, and ECL. All results are averaged over $H \in \{96, 192, 336, 720\}$ and reported as MSE/MAE. Unless otherwise specified, each variant changes only the component named in the row while retaining the remaining NEIN-TSF configuration.

H.1 Core Module Ablation

Table 14 evaluates the progressive contribution of FDRP, SPE, HAEU, and CFUS. FDRP and SPE each improve the vanilla backbone, their combination yields further gains, and the complete model performs best on all four representative datasets.

H.2 FDRP Ablations

Counterfactual construction and supervision. Table 15 examines the two central FDRP design choices. Breaking the matching constraints consistently degrades performance. Replacing teacher-derived damage ranking with intensity ranking or input-deviation ranking is also consistently weaker, supporting forecast consequence rather than perturbation magnitude as the supervision target. Removing the clean forecasting anchor causes the largest degradation in this group.

Table 15: **FDRP ablation: counterfactual construction and forecast-damage supervision.** Results are averaged over $H \in \{96, 192, 336, 720\}$ and reported in MSE/MAE.

Variant	ETTh1	ETTm1	Weather	ECL
Counterfactual construction				
Unmatched counterfactual pairs	0.404/0.420	0.368/0.379	0.239/0.260	0.160/0.248
w/o position matching	0.403/0.420	0.365/0.379	0.238/0.259	0.160/0.247
w/o perturbation-type matching	0.402/0.419	0.365/0.379	0.238/0.260	0.159/0.247
w/o shared direction/noise	0.401/0.418	0.366/0.379	0.237/0.259	0.158/0.247
w/o width matching	0.401/0.418	0.364/0.378	0.236/0.259	0.159/0.247
Forecast-damage supervision				
Intensity ranking	0.405/0.420	0.368/0.380	0.240/0.261	0.162/0.249
Input-deviation ranking	0.406/0.422	0.369/0.381	0.241/0.262	0.162/0.249
Absolute forecast-error supervision	0.402/0.418	0.367/0.379	0.238/0.260	0.160/0.248
Damage regression	0.401/0.418	0.365/0.378	0.238/0.259	0.158/0.246
w/o damage-gap filtering δ_d	0.400/0.417	0.365/0.378	0.236/0.258	0.160/0.248
Online damage teacher	0.404/0.419	0.366/0.381	0.240/0.261	0.161/0.248
w/o forecasting anchor	0.409/0.424	0.371/0.382	0.243/0.263	0.164/0.250

Table 16: **FDRP ablation: perturbation types and context-residual representation.** Results are averaged over $H \in \{96, 192, 336, 720\}$ and reported in MSE/MAE.

Variant	ETTh1	ETTm1	Weather	ECL
Perturbation types				
w/o Jump	0.400/0.417	0.363/0.376	0.236/0.258	0.159/0.247
w/o Noise	0.399/0.418	0.364/0.378	0.236/0.258	0.158/0.246
w/o Erase	0.398/0.416	0.363/0.378	0.236/0.258	0.159/0.246
w/o Drift	0.401/0.418	0.364/0.379	0.238/0.260	0.160/0.248
Jump only	0.405/0.421	0.369/0.381	0.240/0.262	0.163/0.249
Noise only	0.407/0.422	0.370/0.382	0.239/0.260	0.163/0.250
Erase only	0.408/0.423	0.369/0.381	0.242/0.262	0.162/0.249
Drift only	0.404/0.419	0.367/0.380	0.240/0.261	0.161/0.248
Context-residual representation				
X^a only	0.404/0.420	0.369/0.381	0.241/0.262	0.160/0.248
w/o local reference R	0.403/0.419	0.368/0.381	0.239/0.260	0.160/0.248
w/o signed residual	0.401/0.419	0.365/0.378	0.237/0.258	0.159/0.248
w/o absolute residual	0.400/0.418	0.364/0.377	0.236/0.258	0.159/0.247
w/o dynamic mismatch	0.402/0.419	0.366/0.379	0.238/0.260	0.160/0.248
Residual-only representation	0.405/0.420	0.368/0.381	0.240/0.261	0.161/0.248

Perturbation diversity and context-residual representation. Table 16 shows that no single perturbation type matches the complete FDRP configuration. Single-perturbation variants are consistently weaker than the four-type construction. Likewise, removing individual context-residual features degrades performance, while using only X^a or only residual features is clearly weaker than the complete representation.

Table 17: **SPE ablation: historical evidence bank and support fusion.** Results are averaged over $H \in \{96, 192, 336, 720\}$ and reported in MSE/MAE.

Variant	ETTh1	ETTm1	Weather	ECL
Historical evidence bank				
w/o Retrieval	0.400/0.418	0.363/0.377	0.237/0.259	0.159/0.247
w/o Periodicity	0.398/0.417	0.364/0.378	0.239/0.261	0.158/0.247
w/o Local Trend	0.400/0.418	0.365/0.378	0.238/0.260	0.160/0.248
w/o Persistence	0.399/0.416	0.363/0.377	0.236/0.258	0.159/0.247
Retrieval only	0.405/0.420	0.368/0.380	0.241/0.262	0.163/0.249
Periodicity only	0.406/0.422	0.369/0.381	0.244/0.264	0.162/0.250
Local Trend only	0.404/0.419	0.367/0.380	0.242/0.262	0.161/0.248
Persistence only	0.408/0.421	0.369/0.382	0.245/0.265	0.163/0.250
Support and evidence fusion				
No support ($S^k = 1$)	0.406/0.421	0.369/0.380	0.242/0.262	0.162/0.249
w/o support in evidence weighting	0.400/0.418	0.363/0.378	0.236/0.258	0.159/0.247
w/o aggregate support multiplier S	0.403/0.420	0.368/0.380	0.239/0.260	0.161/0.248
Uniform evidence weights	0.401/0.418	0.365/0.378	0.238/0.260	0.160/0.248
Hard evidence selection	0.401/0.419	0.365/0.379	0.239/0.260	0.160/0.248

H.3 SPE Ablations

Historical evidence bank and support fusion. Table 17 shows that removing any evidence source reduces performance and that single-source variants are consistently weaker than the complete evidence bank. Eliminating support entirely causes a larger degradation than changing the fusion rule alone, supporting the role of historical support as a regulator of residual correction.

Bounded and conservative refinement. Table 18 evaluates the residual parameterization, initialization strategy, backbone freezing, and benefit-aware training terms. The largest degradations arise from unbounded residual correction, direct future prediction, removing the benefit mask, removing the no-harm term, or using the fit loss alone. Zero initialization and freezing the pretrained backbone are also consistently beneficial.

H.4 HAEU and CFUS Ablations

Horizon-adaptive utility. Table 19 compares the proposed horizon-wise calibration with fixed, coarse, signed, and unbounded alternatives. Horizon-wise bounded calibration performs best overall, while the four-bin alternative approaches its performance but remains slightly weaker. Signed or unbounded coefficients do not improve forecasting accuracy.

Chronological utility selection. Table 20 compares CFUS with last-checkpoint selection, conventional validation selection, random validation, reuse of the same data for calibration and selection, refitting on $\mathcal{D}_{\mathrm{fwd}}$, and alternative chronological split ratios. The balanced $50{:}50$ split provides the strongest overall result among the tested ratios.

H.5 Hyperparameter Sensitivity

We examine the sensitivity of NEIN-TSF to four representative hyperparameters covering its major components: the FDRP ranking weight λ_{rank}, the SPE correction scale γ, the no-harm weight λ_{nh}, and the HAEU upper bound $\alpha_{\max}$. Unless the parameter under study is varied, all other settings remain fixed at their default values. For λ_{rank}, γ, and λ_{nh}, the reported settings denote multiplicative factors relative to their respective default values.

Table 18: **SPE ablation: bounded and conservative refinement.** Results are averaged over $H \in \{96, 192, 336, 720\}$ and reported in MSE/MAE.

Variant	ETTh1	ETTm1	Weather	ECL
w/o $\tanh$ bound	0.402/0.419	0.367/0.380	0.240/0.261	0.161/0.248
w/o scale factor σ	0.402/0.418	0.366/0.380	0.239/0.260	0.160/0.248
Unbounded residual correction	0.407/0.423	0.370/0.382	0.244/0.265	0.165/0.250
Direct future prediction	0.411/0.424	0.372/0.383	0.246/0.264	0.164/0.252
Xavier initialization	0.400/0.417	0.363/0.377	0.236/0.258	0.157/0.246
Gaussian initialization	0.400/0.417	0.363/0.378	0.236/0.259	0.158/0.247
Jointly fine-tune backbone	0.401/0.419	0.366/0.378	0.239/0.260	0.160/0.247
Fine-tune final backbone layer	0.398/0.417	0.364/0.377	0.236/0.258	0.159/0.246
w/o benefit mask	0.403/0.420	0.367/0.380	0.240/0.261	0.161/0.248
w/o no-harm loss	0.405/0.421	0.369/0.382	0.241/0.262	0.163/0.250
w/o conservative regularizer	0.399/0.418	0.365/0.378	0.236/0.258	0.159/0.247
Fit loss only	0.406/0.422	0.370/0.383	0.243/0.262	0.163/0.250
w/o base-motion feature U^0	0.399/0.416	0.363/0.377	0.236/0.258	0.158/0.247
w/o horizon feature η_h	0.400/0.417	0.364/0.378	0.238/0.259	0.159/0.247
w/o U^0 and η_h	0.402/0.419	0.366/0.379	0.239/0.261	0.160/0.247

Table 19: **HAEU ablation.** Results are averaged over $H \in \{96, 192, 336, 720\}$ and reported in MSE/MAE.

Variant	ETTh1	ETTm1	Weather	ECL
No HAEU ($\alpha_h = 1$)	0.400/0.417	0.363/0.377	0.236/0.258	0.159/0.246
Global scalar α	0.398/0.417	0.362/0.377	0.235/0.258	0.158/0.246
Four-bin α	0.397/0.416	0.361/0.376	0.234/0.257	0.158/0.245
Signed α_h	0.399/0.417	0.363/0.377	0.237/0.258	0.158/0.248
Unbounded α_h	0.402/0.418	0.364/0.379	0.238/0.259	0.161/0.247
w/o ridge regularization	0.399/0.416	0.362/0.377	0.235/0.257	0.158/0.247
Full NEIN-TSF	**0.396/0.415**	**0.360/0.375**	**0.233/0.256**	**0.156/0.245**

Table 20: **CFUS ablation.** Results are averaged over $H \in \{96, 192, 336, 720\}$ and reported in MSE/MAE.

Variant	ETTh1	ETTm1	Weather	ECL
Last checkpoint	0.402/0.418	0.365/0.379	0.239/0.260	0.159/0.248
Conventional validation best	0.401/0.418	0.363/0.378	0.237/0.258	0.158/0.247
Random validation split	0.400/0.417	0.364/0.377	0.236/0.257	0.159/0.247
Same calibration/selection set	0.399/0.417	0.361/0.377	0.235/0.257	0.158/0.246
Refit α on $\mathcal{D}_{\mathrm{fwd}}$	0.399/0.417	0.363/0.377	0.236/0.258	0.158/0.246
$\mathcal{D}_{\mathrm{cal}} : \mathcal{D}_{\mathrm{fwd}} = 25{:}75$	0.397/0.417	0.362/0.377	0.235/0.258	0.158/0.246
$\mathcal{D}_{\mathrm{cal}} : \mathcal{D}_{\mathrm{fwd}} = 75{:}25$	0.399/0.416	0.363/0.377	0.236/0.258	0.159/0.247
Full NEIN-TSF	**0.396/0.415**	**0.360/0.375**	**0.233/0.256**	**0.156/0.245**

Table 21: **Hyperparameter sensitivity of NEIN-TSF.** Results report MSE averaged over $H \in \{96, 192, 336, 720\}$. Avg. ΔMSE denotes the mean relative MSE change across the four datasets with respect to the default configuration. Lower is better. Boldface denotes the default setting.

Parameter	Setting	ETTh1	ETTm1	Weather	ECL	Avg. ΔMSE
λ_{rank}	$0.25\times$	0.402	0.364	0.237	0.159	+1.54%
	$0.5\times$	0.399	0.361	0.235	0.157	+0.62%
	$1.0\times$	**0.396**	**0.360**	**0.233**	**0.156**	**0.00%**
	$2.0\times$	0.397	0.362	0.232	0.156	+0.16%
	$4.0\times$	0.401	0.366	0.236	0.159	+1.40%
γ	$0.25\times$	0.403	0.366	0.238	0.159	+1.90%
	$0.5\times$	0.399	0.362	0.232	0.157	+0.41%
	$1.0\times$	**0.396**	**0.360**	**0.233**	**0.156**	**0.00%**
	$2.0\times$	0.398	0.361	0.235	0.156	+0.33%
	$4.0\times$	0.404	0.368	0.239	0.160	+2.33%
λ_{nh}	$0\times$	0.405	0.369	0.241	0.163	+3.17%
	$0.5\times$	0.399	0.362	0.235	0.158	+0.87%
	$1.0\times$	**0.396**	**0.360**	**0.233**	**0.156**	**0.00%**
	$2.0\times$	0.397	0.361	0.233	0.156	+0.11%
	$4.0\times$	0.400	0.365	0.236	0.159	+1.38%
$\alpha_{\max}$	0.6	0.401	0.364	0.237	0.159	+1.52%
	0.9	0.397	0.361	0.234	0.157	+0.45%
	1.2	**0.396**	**0.360**	**0.233**	**0.156**	**0.00%**
	1.5	0.395	0.361	0.234	0.156	+0.07%
	1.8	0.400	0.363	0.236	0.158	+1.11%

The results reveal a relatively broad stable region around the default configuration. Moderate deviations generally lead to only small performance changes, whereas overly weak or strong settings cause clearer degradation. In particular, removing the no-harm penalty ($\lambda_{\mathrm{nh}} = 0$) increases the average MSE by 3.17%, while an excessively large correction scale ($\gamma = 4\times$) increases it by 2.33%. The default configuration achieves the best overall average performance across the four representative datasets, although individual datasets may slightly benefit from nearby settings. These results suggest that NEIN-TSF does not rely on a narrowly tuned hyperparameter configuration.

H.6 Random-Seed Stability

To quantify the variability of the main results, we report the individual runs corresponding to the three random seeds $\{2022, 2023, 2024\}$. All other experimental settings are kept fixed. Table 22 reports horizon-averaged MSE/MAE for each run together with the mean and sample standard deviation.

Across all seven main benchmarks, NEIN-TSF exhibits low variability across random seeds. The MSE standard deviation ranges from approximately 0.001 to 0.003, while the MAE standard deviation ranges from approximately 0.001 to 0.002. The three-run averages coincide with the standard results reported in the main comparison at the displayed precision, indicating that the reported performance is stable across random initializations and is not driven by a particular seed.

Table 22: **Random-seed stability of NEIN-TSF.** Each entry reports horizon-averaged MSE/MAE. The final column reports mean ± sample standard deviation over three independent runs.

Dataset	Seed 2022	Seed 2023	Seed 2024	Mean ± Std
ETTh1	0.398/0.416	0.393/0.413	0.398/0.416	**0.396±0.003 / 0.415±0.002**
ETTh2	0.335/0.381	0.338/0.385	0.338/0.383	**0.337±0.002 / 0.383±0.002**
ETTm1	0.358/0.374	0.363/0.377	0.359/0.374	**0.360±0.002 / 0.375±0.002**
ETTm2	0.251/0.309	0.248/0.306	0.251/0.309	**0.250±0.002 / 0.308±0.001**
Weather	0.232/0.257	0.235/0.254	0.232/0.257	**0.233±0.002 / 0.256±0.002**
ECL	0.157/0.246	0.155/0.244	0.156/0.245	**0.156±0.001 / 0.245±0.001**
Traffic	0.388/0.268	0.392/0.272	0.393/0.269	**0.391±0.003 / 0.270±0.002**