INSIGHT PRECEDES ACTION: A GEOMETRIC RETHINKING OF TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

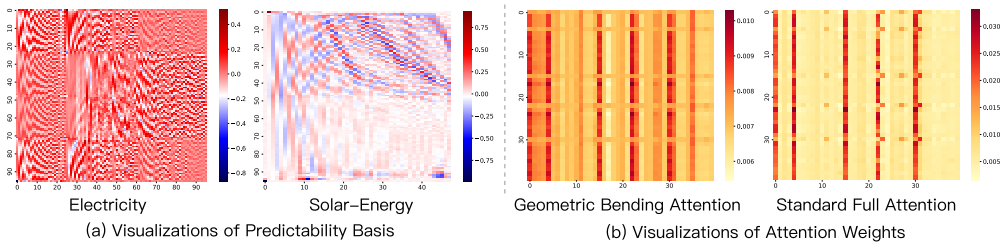
Time series forecasting has advanced rapidly, yet most approaches rely on end-to-end optimization to discover predictive structure in the training data, leaving its potential as an explicit prior underexplored. Meanwhile, attention-based models typically determine cross-channel interaction through query–key similarity, without directly considering the value representations being aggregated. These considerations lead to two key questions: (i) how can predictive structure be extracted directly from the training data and provided to the model as prior knowledge, and (ii) should the information being aggregated also contribute to determining cross-channel interaction strength? To address these questions, we introduce *Navigator*, which integrates *Foreseer* and *BendingFormer* to rethink forecasting through two complementary geometric structures. A *Predictability Basis* is first constructed from past–future statistical relationships in the training data by identifying linear directions in the historical observation space with strong predictive relevance. Foreseer then uses this basis to re-express historical observations in a predictability-oriented coordinate system for subsequent modeling. BendingFormer introduces *Geometric Bending Attention* (GBA), which measures the directional alignment between query-to-key and key-to-value transitions, allowing value representations to directly inform cross-channel interaction strength. Together, these components embody the principle that *insight precedes action*: predictive structure extracted before training guides temporal representation, while the joint query–key–value geometry guides information aggregation. Extensive experiments on twelve real-world datasets demonstrate state-of-the-art forecasting performance with consistent results across random seeds, while comprehensive analyses support the effectiveness of its two core designs.

1 INTRODUCTION

Time series are ubiquitous across domains such as energy, transportation, and finance, where accurate forecasting plays a critical role in planning, risk management, and operational efficiency (Qiu et al., 2026; Wang et al., 2026a; Gao et al., 2025). Recent deep learning methods have improved forecasting performance by advancing temporal modeling (Cao et al., 2026) and cross-channel interaction (Liu et al., 2025). Yet stronger modeling capacity does not necessarily imply a clearer understanding of what information should guide forecasting, suggesting that effective forecasting should begin with an understanding of the available information before proceeding to forecasting.

Existing research remains largely model-centric, with the information useful for forecasting primarily expected to be learned through end-to-end model optimization. Although this paradigm has long proven effective, it leaves the potential value inherent in the training data underexplored: information that could be characterized directly from the training data must instead be discovered through model optimization, placing an additional learning burden on the model and making its acquisition dependent on the model architecture and optimization process. Inspired by the intuition underlying the ergodic theorem (Breiman, 1992; Gray, 1988), that aggregate statistics over sufficiently long observations can reveal regularities of the underlying process, we look more closely at what can be learned directly from the training data itself. A long sequence in the training data naturally contains numerous historical–future instances, whose aggregate statistical relationships provide a direct means to assess which historical information is more informative about subsequent observations, revealing what we refer to as *predictive structure*. Such structure can then serve as a data-derived

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.



(a) Visualizations of Predictability Basis (b) Visualizations of Attention Weights

Figure 1: (a) Visualizations of Predictability Bases computed from the training data of Electricity and Solar-Energy datasets, where columns denote selected temporal directions, rows denote historical time steps, and colors indicate the corresponding basis coefficients. The Predictability Basis construction and additional visualizations are provided in Sec. 4.2.1 and Fig. 7, respectively. (b) Visualizations of attention weights produced by Geometric Bending Attention (GBA) and standard full attention on Solar-Energy with an input/prediction length of 96/96. For clarity, only the first 40 of 137 channels are shown in the default order. Compared with standard full attention, GBA preserves several prominent interaction patterns while emphasizing additional cross-channel interactions.

prior for forecasting. This motivates our first question: *how can predictive structure be explicitly extracted from the training data and incorporated into forecasting as prior knowledge?*

Beyond exploiting predictive structure in the temporal dimension, cross-channel interaction raises another intriguing question. Attention mechanisms have become a mainstream approach for modeling cross-channel interactions in multivariate time series forecasting (Zhang & Yan, 2023; Zhou et al., 2024; Lin et al., 2025a). In standard attention, interaction strengths are typically determined from pairwise query–key similarity, while value representations are used only in the subsequent weighted aggregation (Vaswani et al., 2017). This separation is effective and widely adopted (Liu et al., 2024), but it leaves the value content to be aggregated outside the scoring process itself. We therefore ask: *if an attention score determines how much information is drawn from a value representation, should that value also contribute to determining the interaction strength?*

To answer the first question, we extract the predictive structure embedded in the training data by examining different linear directions in the historical observation space, referred to as *temporal directions*, and characterizing which of them are more informative about future observations. The temporal directions with greater informativeness are then organized into a *Predictability Basis*, which serves as a data-derived prior for emphasizing forecasting-relevant information while attenuating less informative components. As illustrated in Figure 1(a), the Predictability Bases computed from the training data of the Electricity and Solar-Energy datasets exhibit distinct structured patterns, reflecting dataset-specific predictive structures. Building on the Predictability Basis, we design *Foreseer* to transform historical observations into a predictability-oriented representation before subsequent forecasting. Moreover, Foreseer can be integrated into existing models as a plug-in (see Table 3); for example, incorporating it into iTransformer reduces the MSE by 8.05% on the Electricity dataset with an input/prediction length of 96/96. These plug-in results further validate the effectiveness of prior knowledge extracted directly from the training data in supporting forecasting.

For cross-channel modeling, we investigate how the value representation, which carries the information to be aggregated, can participate in determining the interaction strength. Rather than introducing additional pairwise similarities, we consider the ordered query–key–value configuration and characterize the geometric bending formed by the query-to-key and key-to-value transitions, using their directional compatibility as the interaction criterion. Based on this criterion, we introduce *Geometric Bending Attention* (GBA) and build *BendingFormer* around it for geometry-aware cross-channel modeling. As illustrated in Figure 1(b), GBA retains several prominent interaction patterns visible under standard attention while revealing additional channel interactions that receive limited emphasis under conventional query–key similarity. This consideration of the full query–key–value configuration is also supported by improved forecasting performance. Under the same experimental setting as Figure 1(b), our model with GBA achieves a 15.05% reduction in MSE compared with the variant in which GBA is replaced by standard full attention (see Tables 4 and 11).

Bringing these two perspectives together, we introduce *Navigator*, which integrates *Foreseer* and *BendingFormer*, embodying the principle that *insight precedes action* through a geometric rethinking of forecasting: historical observations are re-expressed within a data-derived temporal geometry

defined by selected temporal directions, while cross-channel interaction is governed by the geometry of the full query–key–value configuration. In summary, our contributions are as follows:

- We introduce a *Predictability Basis* that identifies temporal directions informative about future observations from historical–future statistics in the training data, and develop *Foreseer* to leverage this data-derived prior for predictability-oriented transformation.
- We propose *Geometric Bending Attention*, which evaluates the directional consistency between query-to-key and key-to-value transitions, allowing value representations to participate directly in determining cross-channel interaction strength, and further develop *BendingFormer* for geometry-aware cross-channel modeling.
- We integrate Foreseer and BendingFormer into *Navigator*. Extensive experiments on twelve real-world datasets demonstrate that Navigator achieves state-of-the-art forecasting performance, while comprehensive analyses validate the effectiveness of its two core designs.

2 Related Work

Deep Learning-based Time Series Forecasting. Deep learning has become a dominant paradigm for time series forecasting, with existing methods spanning RNNs (Lin et al., 2025b), CNNs (Liu et al., 2026; Wu et al., 2023), GNNs (Yi et al., 2023a), Transformers (Kim et al., 2026; Liu et al., 2025), and MLPs (Fei et al., 2025; Wang et al., 2024a; Lin et al., 2024a). These methods have achieved substantial progress through diverse architectural designs for modeling historical information and cross-channel dependencies. Despite this progress, existing research remains largely model-centric, with comparatively limited attention paid to what can be learned directly from the training data itself beyond its conventional role in end-to-end optimization. In particular, which temporal directions within the historical window are more informative about future observations has rarely been explicitly characterized and incorporated into forecasting as prior knowledge.

Attention-Based Methods for Cross-Channel Modeling. Cross-channel interaction is another important aspect of multivariate time series forecasting, where attention-based methods have become a mainstream approach (Wang et al., 2026b; Qiu et al., 2025; Yi et al., 2024b). For example, Crossformer (Zhang & Yan, 2023) and CARD (Zhou et al., 2024) model cross-channel interactions over aligned temporal segments or patches. iTransformer (Liu et al., 2024) instead treats the entire historical observations of each channel as a single token and applies self-attention across channels, an approach subsequently adopted or extended by Leddam (Yu et al., 2024), TQNet (Lin et al., 2025a), and TimeBridge (Liu et al., 2025). Despite their different designs, these attention-based methods generally determine attention weights from pairwise query–key similarity, while the corresponding value representation participates only after the weights have been computed. This leaves open the possibility of incorporating the joint geometry of query, key, and value into attention scoring, motivating us to rethink cross-channel interaction beyond pairwise similarity.

3 Preliminaries

3.1 Time Series Forecasting

Given historical observations $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T] \in \mathbb{R}^{N\times T}$ and the corresponding future observations $\mathbf{Y} = [\mathbf{x}_{T+1}, \cdots, \mathbf{x}_{T+\tau}] \in \mathbb{R}^{N\times \tau}$, where $\mathbf{x}_t \in \mathbb{R}^N$ denotes the observations across N channels at time step t, and T and τ denote the lookback window and prediction horizon, respectively. Time series forecasting aims to learn a forecasting function $f_\theta(\cdot)$ parameterized by θ that maps the historical observations $\mathbf{X}$ to the future observations, yielding $\hat{\mathbf{Y}} = f_\theta(\mathbf{X}) \in \mathbb{R}^{N\times\tau}$, where $\hat{\mathbf{Y}}$ denotes the predicted future observations.

3.2 Predictability Analysis

To characterize which information within the historical observations is most relevant to future observations, we consider different temporal directions in the historical window space, each capturing a particular linear combination of observations across time steps (see Appendix A.2 for further discussion). We quantify the predictive relevance of each direction through a notion termed *temporal predictability*. For clarity, we consider a single channel, where $\mathbf{p}_t = [x_{t-T+1}, \ldots, x_t]^\top \in \mathbb{R}^T$ denotes the historical window ending at time t, and $\mathbf{f}_t = [x_{t+1}, \ldots, x_{t+\tau}]^\top \in \mathbb{R}^\tau$ denotes the corresponding future window. We regard a nonzero vector $\mathbf{w} \in \mathbb{R}^T \setminus \{\mathbf{0}\}$ as a temporal direction in the historical window space. Projecting $\mathbf{p}_t$ onto $\mathbf{w}$ yields the scalar $\mathbf{w}^\top \mathbf{p}_t$, which represents the historical information along this temporal direction. To assess the predictive relevance of $\mathbf{w}$, we

consider the strongest linear association between the projected historical information $\mathbf{w}^\top \mathbf{p}_t$ and the future information represented along any temporal direction of the future window, as defined below:

Definition 1 (Temporal Predictability). *Let $\mathbf{p}_t \in \mathbb{R}^T$ and $\mathbf{f}_t \in \mathbb{R}^\tau$ denote paired historical and future windows, respectively. For a nonzero temporal direction $\mathbf{w} \in \mathbb{R}^T \setminus \{\mathbf{0}\}$, its* temporal predictability *is defined as*

$$\mathcal{P}(\mathbf{w}) = \max_{\mathbf{v} \in \mathbb{R}^\tau \setminus \{\mathbf{0}\}} \left| \mathrm{corr}\left(\mathbf{w}^\top \mathbf{p}_t, \mathbf{v}^\top \mathbf{f}_t\right) \right|, \tag{1}$$

where $\mathrm{corr}(\cdot,\cdot)$ denotes the Pearson correlation coefficient (Schober et al., 2018) between the projected historical and future information, and $\mathbf{v} \in \mathbb{R}^\tau \setminus \{\mathbf{0}\}$ ranges over all nonzero temporal directions in the future window space.

Accordingly, temporal directions can be compared based on $\mathcal{P}(\mathbf{w})$, with the most predictive direction obtained by maximizing $\mathcal{P}(\mathbf{w})$ over all nonzero $\mathbf{w} \in \mathbb{R}^T$. The following proposition shows that temporal predictability $\mathcal{P}(\mathbf{w})$ admits a generalized eigenvalue characterization, providing a principled way to identify temporal directions according to their predictive relevance.

Proposition 1 (Closed-form Characterization of Temporal Predictability). *Let $\mathbf{p}_t \in \mathbb{R}^T$ and $\mathbf{f}_t \in \mathbb{R}^\tau$ be centered random vectors with finite second moments, and define $\Sigma_{pp} = \mathbb{E}[\mathbf{p}_t \mathbf{p}_t^\top]$, $\Sigma_{ff} = \mathbb{E}[\mathbf{f}_t \mathbf{f}_t^\top]$, and $\Sigma_{pf} = \mathbb{E}[\mathbf{p}_t \mathbf{f}_t^\top]$, with $\Sigma_{fp} = \Sigma_{pf}^\top$. Assume that Σ_{pp} and Σ_{ff} are positive definite. Then:*

(i) Closed-form Characterization. *For any nonzero temporal direction $\mathbf{w} \in \mathbb{R}^T \setminus \{\mathbf{0}\}$, its temporal predictability satisfies*

$$\mathcal{P}(\mathbf{w})^2 = \frac{\mathbf{w}^\top \Sigma_{pf} \Sigma_{ff}^{-1} \Sigma_{fp} \mathbf{w}}{\mathbf{w}^\top \Sigma_{pp} \mathbf{w}}. \tag{2}$$

(ii) Generalized Eigenvalue Formulation. *Since Eq. (2) is a generalized Rayleigh quotient (Horn & Johnson, 2012; Parlett, 1998), its stationary temporal directions, along which its value has zero first-order variation, satisfy the generalized eigenvalue problem*

$$\Sigma_{pf} \Sigma_{ff}^{-1} \Sigma_{fp} \mathbf{w} = \lambda \Sigma_{pp} \mathbf{w}. \tag{3}$$

For any generalized eigenpair $(\lambda, \mathbf{w})$ satisfying Eq. (3), the corresponding temporal direction satisfies $\lambda = \mathcal{P}(\mathbf{w})^2$. Therefore, larger generalized eigenvalues correspond to temporal directions with higher predictability, with the largest generalized eigenvalue identifying the most predictive temporal directions.

The proof of Proposition 1 is provided in Appendix A.1. Proposition 1 characterizes temporal predictability through a generalized eigenvalue formulation, providing the theoretical foundation for identifying predictive temporal directions and constructing the Predictability Basis in Sec. 4.2.1.

4 METHOD

In this section, we elaborate on our proposed method, *Navigator*, which is designed to make informative structure explicit before and during forecasting. To this end, Navigator leverages a data-derived Predictability Basis to guide temporal representation and geometric bending over the full query–key–value configuration to guide cross-channel interaction.

4.1 OVERVIEW

As shown in Figure 2, *Navigator* consists of two main components, *Foreseer* and *BendingFormer*, followed by a forecasting head. Foreseer contains two submodules, *Prospector* and *Reconstructor*, which share a Predictability Basis precomputed from the training data prior to model training. Given historical observations $\mathbf{X} \in \mathbb{R}^{N \times T}$, Prospector projects $\mathbf{X}$ into a predictability-oriented temporal space, where BendingFormer performs geometry-aware cross-channel interaction through *Geometric Bending Attention* (GBA). Reconstructor then maps the resulting representation back to the original temporal space, and the forecasting head produces $\widehat{\mathbf{Y}} \in \mathbb{R}^{N \times \tau}$. In this way, Navigator unifies data-derived temporal predictability with geometry-aware cross-channel modeling.

4.2 FORESEER: PREDICTABILITY-ORIENTED TEMPORAL TRANSFORMATION

4.2.1 PREDICTABILITY BASIS CONSTRUCTION

Proposition 1 characterizes temporal predictability through the covariance relationship between historical and future observations. Motivated by this result, we construct a *Predictability Basis* shared across channels from the training set to capture historical directions with strong predictive relevance.

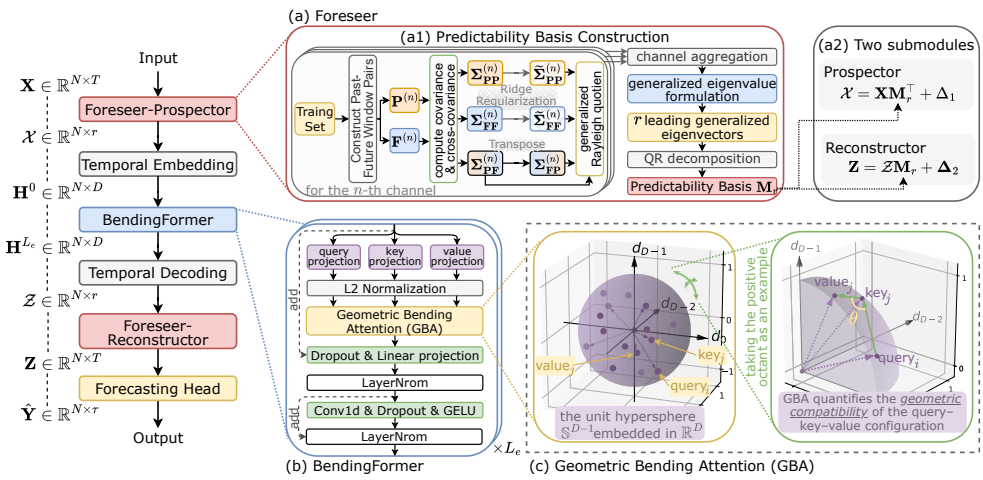


Figure 2: Overview of *Navigator*. (a) *Foreseer*: (a1) a Predictability Basis is constructed from the training data prior to model training; (a2) *Prospector* and *Reconstructor* share this basis to map historical observations into and recover representations from the predictability-oriented temporal space. (b–c) *BendingFormer* performs geometry-aware cross-channel modeling through Geometric Bending Attention (GBA), which determines interaction strengths according to the directional compatibility of geometric bending within the full query–key–value configuration. Additionally, Navigator applies instance normalization (Kim et al., 2021) to the input and output to mitigate non-stationarity (Kim et al., 2021; Liu et al., 2022b); this step is omitted hereafter for brevity.

Let $\mathbf{X}_{\text{train}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_{\text{train}}}] \in \mathbb{R}^{N \times T_{\text{train}}}$ denote the observations in the training set, where T_{train} is its temporal length. For the n-th channel, we construct historical–future window pairs by sliding along the temporal dimension, with $\mathbf{p}_t^{(n)} = [x_{t-T+1}^{(n)}, \ldots, x_t^{(n)}]^\top \in \mathbb{R}^T$ and $\mathbf{f}_t^{(n)} = [x_{t+1}^{(n)}, \ldots, x_{t+\tau}^{(n)}]^\top \in \mathbb{R}^\tau$ for $t = T, \ldots, T_{\text{train}} - \tau$. This yields $S = T_{\text{train}} - T - \tau + 1$ valid pairs, which are collected as $\mathbf{P}^{(n)} = [\mathbf{p}_T^{(n)}, \ldots, \mathbf{p}_{T_{\text{train}}-\tau}^{(n)}] \in \mathbb{R}^{T \times S}$ and $\mathbf{F}^{(n)} = [\mathbf{f}_T^{(n)}, \ldots, \mathbf{f}_{T_{\text{train}}-\tau}^{(n)}] \in \mathbb{R}^{\tau \times S}$. To match the centered formulation in Proposition 1, we subtract the sample mean vector from the S historical windows and future windows, respectively. The empirical covariance matrices are then estimated as $\mathbf{\Sigma}_{\mathbf{PP}}^{(n)} = \frac{1}{S}\mathbf{P}^{(n)}\mathbf{P}^{(n)\top} \in \mathbb{R}^{T \times T}$, $\mathbf{\Sigma}_{\mathbf{FF}}^{(n)} = \frac{1}{S}\mathbf{F}^{(n)}\mathbf{F}^{(n)\top} \in \mathbb{R}^{\tau \times \tau}$, $\mathbf{\Sigma}_{\mathbf{PF}}^{(n)} = \frac{1}{S}\mathbf{P}^{(n)}\mathbf{F}^{(n)\top} \in \mathbb{R}^{T \times \tau}$, and $\mathbf{\Sigma}_{\mathbf{FP}}^{(n)} = \mathbf{\Sigma}_{\mathbf{PF}}^{(n)\top} \in \mathbb{R}^{\tau \times T}$.

Following Proposition 1, we characterize temporal predictability for each channel using its covariance structure. For numerical stability, we apply ridge regularization (Hastie, 2020) to the covariance matrices involved in the inversion, yielding $\widetilde{\mathbf{\Sigma}}_{\mathbf{PP}}^{(n)} = \mathbf{\Sigma}_{\mathbf{PP}}^{(n)} + \epsilon \mathbf{I}_T$ and $\widetilde{\mathbf{\Sigma}}_{\mathbf{FF}}^{(n)} = \mathbf{\Sigma}_{\mathbf{FF}}^{(n)} + \epsilon \mathbf{I}_\tau$, where $\mathbf{I}_T$ and $\mathbf{I}_\tau$ are identity matrices of size T and τ, respectively, and $\epsilon > 0$ is a small regularization coefficient. For concise notation in the subsequent cross-channel aggregation, we define $\mathbf{A}^{(n)} = \mathbf{\Sigma}_{\mathbf{PF}}^{(n)} \left(\widetilde{\mathbf{\Sigma}}_{\mathbf{FF}}^{(n)}\right)^{-1} \mathbf{\Sigma}_{\mathbf{FP}}^{(n)}$ and $\mathbf{B}^{(n)} = \widetilde{\mathbf{\Sigma}}_{\mathbf{PP}}^{(n)}$, where $\mathbf{A}^{(n)}, \mathbf{B}^{(n)} \in \mathbb{R}^{T \times T}$ are the matrices defining the numerator and denominator quadratic forms, respectively, in the generalized Rayleigh quotient of Eq. (2).

To obtain temporal directions shared across channels, we aggregate the channel-wise formulations as $\overline{\mathbf{A}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{A}^{(n)}$ and $\overline{\mathbf{B}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{B}^{(n)}$, where $\overline{\mathbf{A}}, \overline{\mathbf{B}} \in \mathbb{R}^{T \times T}$. This aggregation combines the historical–future covariance information from all channels and yields a unified generalized eigenvalue problem. Since $\overline{\mathbf{A}}$ is symmetric positive semidefinite and $\overline{\mathbf{B}}$ is positive definite, the shared temporal directions are obtained by solving $\overline{\mathbf{A}}\mathbf{w}_i = \gamma_i \overline{\mathbf{B}}\mathbf{w}_i$, where $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_T$, and γ_i and $\mathbf{w}_i \in \mathbb{R}^T$ denote the i-th generalized eigenvalue and its corresponding generalized eigenvector, respectively. According to Proposition 1, larger generalized eigenvalues correspond to temporal directions with stronger predictive relevance. We therefore retain the leading r generalized eigenvectors and collect them as $\mathbf{W}_r = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r] \in \mathbb{R}^{T \times r}$, where $1 \leq r \leq T$. The generalized eigenvectors can be chosen to satisfy $\mathbf{w}_i^\top \overline{\mathbf{B}}\mathbf{w}_j = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta (Parlett, 1998). However, this $\overline{\mathbf{B}}$-weighted orthogonality does not generally imply Euclidean orthogonality, which is desirable for the subsequent projection and reconstruction. We therefore apply QR de-

composition (Trefethen & Bau, 1997) to $\mathbf{W}_r$, yielding $\mathbf{W}_r = \mathbf{Q}_r\mathbf{R}_r$, where $\mathbf{Q}_r \in \mathbb{R}^{T\times r}$ satisfies $\mathbf{Q}_r^\top\mathbf{Q}_r = \mathbf{I}_r$ and $\mathbf{R}_r \in \mathbb{R}^{r\times r}$ is upper triangular. Since QR decomposition preserves the column space of $\mathbf{W}_r$, $\mathbf{Q}_r$ spans the same r-dimensional predictability subspace. We finally define the *Predictability Basis* as $\mathbf{M}_r = \mathbf{Q}_r^\top \in \mathbb{R}^{r\times T}$, whose rows form an orthonormal basis of this subspace. Once constructed from the training set, $\mathbf{M}_r$ is kept fixed during validation and testing.

4.2.2 PROSPECTOR

Rather than treating all historical observations equally, Prospector transforms the input along the predictive temporal directions contained in the Predictability Basis. Given $\mathbf{M}_r \in \mathbb{R}^{r\times T}$, the input $\mathbf{X} \in \mathbb{R}^{N\times T}$ is projected as $\mathcal{X} = \mathbf{X}\mathbf{M}_r^\top + \mathbf{\Delta}_1 \in \mathbb{R}^{N\times r}$, where $\mathbf{\Delta}_1 \in \mathbb{R}^{N\times r}$ is a learnable bias for channel-specific adaptation. This projection maps $\mathbf{X}$ into a compact r-dimensional predictability-oriented representation for subsequent cross-channel modeling.

4.2.3 RECONSTRUCTOR

After cross-channel interaction in BendingFormer (Section 4.3), Reconstructor maps the resulting representation back to the original temporal space. Given the decoded BendingFormer output $\mathcal{Z} \in \mathbb{R}^{N\times r}$, Reconstructor applies the Predictability Basis shared with Prospector to obtain $\mathbf{Z} = \mathcal{Z}\mathbf{M}_r + \mathbf{\Delta}_2 \in \mathbb{R}^{N\times T}$, where $\mathbf{\Delta}_2$ provides channel-specific flexibility during reconstruction. By sharing $\mathbf{M}_r$ with Prospector, Reconstructor maps the refined representation from the same predictability-oriented coordinate system back to the original temporal space.

4.3 BENDINGFORMER: GEOMETRY-AWARE CROSS-CHANNEL MODELING

Conventional attention determines cross-channel interaction strengths from pairwise query–key affinity, without explicitly considering the value representations involved in information aggregation. To address this limitation, we introduce *BendingFormer*, whose core component, *Geometric Bending Attention* (GBA), measures geometric compatibility from the joint configuration of query, key, and value representations. By evaluating the directional consistency within this configuration, GBA assigns larger attention weights to interactions with stronger geometric compatibility, enabling geometry-aware information exchange across channels.

4.3.1 GEOMETRIC BENDING ATTENTION

Given the predictability-oriented representation $\mathcal{X} \in \mathbb{R}^{N\times r}$ produced by Prospector, we first map each channel representation into a D-dimensional latent space through a linear embedding, yielding $\mathbf{H} = \mathrm{Linear}(\mathcal{X}) \in \mathbb{R}^{N\times D}$, where D denotes the hidden dimension. We then obtain the query, key, and value representations through three learnable linear projections, i.e., $\mathbf{Q} = \mathrm{Linear}_Q(\mathbf{H})$, $\mathbf{K} = \mathrm{Linear}_K(\mathbf{H})$, and $\mathbf{V} = \mathrm{Linear}_V(\mathbf{H})$, with $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times D}$. To characterize their geometric relationships, the query, key, and value representations are ℓ_2-normalized onto the unit hypersphere $\mathbb{S}^{D-1} \subset \mathbb{R}^D$. Specifically, $\mathbf{q}_i = \mathbf{Q}_i/\|\mathbf{Q}_i\|_2$, $\mathbf{k}_j = \mathbf{K}_j/\|\mathbf{K}_j\|_2$, and $\mathbf{v}_j = \mathbf{V}_j/\|\mathbf{V}_j\|_2$, where $\mathbf{q}_i, \mathbf{k}_j, \mathbf{v}_j \in \mathbb{S}^{D-1}$ and $i, j \in \{1, \ldots, N\}$. This normalization removes the influence of vector magnitudes, allowing the geometric compatibility to be determined by their directional relationships.

For the i-th query and the j-th key–value pair, we consider the ordered geometric configuration $\mathbf{q}_i \to \mathbf{k}_j \to \mathbf{v}_j$. Specifically, we define two consecutive displacement vectors, $\mathbf{e}_{ij}^{qk} = \mathbf{k}_j - \mathbf{q}_i \in \mathbb{R}^D$ and $\mathbf{e}_j^{kv} = \mathbf{v}_j - \mathbf{k}_j \in \mathbb{R}^D$, representing the query-to-key and key-to-value transitions, respectively. Together, these transitions form an ordered geometric path whose direction may change at $\mathbf{k}_j$. We term this directional transition induced by the query–key–value configuration *geometric bending*. The relative orientation of the two displacement vectors characterizes the consistency of this bending, indicating whether the value continues or deviates from the geometric direction established by the query–key relation. Accordingly, we measure this directional consistency using the cosine similarity between the two displacement vectors:

$$c_{ij} = \frac{(\mathbf{e}_{ij}^{qk})^\top \mathbf{e}_j^{kv}}{\|\mathbf{e}_{ij}^{qk}\|_2 \|\mathbf{e}_j^{kv}\|_2} = \frac{(\mathbf{k}_j - \mathbf{q}_i)^\top (\mathbf{v}_j - \mathbf{k}_j)}{\|\mathbf{k}_j - \mathbf{q}_i\|_2 \|\mathbf{v}_j - \mathbf{k}_j\|_2}. \tag{4}$$

Here, $c_{ij} \in [-1, 1]$, with larger values indicating that the key-to-value transition is more aligned with the query-to-key transition and thus exhibits stronger geometric compatibility; collecting all c_{ij} for $i, j \in \{1, \ldots, N\}$ yields the geometric compatibility matrix $\mathbf{C} = [c_{ij}] \in \mathbb{R}^{N\times N}$.

We further introduce a learnable positive scaling factor α and define the attention scores as $\mathbf{S} = \alpha\mathbf{C}$, where $\alpha > 0$. Since α is positive, this scaling preserves the relative ranking of geometric

Table 1: Performance comparison with lookback length $T = 96$. Prediction horizons are $\tau \in \{12, 24, 48, 96\}$ for PEMS and $\tau \in \{96, 192, 336, 720\}$ for the other datasets. Results are averaged over four horizons. The best and second-best results are highlighted in **red** and **blue**, respectively. Full and additional baseline results are reported in Tables 7 and 8.

Models	**Navigator (ours)**		PatchLGA (2026)		DistDF (2026a)		VPNet (2026)		RAFT (2025)		FreDF (2025)		TimeXer (2024b)		FilterNet (2024a)		iTransformer (2024)		PatchTST (2023)		FreTS (2023b)	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.422**	**0.424**	0.437	0.432	0.430	0.429	0.434	**0.427**	**0.428**	0.433	0.437	0.435	0.437	0.437	0.441	0.439	0.454	0.447	0.469	0.455	0.488	0.474
ETTh2	0.364	**0.387**	**0.361**	0.390	0.367	0.393	**0.356**	**0.383**	0.382	0.410	0.371	0.396	0.368	0.396	0.383	0.407	0.383	0.407	0.387	0.407	0.550	0.515
ETTm1	**0.373**	**0.376**	0.396	0.393	0.378	0.394	**0.376**	**0.382**	0.381	0.400	0.392	0.399	0.382	0.397	0.392	0.401	0.407	0.410	0.387	0.400	0.407	0.415
ETTm2	**0.271**	**0.312**	0.279	0.325	0.277	0.321	**0.270**	**0.312**	0.281	0.330	0.278	**0.319**	0.274	0.322	0.285	0.328	0.288	0.332	0.281	0.326	0.335	0.379
Electricity	**0.161**	**0.250**	0.176	0.265	0.172	0.267	**0.162**	**0.251**	0.175	0.272	0.170	0.259	0.171	0.270	0.173	0.268	0.178	0.270	0.205	0.290	0.209	0.297
Traffic	0.440	**0.249**	0.463	0.309	**0.417**	0.279	0.421	**0.273**	**0.414**	0.284	0.421	0.279	0.466	0.287	0.463	0.310	0.428	0.282	0.481	0.304	0.552	0.348
Weather	**0.237**	**0.259**	0.253	0.273	0.248	0.275	**0.238**	**0.261**	0.270	0.309	0.254	0.274	0.241	0.271	0.245	0.272	0.258	0.278	0.259	0.281	0.340	0.363
Solar-Energy	**0.192**	**0.212**	0.218	0.267	0.207	0.255	**0.204**	**0.207**	0.301	0.303	0.211	0.262	0.237	0.302	0.235	0.266	0.233	0.262	0.270	0.307	0.268	0.322
PEMS03	**0.091**	**0.193**	0.130	0.245	0.123	0.233	0.122	0.232	0.144	0.230	0.113	0.219	**0.112**	**0.214**	0.145	0.251	0.113	0.221	0.180	0.291	0.167	0.273
PEMS04	**0.081**	**0.180**	0.100	0.205	**0.092**	**0.195**	0.111	0.219	0.104	0.210	0.096	0.202	0.105	0.209	0.146	0.258	0.111	0.221	0.195	0.307	0.196	0.302
PEMS07	**0.069**	**0.156**	0.092	0.193	**0.084**	**0.177**	0.094	0.198	0.094	0.193	0.089	0.182	0.085	0.182	0.123	0.229	0.101	0.204	0.211	0.303	0.204	0.298
PEMS08	**0.123**	**0.220**	0.165	0.251	0.151	0.244	0.159	0.245	0.151	0.234	**0.141**	0.238	0.175	0.250	0.172	0.260	0.150	**0.226**	0.280	0.321	0.205	0.295
1^{st} Count	9	10	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0

compatibility scores while allowing the model to adjust the sharpness of the resulting attention distribution. The attention matrix is then obtained by row-wise softmax, $\mathbf{A} = \mathrm{Softmax}(\mathbf{S})$, and the GBA output is computed as $\widetilde{\mathbf{H}} = \mathbf{AV} \in \mathbb{R}^{N\times D}$. In this way, GBA determines cross-channel interaction strength according to the geometric bending jointly formed by the query, key, and value representations, allowing geometric insight from the full query–key–value configuration to precede information aggregation.

4.3.2 BENDINGFORMER

BendingFormer adopts a standard Transformer encoder architecture (Liu et al., 2024; Nie et al., 2023), with self-attention replaced by GBA. Let $\mathbf{H}^{(\ell-1)} \in \mathbb{R}^{N\times D}$ denote the input to the ℓ-th encoder layer. First, GBA takes $\mathbf{H}^{(\ell-1)}$ as input and performs geometry-aware cross-channel interaction. The resulting $\widetilde{\mathbf{H}}^{(\ell-1)}$ is then processed by the standard Transformer post-attention operations (Liu et al., 2024; Nie et al., 2023), including residual connections, layer normalization, and feed-forward transformation, as illustrated in Figure 2(b). Starting from $\mathbf{H}^{(0)} = \mathrm{Linear}(\mathcal{X}) \in \mathbb{R}^{N\times D}$, BendingFormer stacks L_e encoder layers and produces $\mathbf{H}^{(L_e)} \in \mathbb{R}^{N\times D}$, progressively refining channel representations through geometry-aware cross-channel interaction.

4.4 FORECASTING HEAD

After geometry-aware cross-channel interaction in BendingFormer, the resulting hidden representation is mapped back to the original temporal space for forecasting. Specifically, $\mathbf{H}^{(L_e)} \in \mathbb{R}^{N\times D}$ is first decoded into an r-dimensional predictability-oriented representation through a linear projection, yielding $\mathcal{Z} = \mathrm{Linear}(\mathbf{H}^{(L_e)}) \in \mathbb{R}^{N\times r}$. Reconstructor (Section 4.2.3) then maps $\mathcal{Z}$ back to the original temporal space, producing $\mathbf{Z} \in \mathbb{R}^{N\times T}$. Finally, a feed-forward network composed of two linear layers with a GELU activation in between maps $\mathbf{Z}$ to the prediction horizon, yielding $\widehat{\mathbf{Y}} = \mathrm{FFN}(\mathbf{Z}) \in \mathbb{R}^{N\times\tau}$.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Baselines and Datasets We compare Navigator against 23 representative state-of-the-art forecasting models, including (i) Transformer-based methods: PatchLGA (Kim et al., 2026), TimeBridge (Liu et al., 2025), TQNet (Lin et al., 2025a), Leddam (Yu et al., 2024), CARD (Zhou et al., 2024), TimeXer (Wang et al., 2024b), iTransformer (Liu et al., 2024), Fredformer (Piao et al., 2024), PAttn (Tan et al., 2024), PatchTST (Nie et al., 2023), Crossformer (Zhang & Yan, 2023); (ii) Linear-based methods: DLinear (Zeng et al., 2023), TimeMixer (Wang et al., 2024a), RAFT (Han et al., 2025), SparseTSF (Lin et al., 2024b), CycleNet (Lin et al., 2024a); (iii) Frequency-based methods: FilterNet (Yi et al., 2024a), FreTS (Yi et al., 2023b), FITS (Xu et al., 2024) (iv) CNN-based methods: VPNet (Liu et al., 2026), TimesNet (Wu et al., 2023); (v) Different training frameworks: DistDF (Wang et al., 2026a), FreDF (Wang et al., 2025). To assess the effectiveness of our model, we conduct experiments on eight benchmark datasets for long-term forecasting, including four ETT subsets (Zhou et al., 2021), Electricity (Wu et al., 2021), Traffic (Wu et al., 2021), Solar-Energy (Zhou et al., 2021), and Weather (Wu et al., 2021), as well as four PEMS subsets (Liu et al., 2022a)

Table 3: Generalization performance of Foreseer across different models with both the lookback window and prediction horizon fixed at 96.

Backbones	iTransformer						PatchTST						PAttn					
Setups	Original		+ Foreseer		Imporve		Original		+ Foreseer		Imporve		Original		+ Foreseer		Imporve	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	0.327	0.363	0.318	0.357	2.75%	1.65%	0.326	0.364	0.320	0.358	1.84%	1.65%	0.331	0.369	0.327	0.365	1.21%	1.08%
ETTm2	0.189	0.270	0.179	0.260	5.29%	3.70%	0.177	0.262	0.172	0.255	2.82%	2.67%	0.181	0.263	0.174	0.256	3.87%	2.66%
ETTh1	0.386	0.403	0.386	0.403	-	-	0.381	0.404	0.371	0.391	2.62%	3.22%	0.390	0.405	0.373	0.392	4.36%	3.21%
ETTh2	0.299	0.350	0.293	0.343	2.01%	2.00%	0.295	0.351	0.295	0.346	-	1.42%	0.309	0.358	0.296	0.344	4.21%	3.91%
Weather	0.178	0.219	0.161	0.206	9.55%	5.94%	0.176	0.217	0.166	0.210	5.68%	3.23%	0.173	0.215	0.166	0.210	4.05%	2.33%
Electricity	0.149	0.242	0.137	0.234	8.05%	3.31%	0.164	0.256	0.157	0.253	4.27%	1.17%	0.182	0.269	0.168	0.261	7.69%	2.97%
Solar-Energy	0.214	0.246	0.203	0.247	5.14%	-	0.241	0.282	0.236	0.284	2.07%	-	0.222	0.262	0.213	0.262	4.05%	-
PEMSD3	0.168	0.277	0.139	0.224	17.26%	19.13%	0.272	0.374	0.242	0.336	11.03%	10.16%	0.426	0.460	0.351	0.403	17.61%	12.39%
PEMSD4	0.151	0.263	0.133	0.226	11.92%	14.07%	0.296	0.392	0.208	0.312	29.73%	20.41%	0.555	0.528	0.378	0.430	31.89%	18.56%

for short-term forecasting. Additional details on the baselines and datasets are presented in Appendix B.1 and Appendix B.2, while the implementation settings are described in Appendix B.3.

5.2 Main Results

As shown in Table 1, Navigator achieves strong performance across both long-term and short-term forecasting benchmarks, obtaining the best result in 19 out of 24 dataset–metric comparisons. The robustness results in Table 7, based on five runs with different random seeds, further show that these improvements are stable rather than arising from favorable initialization. Such consistent gains across datasets with diverse temporal dynamics and channel structures highlight the effectiveness of the two key designs in Navigator. Foreseer emphasizes temporal directions with stronger predictive relevance, reducing the reliance on indiscriminately modeling the entire historical window. Meanwhile, BendingFormer makes the value representation part of the interaction criterion itself by evaluating the directional compatibility of the geometric bending jointly formed by the query, key, and value representations, rather than relying solely on query–key similarity. Together, these two components provide a more targeted treatment of both temporal representation and cross-channel interaction, which contributes to the strong and robust performance of Navigator across different forecasting scenarios.

5.3 Model Analysis

Table 2: Ablation studies of Foreseer. The reported results are averaged over four prediction horizons. Detailed results are provided in Table 10.

Module variants	Weather		Electricity		Solar-Energy		Traffic		PEMSD3		PEMSD4	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
***w* Foreseer**	**0.237**	**0.259**	**0.161**	**0.250**	**0.192**	**0.212**	**0.440**	**0.249**	**0.091**	**0.193**	**0.081**	**0.180**
w/o Foreseer	0.242	0.263	0.165	0.253	0.203	0.220	0.446	0.252	0.106	0.211	0.102	0.205
w/o Deltas	0.241	0.262	0.165	0.253	0.204	0.222	0.448	0.252	0.111	0.217	0.108	0.213
r Linear	0.243	0.263	0.166	0.255	0.202	0.222	0.450	0.254	0.106	0.212	0.103	0.206
r Channel-wise	0.240	0.263	0.181	0.267	0.194	0.219	0.500	0.303	0.118	0.224	0.097	0.200

Ablation Studies of Foreseer. As shown in Table 2, removing Foreseer (*w/o* Foreseer) consistently degrades performance across all six datasets, confirming the benefit of transforming historical observations according to their predictive relevance before subsequent modeling. Replacing the Predictability-Basis-guided transformations with standard learnable linear layers (*r Linear*) also leads to clear performance drops, indicating that the gains of Foreseer do not simply arise from an additional projection layer, but from the predictive structure derived from the training data. Removing the bias terms $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ (*w/o Deltas*) further deteriorates performance, supporting their role in adapting the shared transformation to channel-specific characteristics. Finally, constructing an independent basis for each channel (*r Channel-wise*) is consistently inferior to the shared Predictability Basis, with particularly pronounced degradation on Electricity and Traffic datasets. This suggests that aggregating predictive structure across channels provides a more stable and generalizable temporal basis than estimating each channel independently. Overall, these results validate both the predictability-oriented design of Foreseer and the specific choice of a shared basis with channel-adaptive bias terms.

Foreseer as Plug-in. Table 3 evaluates the generalizability of Foreseer by integrating it into three different forecasting backbones. Across iTransformer (Liu et al., 2024), PatchTST (Nie et al., 2023), and PAttn (Tan et al., 2024), Foreseer improves performance on the vast majority of dataset–metric combinations, indicating that its benefit is not tied to the specific architecture of Navigator. The gains are especially pronounced on the PEMS datasets, where MSE reductions exceed 10% for all three backbones and reach nearly 30% on PEMS04 for PatchTST and PAttn. Although a few cases show no improvement or slight degradation, this may arise from a mismatch between the data-derived temporal prior introduced by Foreseer and the representation characteristics of certain backbones. Nevertheless, these results suggest that the Predictability Basis captures transferable predictive structure from the training data and can serve as a model-agnostic temporal prior, enhancing different forecasting architectures without modifying their core modeling mechanisms.

Ablation Studies of Geometric Bending Attention. As shown in Table 4, we replace GBA with four representative attention mechanisms: Full Attention (Vaswani et al., 2017) (*r* Full Attn.), ProbSparse Attention (Zhou et al., 2021) (*r* Prob Attn.), LSH Attention (Kitaev et al., 2020) (*r* LSH Attn.), and Flow Attention (Wu et al., 2022) (*r* Flow Attn.). Replacing GBA generally degrades forecasting performance across the evaluated datasets, demonstrating the effectiveness of its geometry-aware interaction criterion. Unlike conventional attention scoring based solely on query–key relations, GBA incorporates the value representation into the scoring process through the joint query–key–value geometry. This allows the model to assess not only the relation between query and key representations, but also whether the information carried by the value is directionally compatible with the corresponding interaction. The resulting gains suggest that effective cross-channel modeling should consider not only pairwise relationships, but also the geometric compatibility of the information to be aggregated.

Table 4: Ablation studies of Geometric Bending Attention (GBA), where we replace GBA with other attention mechanisms. The reported results are averaged over four prediction horizons. Detailed results are provided in Table 11.

Datasets	ETTm1		Weather		Solar-Energy		Traffic	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GBA (ours)	**0.373**	**0.376**	**0.237**	**0.259**	**0.192**	0.212	**0.440**	**0.249**
r Full Attn.	0.375	0.377	0.243	0.264	0.213	0.221	0.448	0.250
r Prob Attn.	0.378	0.381	0.242	0.263	**0.192**	**0.211**	0.445	0.251
r LSH Attn.	0.378	0.381	0.241	0.262	0.195	0.214	0.451	0.250
r Flow Attn.	0.379	0.382	0.240	0.262	0.214	0.219	0.448	0.251

5.4 Efficiency Analysis

To jointly assess forecasting accuracy and computational efficiency, we conduct efficiency experiments on the Electricity dataset with 321 channels, as summarized in Table 5. For a fair comparison, all models are evaluated with a batch size of 32 under the same training framework, except DistDF and FreDF, which follow their original training pipelines. From the results, Navigator achieves a favorable balance between predictive performance and computational cost. Compared with lightweight models such as FilterNet and RAFT, it incurs only moderate increases in parameter count and MACs while delivering substantially better forecasting accuracy. Meanwhile, compared with Transformer-based models such as TimeXer, iTransformer, and PatchTST, Navigator achieves stronger predictive performance with lower overall computational overhead, resulting in a more favorable accuracy-efficiency trade-off. Additional efficiency results on the ETTm1 dataset (7 channels) and the Weather dataset (21 channels) are reported in Tables 15 and 16, respectively, further demonstrating the efficiency of Navigator across datasets with different channel scales.

Table 5: Efficiency comparison between Navigator and other models on the Electricity dataset with lookback window T = 96 and prediction horizon τ = 96. Training Time denotes the average time required per epoch for the model.

Model	Parameters	MACs	Training Time	Memory Usage	MSE	MAE
PatchLGA	42.13 K	105.54 M	21.4 s	2908 MB	0.149	0.241
DistDF	8.62 M	11.62 G	103.7 s	8528 MB	0.137	0.235
VPNet	8.2 M	145.8 M	27.3 s	2270 MB	0.135	0.224
RAFT	44.26 K	14.21 M	10.3 s	1476 MB	0.152	0.257
FreDF	4.83 M	1.55 G	18.6 s	4256 MB	0.144	0.233
FilterNet	680.67 K	201.03 M	9.1 s	1894 MB	0.147	0.245
FreTS	3.24 M	1.02 G	42.7 s	10168 MB	0.189	0.277
iTransformer	4.83 M	1.55 G	17.3 s	3736 MB	0.148	0.240
PatchTST	5.33M	18.45 G	121.3 s	11540 MB	0.181	0.270
TimeXer	8.46 M	11.79 G	82.5 s	9180 MB	0.140	0.242
Navigator(ours)	5.0 M	2.06G	15.8s	3140MB	0.133	0.223

6 Conclusion and Limitation

This work advocates a shift from model-centric forecasting toward a view in which informative structure is first extracted from the data and then used to guide model computation. Through *Navigator*, we show that forecasting can benefit from structure at two levels: predictive organization within the historical observation space and geometry-aware interaction among channels. Experimental results suggest that effective forecasting depends not only on model capacity, but also on what information is emphasized and how information exchange is structured. More broadly, these findings support the principle that *insight precedes action*: uncovering useful structure from the data can provide a stronger foundation for prediction than relying on end-to-end optimization alone.

Despite these results, several limitations remain. The Predictability Basis captures predictive structure through second-order statistics and is fixed after being estimated from the training data, which may limit its ability to represent nonlinear dependencies or adapt to distribution shifts. Meanwhile, GBA models cross-channel interaction through individual query–key–value configurations without explicitly capturing higher-order relations among multiple channels. These limitations motivate future work on nonlinear and adaptive predictive priors, as well as higher-order geometric mechanisms for cross-channel modeling.

AI USE STATEMENT

In this work, generative AI tools were employed solely to assist with grammar correction, spelling correction, and language refinement. No generative AI tools were used for other research-related activities, such as formulating research ideas, developing methodologies or experimental designs, implementing methods, performing theoretical analysis, interpreting experimental results, or producing scientific content. All AI-assisted revisions were carefully examined by the authors, who assume full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses solely on time series forecasting and uses only publicly available benchmark datasets, without involving human-subject research or the collection of personally identifiable information. Therefore, we are not aware of any significant ethical risks or concerns associated with this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed methods in the main text, with implementation details and dataset descriptions included in the appendix. The source code is submitted together with the supplementary materials and will be made publicly available upon publication.

REFERENCES

Leo Breiman. *Probability*. SIAM, 1992.

Fanpu Cao, Lu Dai, Jindong Han, and Hui Xiong. Enhancing multivariate time series forecasting with global temporal retrieval. *arXiv preprint arXiv:2602.10847*, 2026.

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

Haotian Gao, Zheng Dong, Jiawei Yong, Shintaro Fukushima, Kenjiro Taura, and Renhe Jiang. How different from the past? spatio-temporal time series forecasting with self-supervised deviation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=TgGH1bY6kl.

Robert M. Gray. *Probability, Random Processes, and Ergodic Properties*, volume None of *Springer Books*. Springer, none edition, June 1988. doi: 10.1007/978-1-4757-2024-2. URL https://ideas.repec.org/b/spr/sprbok/978-1-4757-2024-2.html.

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

Trevor Hastie. Ridge regularization: An essential concept in data science. *Technometrics*, 62(4): 426–433, 2020.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Dongbin Kim, Youngjoo Park, Woojin Jeong, and Jaewook Lee. Local geometry attention for time series forecasting under realistic corruptions. In *The Fourteenth International Conference on Learning Representations*, 2026.

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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024a.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. Sparsetsf: Modeling long-term time series forecasting with 1k parameters. *arXiv preprint arXiv:2405.00946*, 2024b.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. *arXiv preprint arXiv:2505.12917*, 2025a.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Feiyu Zhao, Ruichao Mo, and Haotong Zhang. Segrnn: Segment recurrent neural network for long-term time series forecasting. *IEEE Internet of Things Journal*, 2025b.

Kun Liu, Renjun Jia, Ruifeng Yang, Xirui Zeng, Yuqi Liang, and Cen Chen. Are global dependencies necessary? scalable time series forecasting via local cross-variate modeling. In *International Conference on Learning Representations*, volume 2026, pp. 102794–102817, 2026.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in neural information processing systems*, 35:5816–5828, 2022a.

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. *International Conference on Machine Learning*, 2025.

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

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

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Beresford N Parlett. *The symmetric eigenvalue problem*. SIAM, 1998.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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.

Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Junkai Lu, Jilin Hu, Chenjuan Guo, Christian S Jensen, and Bin Yang. A comprehensive survey of deep learning for multivariate time series forecasting: A channel strategy perspective. *arXiv preprint arXiv:2502.10721*, 2025.

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

Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

J Michael Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, 37:60162–60191, 2024.

Lloyd N Trefethen and David Bau, III. *Numerical linear algebra*. SIAM, 1997.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hao Wang, Lichen Pan, Yuan Shen, Zhichao Chen, Degui Yang, Yifei Yang, Sen Zhang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. Fredf: Learning to forecast in the frequency domain. In *International Conference on Learning Representations*, volume 2025, pp. 6893–6922, 2025.

Hao Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zi Ciu Chan, Qingsong Wen, Haoxuan Li, and Zhouchen Lin. DistDF: Time-series forecasting needs joint-distribution wasserstein alignment. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=VrdLwUmzBy.

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

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Yunzhong Qiu, Haoran Zhang, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 2024b.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b.

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

Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. *arXiv preprint arXiv:2202.06258*, 2022.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Zhijian Xu, Ailing Zeng, and Qiang Xu. Fits: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, volume 2024, pp. 26295–26318, 2024.

Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. *Advances in neural information processing systems*, 36:69638–69660, 2023a.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36:76656–76679, 2023b.

Kun Yi, Jingru Fei, Qi Zhang, Hui He, Shufeng Hao, Defu Lian, and Wei Fan. Filternet: Harnessing frequency filters for time series forecasting. *Advances in Neural Information Processing Systems*, 37:55115–55140, 2024a.

Kun Yi, Qi Zhang, Hui He, Kaize Shi, Liang Hu, Ning An, and Zhendong Niu. Deep coupling network for multivariate time series forecasting. *ACM Transactions on Information Systems*, 42 (5):1–28, 2024b.

Guoqi Yu, Jing Zou, Xiaowei Hu, Angelica I Aviles-Rivero, Jing Qin, and Shujun Wang. Revitalizing multivariate time series forecasting: Learnable decomposition with inter-series dependencies and intra-series variations modeling. *arXiv preprint arXiv:2402.12694*, 2024.

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*, volume 37, pp. 11121–11128, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Heng Zhou, Xin Sun, and Chao Li. Anti-aliasing matters: A dynamic network for time series forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=fvfJWiTrfF`.

Tian Zhou, Qingsong Wen, Jinyang Gao, Bolin Ding, Rong Jin, et al. Card: Channel aligned robust blend transformer for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11245–11283, 2024.

APPENDIX CONTENTS

A Theoretical Supplement **15**

A.1 Proof of Proposition 1 15

A.2 Discussion on Linear Temporal Directions 17

B Experimental Details **18**

B.1 Baselines 18

B.2 Datasets 20

B.3 Implementation Details 21

C Additional Results **21**

C.1 Full Forecasting Results 21

C.2 Full Ablation Results 21

C.3 Comparison of Cross-Channel Modeling Methods 22

C.4 Extending Geometric Bending Attention to Temporal Modeling 24

C.5 Proportion of Training Data Used to Compute the Predictability Basis 25

C.6 More Efficiency Analysis 25

C.7 Parameter Sensitivity Analysis 26

D Analysis of the Predictability Basis **30**

D.1 Effect of Predictability Basis Dimension r 30

D.2 Visualizations of Predictability Basis $\mathbf{M}_r$ 31

E Visualizations of Predictions **31**

A Theoretical Supplement

A.1 Proof of Proposition 1

Proposition 1 (Closed-form Characterization of Temporal Predictability)**.** *Let $\mathbf{p}_t \in \mathbb{R}^T$ and $\mathbf{f}_t \in \mathbb{R}^\tau$ be centered random vectors with finite second moments, and define $\Sigma_{pp} = \mathbb{E}[\mathbf{p}_t\mathbf{p}_t^\top]$, $\Sigma_{ff} = \mathbb{E}[\mathbf{f}_t\mathbf{f}_t^\top]$, and $\Sigma_{pf} = \mathbb{E}[\mathbf{p}_t\mathbf{f}_t^\top]$, with $\Sigma_{fp} = \Sigma_{pf}^\top$. Assume that Σ_{pp} and Σ_{ff} are positive definite. Then:*

(i) Closed-form Characterization. *For any nonzero temporal direction $\mathbf{w} \in \mathbb{R}^T \setminus \{\mathbf{0}\}$, its temporal predictability satisfies*

$$\mathcal{P}(\mathbf{w})^2 = \frac{\mathbf{w}^\top \Sigma_{pf}\Sigma_{ff}^{-1}\Sigma_{fp}\mathbf{w}}{\mathbf{w}^\top \Sigma_{pp}\mathbf{w}}. \tag{A.1}$$

(ii) Generalized Eigenvalue Formulation. *Since Eq. (A.1) is a generalized Rayleigh quotient, its stationary temporal directions, along which its value has zero first-order variation, satisfy the generalized eigenvalue problem*

$$\Sigma_{pf}\Sigma_{ff}^{-1}\Sigma_{fp}\mathbf{w} = \lambda\Sigma_{pp}\mathbf{w}. \tag{A.2}$$

For any generalized eigenpair $(\lambda, \mathbf{w})$ satisfying Eq. (A.2), the corresponding temporal direction satisfies $\lambda = \mathcal{P}(\mathbf{w})^2$. Therefore, larger generalized eigenvalues correspond to temporal directions with higher predictability, with the largest generalized eigenvalue identifying the most predictive temporal directions.

Proof. For any nonzero temporal directions $\mathbf{w} \in \mathbb{R}^T \setminus \{\mathbf{0}\}$ and $\mathbf{v} \in \mathbb{R}^\tau \setminus \{\mathbf{0}\}$, since $\mathbf{p}_t$ and $\mathbf{f}_t$ are centered, we have $\mathrm{Var}(\mathbf{w}^\top\mathbf{p}_t) = \mathbf{w}^\top\Sigma_{pp}\mathbf{w}$, $\mathrm{Var}(\mathbf{v}^\top\mathbf{f}_t) = \mathbf{v}^\top\Sigma_{ff}\mathbf{v}$ and $\mathrm{Cov}\left(\mathbf{w}^\top\mathbf{p}_t, \mathbf{v}^\top\mathbf{f}_t\right) = \mathbf{w}^\top\Sigma_{pf}\mathbf{v}$. Since $\Sigma_{pp} \succ 0$ and $\Sigma_{ff} \succ 0$, both projected variables have strictly positive variance for any nonzero $\mathbf{w}$ and $\mathbf{v}$. Therefore,

$$\mathrm{corr}\left(\mathbf{w}^\top\mathbf{p}_t, \mathbf{v}^\top\mathbf{f}_t\right)^2 = \frac{\left(\mathbf{w}^\top\Sigma_{pf}\mathbf{v}\right)^2}{\left(\mathbf{w}^\top\Sigma_{pp}\mathbf{w}\right)\left(\mathbf{v}^\top\Sigma_{ff}\mathbf{v}\right)}. \tag{A.3}$$

Since the square function is monotonically increasing on $[0, \infty)$, Definition 1 and Eq. (A.3) give, for any fixed nonzero $\mathbf{w}$,

$$\mathcal{P}(\mathbf{w})^2 = \frac{1}{\mathbf{w}^\top\Sigma_{pp}\mathbf{w}} \max_{\mathbf{v}\neq\mathbf{0}} \frac{\left(\mathbf{w}^\top\Sigma_{pf}\mathbf{v}\right)^2}{\mathbf{v}^\top\Sigma_{ff}\mathbf{v}}. \tag{A.4}$$

Because $\Sigma_{ff} \succ 0$, its symmetric square root $\Sigma_{ff}^{1/2}$ and inverse square root $\Sigma_{ff}^{-1/2}$ are well defined. Let $\mathbf{u} = \Sigma_{ff}^{1/2}\mathbf{v}$, so that $\mathbf{v} = \Sigma_{ff}^{-1/2}\mathbf{u}$. Then

$$\mathbf{v}^\top\Sigma_{ff}\mathbf{v} = \mathbf{u}^\top\mathbf{u}, \tag{A.5}$$

and

$$\mathbf{w}^\top\Sigma_{pf}\mathbf{v} = \left(\Sigma_{ff}^{-1/2}\Sigma_{fp}\mathbf{w}\right)^\top\mathbf{u}. \tag{A.6}$$

Hence,

$$\frac{\left(\mathbf{w}^\top\Sigma_{pf}\mathbf{v}\right)^2}{\mathbf{v}^\top\Sigma_{ff}\mathbf{v}} = \frac{\left[\left(\Sigma_{ff}^{-1/2}\Sigma_{fp}\mathbf{w}\right)^\top\mathbf{u}\right]^2}{\mathbf{u}^\top\mathbf{u}}. \tag{A.7}$$

By the Cauchy–Schwarz inequality (Steele, 2004),

$$\left[\left(\Sigma_{ff}^{-1/2}\Sigma_{fp}\mathbf{w}\right)^\top\mathbf{u}\right]^2 \leq \left\|\Sigma_{ff}^{-1/2}\Sigma_{fp}\mathbf{w}\right\|_2^2 \|\mathbf{u}\|_2^2. \tag{A.8}$$

Therefore,

$$\frac{\left(\mathbf{w}^\top\Sigma_{pf}\mathbf{v}\right)^2}{\mathbf{v}^\top\Sigma_{ff}\mathbf{v}} \leq \left\|\Sigma_{ff}^{-1/2}\Sigma_{fp}\mathbf{w}\right\|_2^2 = \mathbf{w}^\top\Sigma_{pf}\Sigma_{ff}^{-1}\Sigma_{fp}\mathbf{w}. \tag{A.9}$$

If $\Sigma_{fp}\mathbf{w} \neq \mathbf{0}$, equality in Eq. (A.9) is attained when

$$\mathbf{u} \propto \Sigma_{ff}^{-1/2}\Sigma_{fp}\mathbf{w}, \tag{A.10}$$

which is equivalent to

$$\mathbf{v} \propto \Sigma_{ff}^{-1}\Sigma_{fp}\mathbf{w}. \tag{A.11}$$

If $\Sigma_{fp}\mathbf{w} = \mathbf{0}$, the numerator in Eq. (A.4) is zero for every nonzero $\mathbf{v}$, and the maximum is therefore also zero. Consequently,

$$\max_{\mathbf{v}\neq\mathbf{0}} \frac{\left(\mathbf{w}^\top\Sigma_{pf}\mathbf{v}\right)^2}{\mathbf{v}^\top\Sigma_{ff}\mathbf{v}} = \mathbf{w}^\top\Sigma_{pf}\Sigma_{ff}^{-1}\Sigma_{fp}\mathbf{w}. \tag{A.12}$$

Substituting this result into Eq. (A.4) yields

$$\mathcal{P}(\mathbf{w})^2 = \frac{\mathbf{w}^\top\Sigma_{pf}\Sigma_{ff}^{-1}\Sigma_{fp}\mathbf{w}}{\mathbf{w}^\top\Sigma_{pp}\mathbf{w}}, \tag{A.13}$$

which proves part (i).

We next prove part (ii). Define

$$A = \Sigma_{pf}\Sigma_{ff}^{-1}\Sigma_{fp}, \qquad B = \Sigma_{pp}. \tag{A.14}$$

Since Σ_{ff}^{-1} is symmetric positive definite, A is symmetric positive semidefinite, because for any $\mathbf{x} \in \mathbb{R}^T$,

$$\mathbf{x}^\top A\mathbf{x} = (\Sigma_{fp}\mathbf{x})^\top\Sigma_{ff}^{-1}(\Sigma_{fp}\mathbf{x}) \geq 0, \tag{A.15}$$

while B is symmetric positive definite. From part (i),

$$\mathcal{P}(\mathbf{w})^2 = R(\mathbf{w}) = \frac{\mathbf{w}^\top A\mathbf{w}}{\mathbf{w}^\top B\mathbf{w}}, \tag{A.16}$$

which is a generalized Rayleigh quotient (Horn & Johnson, 2012; Parlett, 1998).

Differentiating $R(\mathbf{w})$ with respect to $\mathbf{w}$ gives

$$\nabla_{\mathbf{w}}R(\mathbf{w}) = \frac{2\left[(\mathbf{w}^\top B\mathbf{w})A\mathbf{w} - (\mathbf{w}^\top A\mathbf{w})B\mathbf{w}\right]}{(\mathbf{w}^\top B\mathbf{w})^2}. \tag{A.17}$$

At a stationary temporal direction, $\nabla_{\mathbf{w}}R(\mathbf{w}) = \mathbf{0}$. Hence,

$$(\mathbf{w}^\top B\mathbf{w})A\mathbf{w} = (\mathbf{w}^\top A\mathbf{w})B\mathbf{w}. \tag{A.18}$$

Since $B \succ 0$, we have $\mathbf{w}^\top B\mathbf{w} > 0$ for every nonzero $\mathbf{w}$. Therefore,

$$A\mathbf{w} = \frac{\mathbf{w}^\top A\mathbf{w}}{\mathbf{w}^\top B\mathbf{w}}B\mathbf{w} = R(\mathbf{w})B\mathbf{w}. \tag{A.19}$$

Thus every stationary temporal direction satisfies

$$A\mathbf{w} = \lambda B\mathbf{w}, \tag{A.20}$$

where

$$\lambda = R(\mathbf{w}) = \mathcal{P}(\mathbf{w})^2. \tag{A.21}$$

Substituting the definitions of A and B yields

$$\Sigma_{pf}\Sigma_{ff}^{-1}\Sigma_{fp}\mathbf{w} = \lambda\Sigma_{pp}\mathbf{w}, \tag{A.22}$$

which is Eq. (3).

Conversely, let $(\lambda, \mathbf{w})$ be any generalized eigenpair satisfying Eq. (3), with $\mathbf{w} \neq \mathbf{0}$. Left-multiplying by $\mathbf{w}^\top$ gives

$$\mathbf{w}^\top A\mathbf{w} = \lambda\mathbf{w}^\top B\mathbf{w}. \tag{A.23}$$

Since $\mathbf{w}^\top B\mathbf{w} > 0$,

$$\lambda = \frac{\mathbf{w}^\top A\mathbf{w}}{\mathbf{w}^\top B\mathbf{w}} = \mathcal{P}(\mathbf{w})^2. \tag{A.24}$$

Hence the generalized eigenvalue associated with each generalized eigenvector equals the squared temporal predictability of the corresponding temporal direction.

Finally, we show that the largest generalized eigenvalue identifies the most predictive temporal directions. Since $B \succ 0$, let

$$\mathbf{z} = B^{1/2}\mathbf{w}, \tag{A.25}$$

so that $\mathbf{w} = B^{-1/2}\mathbf{z}$. Then Eq. (A.16) becomes

$$R(\mathbf{w}) = \frac{\mathbf{z}^\top B^{-1/2} A B^{-1/2} \mathbf{z}}{\mathbf{z}^\top \mathbf{z}}. \tag{A.26}$$

Define

$$C = B^{-1/2} A B^{-1/2}. \tag{A.27}$$

Since A is symmetric positive semidefinite and $B \succ 0$, C is symmetric positive semidefinite. By the Rayleigh–Ritz theorem,

$$\max_{\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^\top C \mathbf{z}}{\mathbf{z}^\top \mathbf{z}} = \lambda_{\max}(C). \tag{A.28}$$

Moreover, under the change of variables $\mathbf{z} = B^{1/2}\mathbf{w}$, the generalized eigenvalue problem

$$A\mathbf{w} = \lambda B\mathbf{w} \tag{A.29}$$

is equivalent to

$$C\mathbf{z} = \lambda \mathbf{z}. \tag{A.30}$$

Hence, the eigenvalues of C are exactly the generalized eigenvalues of the matrix pair (A, B). Therefore,

$$\max_{\mathbf{w}\neq\mathbf{0}} \mathcal{P}(\mathbf{w})^2 = \lambda_{\max}, \tag{A.31}$$

where $\lambda_{\max}$ denotes the largest generalized eigenvalue of Eq. (3). Since $\mathcal{P}(\mathbf{w}) \geq 0$, maximizing $\mathcal{P}(\mathbf{w})^2$ is equivalent to maximizing $\mathcal{P}(\mathbf{w})$. Thus the generalized eigenvectors associated with $\lambda_{\max}$ identify the most predictive temporal directions.

This completes the proof. □

A.2 Discussion on Linear Temporal Directions

In our predictability analysis, a temporal direction is represented by a linear projection $\mathbf{w}^\top \mathbf{p}_t$ of the historical observations. This choice should not be interpreted as assuming that the relationship between historical and future observations is inherently linear. Rather, our objective is to extract a simple and principled data-derived prior from the training data before nonlinear forecasting is performed.

From a geometric perspective, $\mathbf{p}_t \in \mathbb{R}^T$ lies in the historical observation space, and a vector $\mathbf{w} \in \mathbb{R}^T$ naturally defines a direction in this space. The scalar $\mathbf{w}^\top \mathbf{p}_t$ therefore measures the projection of the historical observations onto this temporal direction. Such a formulation enables us to directly characterize how informative different directions in the historical observation space are about future observations. More importantly, restricting the analysis to linear projections leads to an analytically tractable characterization of temporal predictability. As shown in Proposition 1, the predictability of a temporal direction can be expressed as a generalized Rayleigh quotient, allowing the directions to be identified and ordered through a generalized eigenvalue problem. Consequently, the Predictability Basis can be computed directly from the training data without introducing an additional neural network or optimization objective, and can remain fixed during subsequent model training.

This linear analysis serves only to construct the Predictability Basis rather than to restrict the forecasting model itself. After the historical observations are transformed by Foreseer, subsequent representation learning and cross-channel interaction are performed by the learnable forecasting model. Therefore, the proposed formulation uses linear temporal directions as a tractable means of revealing predictive structure in the historical observation space, while still allowing nonlinear representations and interactions to be learned during forecasting.

B EXPERIMENTAL DETAILS

B.1 BASELINES

To comprehensively evaluate Navigator, we compare it with 23 representative and state-of-the-art forecasting methods spanning diverse architectural paradigms, including Transformer-based, linear-based, frequency-based, and CNN-based approaches, as well as two recently proposed training frameworks. A brief description of each baseline is provided below:

PatchLGA (Kim et al., 2026) is a geometry-aware variant of PatchTST that replaces the standard attention mechanism with Local Geometry Attention (LGA), allowing the model to capture intrinsic temporal structures while improving robustness to corruptions. Implementation details are available at `https://github.com/dongbeank/LGA`. The results of PatchLGA in Table 7 are obtained by running the official implementation five times and reporting the average performance.

TimeBridge (Liu et al., 2025) bridges non-stationarity and dependency modeling in multivariate time series forecasting through Integrated Attention and Cointegrated Attention for short-term and long-term dynamics, respectively. Implementation details are available at `https://github.com/Hank0626/TimeBridge`. The results of TimeBridge in Table 8 follow those reported in Table 6 of the DistDF (Wang et al., 2026a) paper, except for the Solar-Energy, PEMSD4, and PEMSD7 datasets, for which the results are obtained by running the official code five times and reporting the average values.

TQNet Lin et al. (2025a) leverages periodically shifted learnable temporal queries to effectively capture global and sample-specific cross-channel dependencies for multivariate time series forecasting. Implementation details are available at `https://github.com/ACAT-SCUT/TQNet`. The results of TQNet in Table 12 follow those reported in Table 5 of the original TQNet paper.

Leddam Yu et al. (2024) combines learnable trend decomposition with dual attention to jointly capture inter-channel dependencies and temporal variations for multivariate time series forecasting. Implementation details are available at `https://github.com/Levi-Ackman/Leddam`. The results of Leddam in Table 12 follow those reported in Table 15 of the original Leddam paper.

CARD Zhou et al. (2024) employs channel-aligned attention and multi-scale token blending to capture temporal correlations and cross-channel dependencies for time series forecasting. Implementation details are available at `https://github.com/wxie9/CARD`. The results of CARD in Table 12 follow those reported in Table 7 of the original CARD paper, except for the Solar-Energy dataset, for which we run the official code five times and report the average results.

TimeXer (Wang et al., 2024b) is a Transformer-based method for time series forecasting with exogenous variables, which jointly models endogenous temporal dependencies and exogenous correlations through patch-wise self-attention and variate-wise cross-attention. Implementation details are available at `https://github.com/thuml/TimeXer`. The results of TimeXer in Table 7 follow those reported in Table 12 of the original TimeXer paper, except for the Solar-Energy and four PEMS datasets, for which the results are obtained by running the official code five times and reporting the average values.

iTransformer (Liu et al., 2024) inverts the conventional Transformer architecture by applying attention across variates rather than time steps, improving multivariate dependency modeling. Implementation details are available at `https://github.com/thuml/iTransformer`. The results of iTransformer in Table 7 follow those reported in Table 9 and Table 10 of the original iTransformer paper. And the results of iTransformer in Table 9 follow those reported in Table 13 of the original DMANet (Zhou et al., 2026) paper.

Fredformer (Piao et al., 2024) mitigates frequency bias by promoting balanced learning across different frequency bands, preventing dominant high-energy components from overwhelming informative low-amplitude features. Implementation details are available at `https://github.com/chenzRG/Fredformer`. The results of Fredformer in Table 8 follow those reported in Table 12 of the original Fredformer paper, except for the four PEMS datasets, for which the results are obtained by running the official code five times and reporting the average values.

PAttn (Tan et al., 2024) combines patch-wise tokenization with a single self-attention layer and achieves forecasting performance comparable to several LLM-based forecasting

methods. Implementation details are available at https://github.com/BennyTMT/LLMsForTimeSeries. The results of PAttn in Table 8 are obtained by running the official code five times and reporting the average values.

PatchTST (Nie et al., 2023) introduces patch-based tokenization and channel-independent Transformer modeling, which reduce computational complexity, extend the effective look-back window, and improve the learning of temporal representations for time series forecasting. Implementation details are available at https://github.com/yuqinie98/patchtst. The results of PatchTST in Table 7 follow those reported in Table 9 and Table 10 of the original iTransformer (Liu et al., 2024) paper. And the results of PatchTST in Table 9 follow those reported in Table 13 of the original DMANet (Zhou et al., 2026) paper.

Crossformer (Zhang & Yan, 2023) enhances multivariate time series forecasting by jointly modeling cross-time and cross-dimension dependencies through a two-stage attention mechanism and hierarchical multi-scale representation learning. Implementation details are available at https://github.com/Thinklab-SJTU/Crossformer. The results of Crossformer in Table 8 follow those reported in Table 9 and Table 10 of the original iTransformer (Liu et al., 2024) paper.

DLinear (Zeng et al., 2023) models long-term temporal dependencies through direct linear mappings, providing a simple yet highly competitive alternative to Transformer-based forecasting models. Implementation details are available at https://github.com/cure-lab/LTSF-Linear. The results of DLinear in Table 8 follow those reported in Table 9 and Table 10 of the original iTransformer (Liu et al., 2024) paper.

TimeMixer (Wang et al., 2024a) models complex temporal variations through multiscale decomposition and mixing, enabling the aggregation of fine-grained seasonal patterns and coarse-grained trend information for accurate forecasting. Implementation details are available at https://github.com/kwuking/TimeMixer. The results of TimeMixer in Table 8 follow those reported in Table 13 of the original TimeMixer paper, except for the four PEMS datasets, for which the results are obtained by running the official code five times and reporting the average values. And the results of TimeMixer in Table 9 follow those reported in Table 13 of the original DMANet (Zhou et al., 2026) paper.

RAFT (Han et al., 2025) enhances time series forecasting through retrieval augmentation, leveraging future trajectories of historically similar patterns as external inductive biases for prediction. Implementation details are available at https://github.com/archon159/RAFT. The results of RAFT in Table 7 are obtained by running the official code five times and reporting the average values.

SparseTSF (Lin et al., 2024b) is a lightweight forecasting model that decouples periodicity and trend through cross-period sparse forecasting, enabling efficient long-term forecasting with minimal model complexity. Implementation details are available at https://github.com/lss-1138/SparseTSF. The results of SparseTSF in Table 9 follow those reported in Table 8 of the original CycleNet (Lin et al., 2024a) paper.

CycleNet (Lin et al., 2024a) introduces Residual Cycle Forecasting (RCF), which models shared periodic patterns with learnable recurrent cycles and forecasts the corresponding residual components. Implementation details are available at https://github.com/ACAT-SCUT/CycleNet. The results of CycleNet in Table 9 follow those reported in Table 8 of the original CycleNet paper.

FilterNet (Yi et al., 2024a) integrates signal processing principles into time series forecasting by employing learnable frequency filters for adaptive spectral shaping and temporal pattern extraction. Implementation details are available at https://github.com/aikunyi/FilterNet. The results of FilterNet in Table 7 follow those reported in Table 1 of the original FilterNet paper, except for the Solar-Energy and four PEMS datasets, for which the results are obtained by running the official code five times and reporting the average values.

FreTS (Yi et al., 2023b) reformulates time series forecasting in the frequency domain, enabling MLPs to exploit global spectral information and compact signal energy for effective forecasting. Implementation details are available at https://github.com/aikunyi/FreTS. The results of FreTS in Table 7 follow those reported in Table 5 of the original FreDF (Wang et al., 2025) paper, except for the Solar-Energy and four PEMS datasets, for which the results are obtained by running the official code five times and reporting the average values.

FITS (Xu et al., 2024) performs time series analysis through frequency-domain interpolation, achieving competitive forecasting performance with an extremely lightweight architecture. Implementation details are available at https://github.com/VEWOXIC/FITS. The results of FITS in Table 8 are obtained by running the official code five times and reporting the average values.

VPNet (Liu et al., 2026) improves multivariate time series forecasting by modeling locally sufficient cross-variate interactions through a variate–patch field representation and scalable convolutional dependency learning. Implementation details are available at https://github.com/iuaku/VPNet/. The results of VPNet in Table 7 follow those reported in Table 12 of the original VPNet paper, except for four PEMS datasets, whose results are obtained by running the official code five times and reporting the average values.

TimesNet (Wu et al., 2023) transforms 1D time series into 2D representations based on multiple periods, enabling effective modeling of intra-period and inter-period variations through parameter-efficient 2D convolutions. Implementation details are available at https://github.com/thuml/TimesNet. The results of TimesNet in Table 8 follow those reported in Table 5 of the original FreDF (Wang et al., 2025) paper, except for the Solar-Energy and four PEMS datasets, for which the results are obtained by running the official code five times and reporting the average values. And the results of TimesNet in Table 9 follow those reported in Table 13 of the original DMANet (Zhou et al., 2026) paper.

DistDF (Wang et al., 2026a) improves time series forecasting by aligning the conditional distributions of forecasts and labels through a Wasserstein discrepancy-based distribution matching objective. Implementation details are available at https://github.com/Master-PLC/DistDF. The results of DistDF in Table 7 follow those reported in Table 6 of the original DistDF paper, except for the Solar-Energy, PEMSD4, and PEMSD7 datasets, whose results are obtained by running the official code five times and reporting the average values.

FreDF (Wang et al., 2025) improves time series forecasting by mitigating label autocorrelation through frequency-domain forecasting objectives, thereby reducing the bias of direct forecast learning. Implementation details are available at https://github.com/Master-PLC/FreDF. The results of FreDF in Table 7 follow those reported in Table 5 of the original FreDF paper, except for the Solar-Energy and four PEMS datasets, for which the results are obtained by running the official code five times and reporting the average values.

B.2 DATASETS

We evaluate our method on eight widely used datasets for long-term forecasting and four PEMS datasets for short-term forecasting. The dataset statistics are summarized in Table 6, and brief descriptions of each dataset are provided below:

(1) The **ETT** (Electricity Transformer Temperature) datasets (Zhou et al., 2021) consist of operational records collected from two electricity transformers between 2016 and 2018. Each sample contains seven variables, including six load-related measurements and one oil temperature (OT) measurement. The datasets are divided into four subsets: ETTh1 and ETTh2 with hourly sampling frequency, and ETTm1 and ETTm2 with 15-minute sampling frequency.

(2) The **Electricity** dataset (Wu et al., 2021) consists of hourly electricity consumption measurements from 321 clients over the period from 2012 to 2014. The values represent electricity usage in kilowatt-hours (kWh) and are recorded at an hourly resolution.

(3) The **Traffic** dataset (Wu et al., 2021) contains hourly road occupancy measurements collected by 862 sensors deployed on the freeways of the San Francisco Bay Area between 2015 and 2016. Each variable corresponds to the occupancy rate recorded by an individual traffic sensor.

(4) The **Weather** dataset (Wu et al., 2021) contains 21 meteorological variables recorded at 10-minute intervals throughout 2020. The measurements were collected from the Weather Station of the Max Planck Institute for Biogeochemistry and include atmospheric indicators such as temperature, humidity, and visibility.

(5) The **Solar-Energy** dataset (Lai et al., 2018) consists of solar power production data gathered from 137 photovoltaic (PV) plants in 2006, with a sampling frequency of 10 minutes.

Table 6: Dataset statistics. *Dim* denotes the number of channels. *Prediction Length* specifies the forecasting horizon. *Dataset Size* reports the number of time points in the training, validation, and test sets, respectively. *Frequency* denotes the sampling interval.

Dataset	Domain	Dim	Prediction Length	Dataset Size	Frequency
ETTh1, ETTh2	Electricity	7	{96, 192, 336, 720}	(8545, 2881, 2881)	Hourly
ETTm1, ETTm2	Electricity	7	{96, 192, 336, 720}	(34465, 11521, 11521)	15min
Weather	Weather	21	{96, 192, 336, 720}	(36792, 5271, 10540)	10min
Electricity	Electricity	321	{96, 192, 336, 720}	(18317, 2633, 5261)	Hourly
Traffic	Transportation	862	{96, 192, 336, 720}	(12185, 1757, 3509)	Hourly
Solar-Energy	Energy	137	{96, 192, 336, 720}	(36601, 5161, 10417)	10min
PEMS03	Transportation	358	{12, 24, 48, 96}	(15617, 5135, 5135)	5min
PEMS04	Transportation	307	{12, 24, 48, 96}	(10172, 3375, 3375)	5min
PEMS07	Transportation	883	{12, 24, 48, 96}	(16911, 5622, 5622)	5min
PEMS08	Transportation	170	{12, 24, 48, 96}	(10690, 3548, 3548)	5min

(6) The **PEMS** datasets (Liu et al., 2022a) consist of four traffic network datasets, namely PEMS03, PEMS04, PEMS07, and PEMS08, derived from the Caltrans Performance Measurement System (PeMS) in California. Traffic observations are recorded at a 5-minute resolution.

B.3 Implementation Details

All experiments are implemented in PyTorch (Paszke et al., 2019) and conducted on a single NVIDIA RTX 4090 GPU with 24GB of memory. Model parameters are optimized using Adam (Kingma & Ba, 2014), with the learning rate selected from $\{1\mathrm{e}{-4}, 3\mathrm{e}{-4}, 5\mathrm{e}{-4}, 7\mathrm{e}{-4}, 1\mathrm{e}{-3}\}$. Following prior works such as CARD (Zhou et al., 2024) and OLinear (Yue et al., 2026), we adopt a weighted loss function for optimization. The number of training epochs is selected from $\{10, 30, 40, 50\}$, while the batch size is chosen from $\{16, 32, 64\}$. Early stopping is applied based on validation performance, with the patience selected from $\{5, 8, 10\}$. The hidden dimensions d_{model} and d_{ff}, used in both BendingFormer and the forecasting head, are selected from $\{256, 512\}$. The number of encoder layers in BendingFormer, denoted by L_{e}, is selected from $\{1, 2, 3\}$.

C Additional Results

C.1 Full Forecasting Results

Lookback window $T = 96$. Tables 7 and 8 present the detailed results under the lookback window $T = 96$ setting. The former expands the results summarized in Table 1 of the main text, while the latter further includes additional representative baselines for a broader comparison with Navigator. Table 8 also reports the standard deviations of Navigator over five runs with different random seeds. Taken together, these results further confirm the strong forecasting performance and robustness of Navigator.

Increasing Lookback Length. We further examine how model performance evolves as the lookback length increases, as shown in Table 9. With the lookback window extended to $T = 720$, Navigator remains highly competitive across all five datasets and achieves the best average performance on ETTm1, ETTm2, Weather, and Electricity datasets. These results suggest that Navigator maintains strong forecasting performance under longer lookback windows, as Foreseer emphasizes a compact set of more predictive historical information while suppressing a large amount of redundant content.

C.2 Full Ablation Results

Full Ablation Results of Foreseer. Table 10 provides the complete results corresponding to Table 2 in the main text. Across individual prediction horizons, the results exhibit trends consistent with the averaged results reported in the main text. The full results further show that the advantages

Table 7: Performance comparison with a fixed lookback window $T = 96$. The prediction horizons are set as $\tau \in \{12, 24, 48, 96\}$ for PEMS datasets for short-term forecasting, and $\tau \in \{96, 192, 336, 720\}$ for the others for long-term forecasting. The best and second-best results are highlighted in **red** and **blue**, respectively.

Models		**Navigator (ours)**		PatchLGA (2026)		DistDF (2026a)		VPNet (2026)		RAFT (2025)		FreDF (2025)		TimeXer (2024b)		FilterNet (2024a)		iTransformer (2024)		PatchTST (2023)		FreTS (2023b)	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.361**	**0.384**	0.374	0.395	**0.373**	0.393	0.374	**0.386**	0.376	0.401	0.382	0.400	0.382	0.403	0.382	0.402	0.386	0.405	0.414	0.419	0.399	0.412
	192	**0.414**	**0.413**	0.431	0.424	0.428	0.425	0.428	**0.418**	**0.417**	0.424	0.430	0.427	0.429	0.435	0.430	0.429	0.441	0.436	0.460	0.445	0.453	0.443
	336	**0.454**	**0.437**	0.466	0.442	0.466	0.445	**0.464**	**0.439**	**0.454**	0.441	0.474	0.451	0.468	0.448	0.472	0.451	0.487	0.458	0.501	0.466	0.503	0.475
	720	**0.460**	**0.460**	0.476	0.466	**0.453**	**0.453**	0.471	0.465	0.465	0.465	0.463	0.462	0.469	0.461	0.481	0.473	0.503	0.491	0.500	0.488	0.596	0.565
	Avg	**0.422**	**0.424**	0.437	0.432	0.430	0.429	0.434	**0.427**	**0.428**	0.433	0.437	0.435	0.437	0.437	0.441	0.439	0.454	0.447	0.469	0.455	0.488	0.474
ETTh2	96	**0.284**	**0.329**	0.288	0.336	0.287	0.336	**0.284**	**0.330**	0.289	0.342	0.289	0.337	**0.286**	0.338	0.293	0.343	0.297	0.349	0.302	0.348	0.350	0.403
	192	0.359	**0.378**	0.366	0.384	**0.358**	0.381	**0.357**	**0.376**	0.382	0.400	0.363	0.385	0.363	0.389	0.374	0.396	0.380	0.400	0.388	0.400	0.472	0.475
	336	0.401	0.411	**0.380**	**0.408**	0.408	0.421	**0.392**	**0.406**	0.424	0.439	0.419	0.426	0.414	0.423	0.417	0.430	0.428	0.432	0.426	0.433	0.564	0.528
	720	0.410	**0.430**	0.409	0.432	0.416	0.435	**0.391**	**0.418**	0.433	0.457	0.415	0.437	**0.408**	0.432	0.449	0.460	0.427	0.445	0.431	0.446	0.815	0.654
	Avg	0.364	**0.387**	**0.361**	0.390	0.367	0.393	**0.356**	**0.383**	0.382	0.410	0.371	0.396	0.368	0.396	0.383	0.407	0.383	0.407	0.387	0.407	0.550	0.515
ETTm1	96	**0.303**	**0.336**	0.327	0.357	0.316	0.357	**0.313**	**0.342**	0.325	0.369	0.324	0.362	0.318	0.356	0.321	0.361	0.334	0.368	0.329	0.367	0.339	0.374
	192	**0.354**	**0.362**	0.371	0.376	**0.358**	0.380	0.362	**0.371**	0.364	0.388	0.373	0.385	0.362	0.383	0.367	0.387	0.377	0.391	0.367	0.385	0.382	0.397
	336	**0.385**	**0.384**	0.415	0.399	**0.390**	0.404	**0.385**	**0.389**	0.391	0.407	0.402	0.404	0.395	0.407	0.401	0.409	0.426	0.420	0.399	0.410	0.421	0.426
	720	0.450	**0.423**	0.469	0.439	**0.448**	0.437	**0.444**	**0.424**	**0.444**	0.438	0.469	0.444	0.452	0.441	0.477	0.448	0.491	0.459	0.454	0.439	0.485	0.462
	Avg	**0.373**	**0.376**	0.396	0.393	0.378	0.394	**0.376**	**0.382**	0.381	0.400	0.392	0.399	0.382	0.397	0.392	0.401	0.407	0.410	0.387	0.400	0.407	0.415
ETTm2	96	**0.168**	**0.246**	0.175	0.258	0.174	0.256	**0.169**	**0.244**	0.176	0.264	0.173	0.252	0.171	0.256	0.175	0.258	0.180	0.264	0.175	0.259	0.190	0.282
	192	**0.233**	**0.289**	0.242	0.303	0.239	0.298	**0.233**	**0.288**	0.241	0.306	0.241	0.298	**0.237**	0.299	0.240	0.301	0.250	0.309	0.241	0.302	0.260	0.329
	336	**0.293**	**0.328**	0.302	0.343	0.300	0.338	**0.291**	**0.327**	0.302	0.345	0.298	0.334	0.296	0.338	0.311	0.347	0.311	0.348	0.305	0.343	0.373	0.405
	720	**0.389**	**0.386**	0.396	0.397	0.397	0.394	**0.387**	**0.387**	0.404	0.405	0.398	0.393	0.392	0.394	0.414	0.405	0.412	0.407	0.402	0.400	0.517	0.499
	Avg	**0.271**	**0.312**	0.279	0.325	0.277	0.321	**0.270**	**0.312**	0.281	0.330	0.278	**0.319**	0.274	0.322	0.285	0.328	0.288	0.332	0.281	0.326	0.335	0.379
Electricity	96	**0.133**	**0.223**	0.149	0.241	0.137	0.235	**0.135**	**0.224**	0.152	0.257	0.144	0.233	0.140	0.242	0.147	0.245	0.148	0.240	0.181	0.270	0.189	0.277
	192	**0.152**	**0.240**	0.166	0.264	0.159	0.257	**0.151**	**0.239**	0.156	0.257	0.159	0.247	0.157	0.256	0.160	0.250	0.162	0.253	0.188	0.274	0.193	0.282
	336	**0.165**	**0.254**	0.175	0.264	0.178	0.272	**0.166**	**0.257**	0.171	0.272	0.172	0.263	0.176	0.275	0.173	0.267	0.178	0.269	0.204	0.293	0.207	0.296
	720	**0.192**	**0.281**	0.212	0.292	0.212	0.302	**0.196**	**0.284**	0.221	0.300	0.204	0.294	0.211	0.306	0.210	0.309	0.225	0.317	0.246	0.324	0.245	0.332
	Avg	**0.161**	**0.250**	0.176	0.265	0.172	0.267	**0.162**	**0.251**	0.175	0.272	0.170	0.259	0.171	0.270	0.173	0.268	0.178	0.270	0.205	0.290	0.209	0.297
Traffic	96	0.396	**0.230**	0.429	0.295	**0.380**	0.262	**0.384**	**0.258**	0.388	0.265	0.391	0.265	0.428	0.271	0.430	0.294	0.395	0.268	0.462	0.295	0.528	0.341
	192	0.426	**0.243**	0.440	0.298	0.407	0.275	**0.406**	**0.266**	**0.400**	0.278	0.410	0.273	0.448	0.282	0.452	0.307	0.417	0.276	0.466	0.296	0.531	0.338
	336	0.448	**0.252**	0.459	0.314	0.429	0.284	0.429	**0.275**	**0.411**	0.287	**0.424**	0.280	0.473	0.289	0.470	0.316	0.433	0.283	0.482	0.304	0.551	0.345
	720	0.490	**0.271**	0.524	0.330	**0.452**	0.297	0.466	**0.294**	**0.456**	0.305	0.460	0.298	0.516	0.307	0.498	0.323	0.467	0.302	0.514	0.322	0.598	0.367
	Avg	0.440	**0.249**	0.463	0.309	**0.417**	0.279	0.421	**0.273**	**0.414**	0.284	0.421	0.279	0.466	0.287	0.463	0.310	0.428	0.282	0.481	0.304	0.552	0.348
Weather	96	**0.153**	**0.192**	0.174	0.213	0.164	0.209	**0.157**	**0.192**	0.185	0.241	0.164	**0.202**	**0.157**	0.205	0.162	0.207	0.174	0.214	0.177	0.218	0.184	0.239
	192	**0.200**	**0.235**	0.211	0.244	0.212	0.252	0.207	**0.239**	0.237	0.287	0.220	0.253	**0.204**	0.247	0.210	0.250	0.221	0.254	0.225	0.259	0.223	0.275
	336	**0.258**	**0.279**	0.273	0.288	0.270	0.295	**0.258**	**0.281**	0.290	0.327	0.275	0.294	**0.261**	0.290	0.265	0.290	0.278	0.296	0.278	0.297	0.272	0.316
	720	**0.337**	**0.331**	0.352	0.346	0.348	0.345	**0.330**	**0.333**	0.367	0.380	0.356	0.347	0.340	0.341	0.342	0.340	0.358	0.349	0.354	0.348	0.340	0.363
	Avg	**0.237**	**0.259**	0.253	0.273	0.248	0.275	**0.238**	**0.261**	0.270	0.309	0.254	0.274	0.241	0.271	0.245	0.272	0.258	0.278	0.259	0.281	0.340	0.363
Solar-Energy	96	**0.158**	**0.193**	0.189	0.251	0.185	0.237	**0.177**	**0.178**	0.278	0.278	0.188	0.240	0.215	0.295	0.205	0.242	0.203	0.237	0.234	0.286	0.237	0.300
	192	**0.187**	**0.210**	0.213	0.265	0.202	0.250	**0.196**	**0.206**	0.297	0.297	0.206	0.255	0.236	0.301	0.233	0.265	0.233	0.261	0.267	0.310	0.265	0.321
	336	**0.205**	**0.220**	0.226	0.273	**0.212**	0.265	0.216	**0.219**	0.311	0.311	0.217	0.272	0.252	0.307	0.249	0.278	0.248	0.273	0.290	0.315	0.283	0.330
	720	**0.218**	**0.226**	0.243	0.280	0.230	0.268	**0.228**	**0.225**	0.318	0.325	0.231	0.282	0.244	0.305	0.253	0.281	0.249	0.275	0.289	0.317	0.286	0.326
	Avg	**0.192**	**0.212**	0.218	0.267	0.207	0.255	**0.204**	**0.207**	0.301	0.303	0.211	0.262	0.237	0.302	0.235	0.266	0.233	0.262	0.270	0.307	0.268	0.322
PEMS03	12	**0.065**	**0.162**	0.078	0.194	**0.068**	0.174	**0.068**	0.174	0.071	0.175	**0.068**	**0.172**	0.070	0.173	0.071	0.177	0.071	0.174	0.099	0.216	0.083	0.194
	24	**0.078**	**0.179**	0.101	0.219	0.094	0.205	**0.091**	0.201	0.099	0.201	0.096	0.205	0.092	**0.194**	0.102	0.213	0.093	0.201	0.142	0.259	0.127	0.241
	48	**0.096**	**0.201**	0.141	0.261	0.138	0.252	0.135	0.247	0.153	0.243	0.128	0.240	0.129	**0.229**	0.162	0.272	**0.125**	0.236	0.211	0.319	0.204	0.311
	96	**0.123**	**0.228**	0.198	0.306	0.191	0.301	0.195	0.304	0.253	0.301	0.161	0.269	**0.157**	**0.261**	0.244	0.340	0.164	0.275	0.269	0.370	0.255	0.345
	Avg	**0.091**	**0.193**	0.130	0.245	0.123	0.233	0.122	0.232	0.144	0.230	0.113	0.219	**0.112**	**0.214**	0.145	0.251	0.113	0.221	0.180	0.291	0.167	0.273
PEMS04	12	**0.068**	**0.164**	0.079	0.181	**0.071**	**0.167**	0.084	0.182	0.077	0.179	0.073	0.174	0.074	0.178	0.082	0.190	0.078	0.183	0.105	0.224	0.098	0.210
	24	**0.076**	**0.173**	0.090	0.186	**0.080**	**0.181**	0.099	0.206	0.091	0.195	0.085	0.186	0.087	0.195	0.110	0.224	0.095	0.205	0.153	0.275	0.148	0.263
	48	**0.085**	**0.184**	0.103	0.213	**0.097**	**0.205**	0.117	0.225	0.111	0.218	0.101	0.212	0.110	0.214	0.160	0.276	0.120	0.233	0.229	0.339	0.235	0.341
	96	**0.096**	**0.197**	0.129	0.241	**0.120**	**0.228**	0.145	0.261	0.137	0.247	0.125	0.235	0.148	0.251	0.234	0.343	0.150	0.262	0.291	0.389	0.304	0.393
	Avg	**0.081**	**0.180**	0.100	0.205	**0.092**	**0.195**	0.111	0.219	0.104	0.210	0.096	0.202	0.105	0.209	0.146	0.258	0.111	0.221	0.195	0.307	0.196	0.302
PEMS07	12	**0.053**	**0.139**	0.060	0.161	**0.056**	**0.149**	0.060	0.156	0.062	0.159	0.060	0.156	0.057	0.152	0.064	0.163	0.067	0.165	0.095	0.207	0.080	0.190
	24	**0.063**	**0.149**	0.079	0.184	**0.073**	**0.168**	0.075	0.176	0.078	0.178	0.077	0.170	0.079	0.179	0.093	0.200	0.088	0.190	0.150	0.262	0.136	0.250
	48	**0.074**	**0.161**	0.096	0.199	**0.089**	**0.183**	0.103	0.211	0.102	0.203	0.093	0.188	0.099	0.191	0.137	0.248	0.110	0.215	0.253	0.340	0.248	0.340
	96	**0.086**	**0.173**	0.131	0.229	0.119	0.209	0.138	0.247	0.135	0.231	0.124	0.215	**0.107**	**0.205**	0.198	0.306	0.139	0.245	0.346	0.404	0.352	0.413
	Avg	**0.069**	**0.156**	0.092	0.193	**0.084**	**0.177**	0.094	0.198	0.094	0.193	0.089	0.182	0.085	0.182	0.123	0.229	0.101	0.204	0.211	0.303	0.204	0.298
PEMS08	12	**0.074**	**0.170**	0.084	0.189	0.076	0.177	0.079	0.185	0.078	0.181	0.080	0.182	**0.075**	**0.176**	0.080	0.182	0.079	0.182	0.168	0.232	0.095	0.204
	24	**0.095**	**0.197**	0.116	0.223	0.107	0.210	0.106	0.210	**0.102**	0.205	0.118	0.220	**0.102**	**0.201**	0.114	0.219	0.115	0.219	0.224	0.281	0.150	0.259
	48	**0.135**	**0.231**	0.187	0.277	0.171	0.265	0.182	0.272	**0.148**	0.244	0.161	0.258	0.158	0.248	0.184	0.284	0.186	**0.235**	0.321	0.354	0.250	0.341
	96	**0.189**	**0.280**	0.273	0.316	0.250	0.325	0.269	0.311	0.277	0.307	**0.206**	**0.293**	0.366	0.377	0.309	0.356	0.221	0.267	0.408	0.417	0.323	0.376
	Avg	**0.123**	**0.220**	0.165	0.251	0.151	0.244	0.159	0.245	0.151	0.234	**0.141**	0.238	0.175	0.250	0.172	0.260	0.150	**0.226**	0.280	0.321	0.205	0.295
1^{st} Count		44	46	1	0	3	1	13	13	5	0	0	0	0	0	0	0	0	0	0	0	0	0

of Foreseer and its key design choices are generally maintained across different forecasting horizons and datasets, supporting the robustness of the conclusions drawn from Table 2.

Full Ablation Results of Geometric Bending Attention. Table 11 provides the complete results corresponding to Table 4 in the main text. The detailed results show that replacing GBA with alternative attention mechanisms rarely yields sustained improvements across forecasting horizons. Instead, GBA attains the best or near-best performance in most settings, indicating that its query–key–value geometric scoring criterion provides a consistently effective mechanism for cross-channel interaction.

C.3 Comparison of Cross-Channel Modeling Methods

Table 12 presents a comparison between Navigator and representative forecasting methods with explicit cross-channel modeling capabilities, including TimeBridge (Liu et al., 2025), TQNet (Lin

Table 8: Additional performance comparison with a fixed lookback window $T = 96$. The prediction horizons are set as $\tau \in \{12, 24, 48, 96\}$ for PEMS datasets for short-term forecasting, and $\tau \in \{96, 192, 336, 720\}$ for the others for long-term forecasting. The best and second-best results are highlighted in **red** and **blue**, respectively. The reported results with standard deviation of Navigator are averaged from 5 runs (with different random seeds of 2024, 2025, 2026, 2027, 2028).

Models		**Navigator (ours)**		TimeBridge (2025)		TimeMixer (2024a)		FITS (2024)		Fredformer (2024)		PAttn (2024)		Crossformer (2023)		TimesNet (2023)		DLinear (2023)	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.361**±0.000	**0.384**±0.000	**0.373**	0.395	0.375	0.400	0.385	0.394	**0.373**	**0.392**	0.390	0.404	0.423	0.448	0.422	0.433	0.386	0.400
	192	**0.414**±0.001	**0.413**±0.000	**0.428**	0.426	0.429	0.421	0.434	0.422	0.433	**0.420**	0.443	0.435	0.471	0.474	0.465	0.457	0.437	0.432
	336	**0.454**±0.001	**0.437**±0.001	0.471	0.451	0.484	0.458	0.476	**0.444**	**0.470**	**0.437**	0.487	0.462	0.570	0.546	0.492	0.470	0.481	0.459
	720	**0.460**±0.000	**0.460**±0.000	0.495	0.487	0.498	0.482	**0.465**	0.462	0.467	**0.456**	0.507	0.488	0.653	0.621	0.532	0.502	0.519	0.516
	Avg	**0.422**±0.001	**0.424**±0.000	0.442	0.440	0.447	0.440	0.440	0.431	**0.435**	**0.426**	0.457	0.447	0.529	0.522	0.478	0.466	0.456	0.452
ETTh2	96	**0.284**±0.001	**0.329**±0.000	0.294	0.344	**0.289**	0.341	0.292	**0.340**	0.293	0.342	0.309	0.358	0.745	0.584	0.320	0.364	0.333	0.387
	192	**0.359**±0.001	**0.378**±0.001	**0.371**	0.394	0.372	0.392	0.377	0.391	**0.371**	**0.389**	0.383	0.407	0.877	0.656	0.409	0.417	0.477	0.476
	336	0.401±0.001	**0.411**±0.000	0.421	0.429	**0.386**	0.414	0.416	0.425	**0.382**	**0.409**	0.429	0.436	1.043	0.731	0.449	0.451	0.594	0.541
	720	**0.410**±0.001	**0.430**±0.001	0.423	0.443	**0.412**	**0.434**	0.418	0.437	0.415	**0.434**	0.442	0.450	1.104	0.763	0.473	0.474	0.831	0.657
	Avg	**0.364**±0.001	**0.387**±0.001	0.377	0.403	**0.364**	0.395	0.376	0.398	**0.365**	**0.393**	0.391	0.413	0.942	0.684	0.413	0.426	0.559	0.515
ETTm1	96	**0.303**±0.000	**0.336**±0.000	0.323	0.361	**0.320**	**0.357**	0.353	0.375	0.326	0.361	0.331	0.369	0.404	0.426	0.338	0.379	0.345	0.372
	192	**0.354**±0.000	**0.362**±0.000	0.366	0.385	**0.361**	0.381	0.486	0.445	0.363	**0.380**	0.365	0.385	0.450	0.451	0.389	0.400	0.380	0.389
	336	**0.385**±0.000	**0.384**±0.000	0.398	0.408	**0.390**	0.404	0.531	0.475	0.395	**0.403**	0.392	0.404	0.532	0.515	0.429	0.428	0.413	0.413
	720	**0.450**±0.000	**0.423**±0.001	0.461	0.445	0.454	0.441	0.600	0.513	**0.453**	**0.438**	0.456	0.439	0.666	0.589	0.495	0.464	0.474	0.453
	Avg	**0.373**±0.000	**0.376**±0.000	0.387	0.400	**0.381**	**0.395**	0.493	0.452	0.384	**0.395**	0.386	0.399	0.513	0.496	0.413	0.418	0.403	0.407
ETTm2	96	**0.168**±0.000	**0.246**±0.000	0.177	0.259	**0.175**	**0.258**	0.182	0.266	0.177	0.259	0.181	0.263	0.287	0.366	0.185	0.264	0.193	0.292
	192	**0.233**±0.000	**0.289**±0.000	0.243	0.303	**0.237**	**0.299**	0.253	0.312	0.243	0.301	0.245	0.307	0.414	0.492	0.254	0.307	0.284	0.362
	336	**0.293**±0.000	**0.328**±0.000	0.303	0.343	**0.298**	**0.340**	0.313	0.349	0.302	**0.340**	0.315	0.353	0.597	0.542	0.314	0.345	0.369	0.427
	720	**0.389**±0.001	**0.386**±0.000	0.401	0.399	**0.391**	**0.396**	0.416	0.406	0.397	**0.396**	0.415	0.408	1.730	1.042	0.434	0.413	0.554	0.522
	Avg	**0.271**±0.000	**0.312**±0.000	0.281	0.326	**0.275**	**0.323**	0.291	0.333	0.279	0.324	0.289	0.333	0.757	0.610	0.297	0.332	0.350	0.401
Electricity	96	**0.133**±0.001	**0.223**±0.001	**0.142**	**0.239**	0.153	0.247	0.198	0.274	0.147	0.241	0.182	0.269	0.219	0.314	0.171	0.273	0.197	0.282
	192	**0.152**±0.001	**0.240**±0.001	**0.161**	0.257	0.166	**0.256**	0.363	0.422	0.165	0.258	0.179	0.264	0.231	0.322	0.188	0.289	0.196	0.285
	336	**0.165**±0.001	**0.254**±0.001	0.182	0.278	0.185	0.277	0.444	0.490	**0.177**	**0.273**	0.195	0.282	0.246	0.337	0.208	0.304	0.209	0.301
	720	**0.192**±0.001	**0.281**±0.001	0.217	0.309	0.225	0.310	0.532	0.551	**0.213**	**0.304**	0.238	0.317	0.280	0.363	0.289	0.363	0.245	0.333
	Avg	**0.161**±0.001	**0.250**±0.001	**0.176**	0.271	0.182	0.272	0.384	0.434	**0.176**	**0.269**	0.199	0.283	0.244	0.334	0.214	0.307	0.212	0.300
Traffic	96	**0.396**±0.001	**0.230**±0.001	**0.391**	**0.268**	0.462	0.285	0.601	0.361	0.406	0.277	0.470	0.304	0.522	0.290	0.504	0.298	0.650	0.396
	192	**0.426**±0.001	**0.243**±0.001	**0.418**	**0.276**	0.473	0.296	0.603	0.365	**0.426**	0.290	0.494	0.322	0.530	0.293	0.526	0.305	0.598	0.370
	336	**0.448**±0.001	**0.252**±0.002	**0.432**	0.284	0.498	0.296	0.609	0.366	**0.432**	**0.281**	0.507	0.325	0.558	0.305	0.540	0.310	0.605	0.373
	720	0.490±0.001	**0.271**±0.001	**0.464**	0.301	0.506	0.313	0.648	0.387	**0.463**	**0.300**	0.538	0.344	0.589	0.328	0.570	0.324	0.645	0.394
	Avg	0.440±0.001	**0.249**±0.001	**0.426**	**0.282**	0.484	0.297	0.615	0.370	**0.431**	0.287	0.502	0.324	0.550	0.304	0.535	0.309	0.625	0.383
Weather	96	**0.153**±0.000	**0.192**±0.000	0.168	0.211	0.163	0.209	0.196	0.236	0.163	**0.207**	0.172	0.214	**0.158**	0.230	0.178	0.226	0.196	0.255
	192	**0.200**±0.001	**0.235**±0.001	0.214	0.254	0.208	**0.250**	0.240	0.271	0.211	0.251	0.222	0.257	**0.206**	0.277	0.227	0.266	0.237	0.296
	336	**0.258**±0.001	**0.279**±0.001	0.273	0.297	**0.251**	**0.287**	0.292	0.307	0.267	0.292	0.275	0.297	0.272	0.335	0.283	0.305	0.283	0.335
	720	**0.337**±0.001	**0.331**±0.001	0.353	0.347	**0.339**	**0.341**	0.365	0.354	0.343	**0.341**	0.358	0.348	0.398	0.418	0.359	0.355	0.345	0.381
	Avg	**0.237**±0.001	**0.259**±0.001	0.252	0.277	**0.240**	**0.271**	0.273	0.292	0.246	0.272	0.257	0.279	0.259	0.315	0.262	0.288	0.345	0.381
Solar-Energy	96	**0.158**±0.001	**0.193**±0.002	0.192	**0.207**	0.189	0.259	0.319	0.353	**0.185**	0.233	0.222	0.262	0.310	0.331	0.250	0.292	0.290	0.378
	192	**0.187**±0.002	**0.210**±0.002	0.223	**0.231**	**0.222**	0.283	0.367	0.387	0.227	0.253	0.242	0.276	0.734	0.725	0.296	0.318	0.320	0.398
	336	**0.205**±0.002	**0.220**±0.003	0.243	**0.250**	**0.231**	0.292	0.408	0.403	0.246	0.284	0.271	0.296	0.750	0.735	0.319	0.330	0.353	0.415
	720	**0.218**±0.001	**0.226**±0.002	0.250	**0.257**	**0.223**	0.285	0.411	0.395	0.247	0.276	0.259	0.286	0.769	0.765	0.338	0.337	0.356	0.413
	Avg	**0.192**±0.002	**0.212**±0.002	0.227	**0.236**	**0.216**	0.280	0.376	0.384	0.226	0.262	0.249	0.280	0.641	0.639	0.301	0.319	0.330	0.401
PEMS03	12	**0.065**±0.000	**0.162**±0.001	0.070	0.176	0.076	0.188	0.117	0.226	**0.068**	**0.174**	0.084	0.196	0.090	0.203	0.085	0.192	0.122	0.243
	24	**0.078**±0.001	**0.179**±0.000	0.099	0.211	0.113	0.226	0.235	0.324	**0.093**	**0.202**	0.143	0.255	0.121	0.240	0.118	0.223	0.201	0.317
	48	**0.096**±0.002	**0.201**±0.001	0.153	0.267	0.191	0.292	0.541	0.521	**0.146**	**0.258**	0.256	0.345	0.202	0.317	0.155	0.260	0.333	0.425
	96	**0.123**±0.001	**0.228**±0.002	**0.196**	**0.305**	0.288	0.363	1.062	0.790	0.228	0.330	0.426	0.460	0.262	0.367	0.228	0.317	0.457	0.515
	Avg	**0.091**±0.001	**0.193**±0.001	**0.130**	**0.240**	0.167	0.267	0.489	0.465	0.135	0.243	0.227	0.314	0.169	0.281	0.147	0.248	0.278	0.375
PEMS04	12	**0.068**±0.002	**0.164**±0.001	**0.082**	**0.183**	0.092	0.204	0.129	0.239	0.085	0.189	0.106	0.218	0.098	0.218	0.087	0.195	0.148	0.272
	24	**0.076**±0.001	**0.173**±0.001	0.108	**0.213**	0.128	0.243	0.246	0.337	0.117	0.224	0.173	0.283	0.131	0.256	**0.103**	0.215	0.224	0.340
	48	**0.085**±0.003	**0.184**±0.002	0.158	0.263	0.213	0.315	0.568	0.539	0.174	0.276	0.326	0.396	0.205	0.326	**0.136**	**0.250**	0.355	0.437
	96	**0.096**±0.002	**0.197**±0.002	0.232	0.327	0.307	0.384	1.181	0.843	0.273	0.354	0.555	0.528	0.402	0.457	**0.190**	**0.303**	0.452	0.504
	Avg	**0.081**±0.002	**0.180**±0.002	0.145	0.247	0.185	0.287	0.531	0.489	0.162	0.261	0.290	0.356	0.209	0.314	**0.129**	**0.241**	0.295	0.388
PEMS07	12	**0.053**±0.001	**0.139**±0.001	**0.061**	**0.154**	0.073	0.184	0.109	0.222	0.063	0.158	0.083	0.193	0.094	0.200	0.082	0.181	0.115	0.242
	24	**0.063**±0.001	**0.149**±0.003	**0.085**	**0.182**	0.111	0.219	0.230	0.327	0.089	0.192	0.139	0.250	0.139	0.247	0.101	0.204	0.210	0.329
	48	**0.074**±0.002	**0.161**±0.002	**0.128**	**0.226**	0.237	0.328	0.551	0.531	0.136	0.241	0.263	0.347	0.311	0.369	0.134	0.238	0.398	0.458
	96	**0.086**±0.002	**0.173**±0.002	0.186	0.281	0.303	0.354	1.112	0.809	0.197	0.298	0.449	0.467	0.396	0.442	**0.181**	**0.279**	0.594	0.553
	Avg	**0.069**±0.002	**0.156**±0.002	**0.115**	**0.211**	0.181	0.271	0.500	0.472	0.121	0.222	0.234	0.314	0.235	0.315	0.124	0.225	0.329	0.395
PEMS08	12	**0.074**±0.002	**0.170**±0.002	**0.080**	**0.184**	0.091	0.201	0.122	0.233	0.081	0.185	0.096	0.204	0.165	0.214	0.112	0.212	0.154	0.276
	24	**0.095**±0.002	**0.197**±0.001	0.119	0.224	0.137	0.246	0.236	0.330	**0.112**	**0.214**	0.167	0.275	0.215	0.260	0.141	0.238	0.248	0.353
	48	**0.135**±0.001	**0.231**±0.003	0.198	0.289	0.265	0.343	0.562	0.540	**0.174**	**0.267**	0.290	0.362	0.315	0.355	0.198	0.283	0.440	0.470
	96	**0.189**±0.002	**0.280**±0.001	0.303	**0.322**	0.410	0.407	1.216	0.846	**0.277**	0.335	0.519	0.490	0.377	0.397	0.320	0.351	0.674	0.565
	Avg	**0.123**±0.002	**0.220**±0.002	0.175	0.255	0.226	0.299	0.534	0.487	**0.161**	**0.250**	0.268	0.333	0.268	0.307	0.193	0.271	0.379	0.416
1^{st} Count		53	58	4	0	2	0	0	0	3	3	0	0	0	0	1	0	0	0

et al., 2025a), Leddam (Yu et al., 2024), CARD (Zhou et al., 2024), iTransformer (Liu et al., 2024), and Crossformer (Zhang & Yan, 2023). Navigator achieves the best average performance on most datasets and consistently delivers competitive results across different prediction horizons. The improvement is observed against a diverse set of models that employ different strategies for capturing cross-channel dependencies, indicating that the proposed approach is not limited to outperforming a particular type of interaction mechanism. These results further support the effectiveness of GBA in modeling cross-channel interactions through query–key–value geometric compatibility. Combined with Foreseer, which emphasizes more predictive historical information before cross-channel inter-

Table 9: Forecasting results with a long lookback window of $T = 720$ and prediction horizons $\tau \in \{96, 192, 336, 720\}$. The best and second-best results are highlighted in **red** and **blue**, respectively.

Models		**Navigator (ours)**		RAFT (2025)		SparseTSF (2024b)		CycleNet (2024a)		iTransformer (2024)		TimeMixer (2024a)		PatchTST (2023)		TimesNet (2023)	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	**0.290**	**0.340**	0.302	0.349	0.312	0.354	0.307	0.353	0.300	0.353	0.293	0.345	**0.289**	**0.343**	0.340	0.378
	192	**0.321**	**0.359**	**0.329**	**0.367**	0.347	0.376	0.337	0.371	0.341	0.380	0.335	0.372	**0.329**	0.368	0.392	0.404
	336	**0.351**	**0.378**	**0.355**	**0.383**	0.367	0.386	0.364	0.387	0.374	0.396	0.368	0.386	0.362	0.390	0.423	0.426
	720	**0.410**	**0.412**	**0.406**	0.413	0.419	0.413	**0.410**	**0.411**	0.429	0.430	0.426	0.417	0.416	0.423	0.475	0.453
	Avg	**0.343**	**0.372**	**0.348**	**0.378**	0.361	0.382	0.355	0.381	0.361	0.390	0.356	0.380	0.349	0.381	0.408	0.415
ETTm2	96	**0.159**	**0.248**	0.164	0.256	**0.163**	0.252	**0.159**	**0.249**	0.175	0.266	0.165	0.256	0.165	0.255	0.189	0.265
	192	**0.212**	**0.284**	0.219	0.296	0.217	0.290	**0.214**	**0.289**	0.242	0.312	0.225	0.298	0.221	0.293	0.254	0.310
	336	**0.268**	**0.322**	0.275	0.336	**0.270**	0.327	**0.268**	**0.326**	0.282	0.337	0.277	0.332	0.276	0.327	0.313	0.345
	720	**0.349**	**0.374**	0.359	0.392	**0.352**	**0.379**	0.353	0.384	0.375	0.394	0.360	0.387	0.362	0.381	0.413	0.402
	Avg	**0.247**	**0.307**	0.254	0.320	0.251	**0.312**	**0.249**	**0.312**	0.269	0.327	0.257	0.318	0.256	0.314	0.292	0.331
Weather	96	**0.145**	**0.191**	0.165	0.222	0.169	0.223	0.164	0.220	0.157	0.207	**0.147**	0.198	0.149	**0.196**	0.168	0.214
	192	**0.187**	**0.231**	0.211	0.264	0.214	0.262	0.209	0.258	0.200	0.248	0.192	0.243	**0.191**	**0.239**	0.219	0.262
	336	**0.235**	**0.270**	0.260	0.302	0.257	0.293	0.255	0.292	0.252	0.287	0.247	0.284	**0.242**	**0.279**	0.278	0.302
	720	**0.303**	**0.318**	0.327	0.355	0.321	0.340	0.320	0.338	0.320	0.336	0.318	**0.330**	**0.312**	**0.330**	0.353	0.351
	Avg	**0.218**	**0.253**	0.241	0.286	0.240	0.280	0.237	0.277	0.232	0.270	0.226	0.264	**0.224**	**0.261**	0.255	0.282
Electricity	96	**0.130**	**0.223**	0.133	0.232	0.138	0.233	**0.128**	**0.223**	0.134	**0.230**	0.153	0.256	0.143	0.247	0.169	0.271
	192	**0.146**	**0.238**	0.149	0.247	0.151	0.244	**0.143**	**0.237**	0.154	0.250	0.168	0.269	0.158	0.260	0.180	0.280
	336	**0.160**	**0.253**	0.161	0.259	0.166	0.260	**0.159**	**0.254**	0.169	0.265	0.189	0.291	0.168	0.267	0.204	0.304
	720	**0.183**	**0.272**	0.197	0.297	0.205	0.293	0.197	**0.287**	**0.194**	0.288	0.228	0.320	0.214	0.307	0.205	0.304
	Avg	**0.155**	**0.247**	0.160	0.259	0.165	0.258	**0.157**	**0.250**	0.163	0.258	0.184	0.284	0.171	0.270	0.190	0.290
Solar-Energy	96	**0.175**	**0.200**	0.192	0.251	0.195	0.243	0.194	0.255	0.190	0.244	0.179	**0.232**	**0.170**	0.234	0.198	0.270
	192	**0.188**	**0.215**	0.247	0.323	0.215	0.254	0.205	**0.251**	**0.193**	0.257	0.201	0.259	0.204	0.302	0.206	0.276
	336	0.204	**0.223**	0.240	0.300	0.232	0.262	0.218	0.257	**0.203**	0.266	**0.190**	**0.256**	0.212	0.293	0.208	0.284
	720	**0.210**	**0.226**	0.246	0.311	0.237	0.263	0.239	0.278	0.223	0.281	**0.203**	**0.261**	0.215	0.307	0.232	0.294
	Avg	**0.194**	**0.216**	0.231	0.296	0.220	0.256	0.214	0.260	0.202	0.262	**0.193**	**0.252**	0.200	0.284	0.211	0.281

action, GBA enables Navigator to achieve strong and stable forecasting performance across datasets with different channel characteristics.

C.4 Extending Geometric Bending Attention to Temporal Modeling

Although GBA is instantiated for cross-channel interaction in Navigator, its underlying geometric scoring mechanism is not inherently restricted to the channel dimension. The key idea of GBA is to determine interaction strength from the geometric compatibility jointly formed by query, key, and value representations. Such a formulation is applicable whenever interactions among a set of tokens need to be modeled. We therefore further investigate whether this geometric attention principle can also be extended to temporal modeling, providing additional evidence on the applicability of GBA beyond its original cross-channel setting. To this end, we construct a temporal variant, termed *GBA-Time*. Specifically, *GBA-Time* adopts the temporal patching and embedding scheme of PatchTST (Nie et al., 2023) to construct temporal tokens, and replaces the standard attention applied to these tokens with our proposed geometric attention mechanism. In contrast, *GBA-Channel*, the original GBA used in Navigator, operates on channel tokens. This design allows us to examine whether the proposed geometric attention mechanism can be effectively extended to temporal modeling while retaining a well-established temporal tokenization scheme.

As shown in Table 13, GBA-Time achieves performance comparable to the original GBA-Channel across the six datasets. The performance gap is generally small on ETTh1, ETTh2, ETTm1, and ETTm2, indicating that the proposed geometric scoring criterion remains effective when transferred from cross-channel interaction to temporal interaction. For example, on ETTh2 dataset, the average MSE changes only from 0.364 to 0.366, while the corresponding MAE changes from 0.387 to 0.388. Meanwhile, GBA-Channel generally retains better overall performance, particularly on Weather and Electricity. One possible explanation is that, compared with the four ETT datasets, Weather and Electricity contain substantially more channels, making the information conveyed by cross-channel interactions potentially more important for forecasting. Consequently, explicitly applying GBA across channel tokens may better exploit such dependencies on these datasets. In addition, GBA-Time introduces moderately more parameters due to the architectural modifications required for temporal tokenization and modeling. Overall, these results suggest that the effectiveness of GBA is not tied exclusively to cross-channel modeling. Its query–key–value geometric scoring principle

Table 10: Full ablation results of Foreseer on six datasets. The lookback window is fixed at $T = 96$. The prediction horizons are set to $\tau \in \{12, 24, 48, 96\}$ for PEMSD3 and PEMSD4, and $\tau \in \{96, 192, 336, 720\}$ for the remaining datasets.

Module variants		*w* Foreseer		*w/o* Foreseer		*w/o* Deltas		*r* Linear		*r* Channel-wise	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.153**	**0.192**	0.159	0.196	0.158	0.195	0.162	0.199	0.155	0.194
	192	**0.200**	**0.235**	0.202	0.238	0.201	0.237	0.203	0.237	0.205	0.241
	336	**0.258**	**0.279**	0.262	0.283	0.262	0.280	0.261	**0.279**	0.261	0.282
	720	**0.337**	**0.331**	0.345	0.336	0.342	0.335	0.345	0.336	0.340	0.333
	Avg	**0.237**	**0.259**	0.242	0.263	0.241	0.262	0.243	0.263	0.240	0.263
Electricity	96	**0.133**	**0.223**	0.137	0.227	0.135	0.225	0.136	0.226	0.147	0.238
	192	**0.152**	**0.240**	0.155	0.243	0.156	0.243	0.156	0.245	0.169	0.255
	336	**0.165**	**0.254**	0.168	0.257	0.170	0.259	0.170	0.258	0.185	0.272
	720	**0.192**	**0.281**	0.199	0.286	0.199	0.285	0.202	0.289	0.222	0.302
	Avg	**0.161**	**0.250**	0.165	0.253	0.165	0.253	0.166	0.255	0.181	0.267
Solar-Energy	96	**0.158**	**0.193**	0.176	0.202	0.185	0.206	0.178	0.200	0.160	0.195
	192	**0.187**	**0.210**	0.192	0.217	0.193	0.219	0.199	0.222	0.185	0.220
	336	**0.205**	**0.220**	0.207	0.229	0.212	0.232	0.212	0.233	0.209	0.229
	720	**0.218**	**0.226**	0.236	0.232	0.225	0.232	0.220	0.233	0.221	0.232
	Avg	**0.192**	**0.212**	0.203	0.220	0.204	0.222	0.202	0.222	0.194	0.219
Traffic	96	**0.396**	**0.230**	0.404	0.234	0.405	0.233	0.410	0.235	0.464	0.287
	192	**0.426**	**0.243**	0.428	0.246	0.433	0.247	0.436	0.249	0.484	0.295
	336	**0.448**	**0.252**	0.454	0.254	0.457	0.255	0.454	0.256	0.504	0.302
	720	**0.490**	**0.271**	0.498	0.272	0.497	0.273	0.500	0.275	0.549	0.326
	Avg	**0.440**	**0.249**	0.446	0.252	0.448	0.252	0.450	0.254	0.500	0.303
PEMSD3	12	0.065	**0.162**	**0.064**	0.164	0.071	0.173	0.065	0.165	0.079	0.183
	24	**0.078**	**0.179**	0.085	0.189	0.091	0.197	0.086	0.192	0.098	0.206
	48	**0.096**	**0.201**	0.120	0.228	0.123	0.232	0.120	0.229	0.131	0.238
	96	**0.123**	**0.228**	0.155	0.262	0.158	0.267	0.154	0.262	0.162	0.269
	Avg	**0.091**	**0.193**	0.106	0.211	0.111	0.217	0.106	0.212	0.118	0.224
PEMSD4	12	**0.068**	**0.164**	0.073	0.172	0.078	0.180	0.073	0.173	0.075	0.176
	24	**0.076**	**0.173**	0.088	0.191	0.095	0.200	0.089	0.192	0.085	0.189
	48	**0.085**	**0.184**	0.111	0.216	0.116	0.223	0.111	0.217	0.099	0.203
	96	**0.096**	**0.197**	0.137	0.241	0.142	0.247	0.137	0.242	0.128	0.231
	Avg	**0.081**	**0.180**	0.102	0.205	0.108	0.213	0.103	0.206	0.097	0.200

can be transferred to temporal interaction while preserving competitive forecasting performance, indicating that GBA may serve as a more general attention mechanism for modeling interactions along different dimensions.

C.5 Proportion of Training Data Used to Compute the Predictability Basis

Since the Predictability Basis is constructed directly from the training data, it is important to examine its sensitivity to the amount of data used for its computation. To this end, we vary the proportion of the training set used to construct the Predictability Basis, denoted as *M_ratio*, while training the forecasting model on the full training set in all experiments. As shown in Table 14, performance remains highly stable as *M_ratio* decreases from 100% to 10%, with only minor variations in MSE and MAE across all six datasets. Even a small fraction of the training data is sufficient to construct a Predictability Basis that achieves performance comparable to that obtained using the full training set. These results demonstrate the robustness of the Predictability Basis to the amount of data used for its construction, suggesting that stable historical–future relationships can be captured without requiring extensive training data.

C.6 More Efficiency Analysis

To further assess the computational efficiency of Navigator across different channel scales, we conduct experiments on three datasets with substantially different numbers of channels: ETTm1 (7 channels, Table 15), Weather (21 channels, Table 16), and Electricity (321 channels, Table 5). These

Table 11: Full ablation studies of Geometric Bending Attention (GBA), where we replace GBA with other attention mechanisms. The lookback window is fixed at $T = 96$, and the prediction horizons are set to $\tau \in \{96, 192, 336, 720\}$.

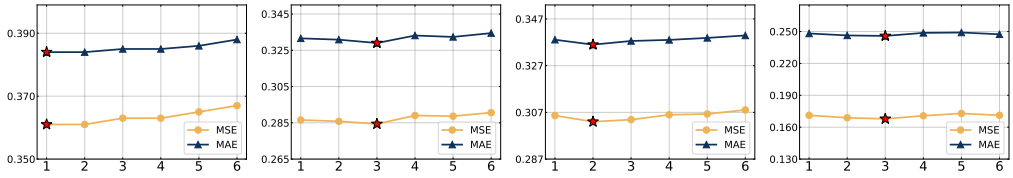
Datasets		ETTm1		Weather		Solar-Energy		Traffic	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
GBA (ours)	96	**0.303**	**0.336**	0.153	0.192	**0.158**	0.193	**0.396**	**0.230**
	192	0.354	**0.362**	**0.200**	**0.235**	0.187	0.210	**0.426**	**0.243**
	336	**0.385**	**0.384**	0.258	**0.279**	0.205	**0.220**	**0.448**	**0.252**
	720	**0.450**	**0.423**	**0.337**	**0.331**	**0.218**	0.226	**0.490**	**0.271**
	Avg	**0.373**	**0.376**	**0.237**	**0.259**	**0.192**	0.212	**0.440**	**0.249**
r Full Attention	96	0.306	0.339	0.160	0.198	0.186	0.198	0.405	0.231
	192	**0.353**	**0.362**	0.206	0.241	0.216	0.221	0.430	0.245
	336	0.386	0.385	0.261	0.282	0.219	0.230	0.457	0.253
	720	0.453	**0.423**	0.344	0.336	0.229	0.234	0.499	**0.271**
	Avg	0.375	0.377	0.243	0.264	0.213	0.221	0.448	0.250
r Prob Attention	96	0.310	0.341	0.155	0.194	**0.158**	**0.191**	0.397	**0.230**
	192	0.357	0.366	0.205	0.241	**0.179**	**0.206**	0.429	0.245
	336	0.389	0.388	0.266	0.283	0.206	0.222	0.455	0.254
	720	0.457	0.427	0.340	0.333	0.226	0.226	0.500	0.273
	Avg	0.378	0.381	0.242	0.263	**0.192**	**0.211**	0.445	0.251
r LSH Attention	96	0.306	0.339	0.153	0.192	0.172	0.199	0.405	**0.230**
	192	0.357	0.366	0.208	0.242	0.187	0.211	0.440	0.245
	336	0.390	0.389	0.262	0.281	**0.202**	0.221	0.459	0.254
	720	0.459	0.428	0.341	0.334	**0.218**	**0.225**	0.500	0.272
	Avg	0.378	0.381	0.241	0.262	0.195	0.214	0.451	0.250
r Flow Attention	96	0.309	0.342	**0.152**	**0.190**	0.196	0.200	0.398	0.231
	192	0.359	0.368	0.206	0.240	0.209	0.219	0.433	0.246
	336	0.391	0.390	**0.257**	0.280	0.218	0.228	0.456	0.254
	720	0.457	0.428	0.346	0.337	0.231	0.230	0.503	0.273
	Avg	0.379	0.382	0.240	0.262	0.214	0.219	0.448	0.251

datasets represent relatively small-, medium-, and large-scale multivariate forecasting scenarios, respectively. All experiments use a lookback window of $T = 96$ and a prediction horizon of $\tau = 96$. The batch size is set to 64 for ETTm1 and 32 for Weather and Electricity.

As shown in Tables 15, 16, and 5, Navigator consistently achieves a favorable trade-off between forecasting accuracy and computational cost across different channel scales. On ETTm1, Navigator delivers strong predictive performance with moderate computational requirements. As the number of channels increases to 21 on Weather, Navigator maintains superior forecasting accuracy while requiring substantially fewer MACs and less memory than Transformer-based baselines. This trend remains evident on Electricity with 321 channels: despite the increased computational demand associated with the larger channel scale, Navigator remains considerably more efficient than Transformer-based models in terms of MACs and training time, while incurring only moderate additional cost compared with lightweight approaches such as FilterNet and RAFT. Overall, these results show that Navigator preserves a favorable accuracy–efficiency trade-off as the number of channels increases, supporting its applicability to time series with different channel scales.

C.7 Parameter Sensitivity Analysis

The number of encoder layers L_e. Figure 3 investigates the sensitivity of Navigator to the number of encoder layers in BendingFormer. Across the four ETT datasets, forecasting performance varies only slightly as L_e increases from 1 to 6, indicating that Navigator is generally insensitive to this hyperparameter. The best results are typically achieved with a relatively small number of layers, while further increasing the depth brings little additional benefit and may even lead to slight performance degradation. This suggests that effective cross-channel interaction can be achieved with a shallow BendingFormer, without relying on a deep encoder architecture.



(a) ETTh1 (b) ETTh2 (c) ETTm1 (d) ETTm2

Figure 3: Sensitivity analysis of the number of encoder layers L_e under the setting of lookback window $T = 96$ and prediction horizon $\tau = 96$.

Hidden dimensions d_{model} and d_{ff} Figures 4 and 5 examine the sensitivity of Navigator to the hidden dimensions d_{model} and d_{ff}, respectively. For d_{model}, increasing the dimension generally improves performance at first, while further enlargement brings only marginal gains or even slight degradation on some datasets. In comparison, the results are less sensitive to d_{ff}, with MSE and MAE remaining

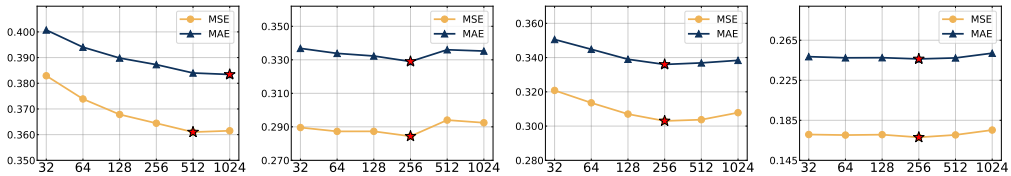
Table 12: Performance comparison of representative models with cross-channel modeling capabilities for long-term forecasting. The lookback window is fixed at $T = 96$, and the prediction horizons are set to $\tau \in \{96, 192, 336, 720\}$. Results are averaged over the four prediction horizons. The best and second-best results are highlighted in **red** and **blue**, respectively.

Models		**Navigator**		TimeBridge		TQNet		Leddam		CARD		iTransformer		Crossformer	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.361**	**0.384**	0.373	0.395	**0.371**	0.393	0.377	0.394	0.383	**0.391**	0.386	0.405	0.423	0.448
	192	**0.414**	**0.413**	0.428	0.426	0.428	0.426	**0.424**	0.422	0.435	**0.420**	0.441	0.436	0.471	0.474
	336	**0.454**	**0.437**	0.471	0.451	0.476	0.446	**0.459**	**0.442**	0.479	**0.442**	0.487	0.458	0.570	0.546
	720	**0.460**	**0.460**	0.495	0.487	0.487	0.470	**0.463**	**0.459**	0.471	0.461	0.503	0.491	0.653	0.621
	Avg	**0.422**	**0.424**	0.442	0.440	0.441	0.434	**0.431**	**0.429**	0.442	**0.429**	0.454	0.447	0.529	0.522
ETTh2	96	**0.284**	**0.329**	0.294	0.344	0.295	0.343	0.292	0.343	**0.281**	**0.330**	0.297	0.349	0.745	0.584
	192	**0.359**	**0.378**	0.371	0.394	0.367	0.393	0.367	0.389	**0.363**	**0.381**	0.380	0.400	0.877	0.656
	336	**0.401**	**0.411**	0.421	0.429	0.417	0.427	0.412	0.424	**0.411**	**0.418**	0.428	0.432	1.043	0.731
	720	**0.410**	**0.430**	0.423	0.443	0.433	0.446	0.419	0.438	**0.416**	**0.431**	0.427	0.445	1.104	0.763
	Avg	**0.364**	**0.387**	0.377	0.403	0.378	0.402	0.373	0.399	**0.368**	**0.390**	0.383	0.407	0.942	0.684
ETTm1	96	**0.303**	**0.336**	0.323	0.361	**0.311**	**0.353**	0.319	0.359	0.316	0.347	0.334	0.368	0.404	0.426
	192	**0.354**	**0.362**	0.366	0.385	**0.356**	0.378	0.369	0.383	0.363	**0.370**	0.377	0.391	0.450	0.451
	336	**0.385**	**0.384**	0.398	0.408	**0.390**	0.401	0.394	0.402	0.392	**0.390**	0.426	0.420	0.532	0.515
	720	**0.450**	**0.423**	0.461	0.445	**0.452**	0.440	0.460	0.442	0.458	**0.425**	0.491	0.459	0.666	0.589
	Avg	**0.373**	**0.376**	0.387	0.400	**0.377**	0.393	0.386	0.397	0.383	**0.384**	0.407	0.410	0.513	0.496
ETTm2	96	**0.168**	**0.246**	0.177	0.259	0.173	0.256	0.176	0.257	**0.169**	**0.248**	0.180	0.264	0.287	0.366
	192	**0.233**	**0.289**	0.243	0.303	0.238	0.298	0.243	0.303	**0.234**	**0.292**	0.250	0.309	0.414	0.492
	336	**0.293**	**0.328**	0.303	0.343	0.301	0.340	0.303	0.341	**0.294**	**0.339**	0.311	0.348	0.597	0.542
	720	**0.389**	**0.386**	0.401	0.399	0.397	0.396	0.400	0.398	**0.390**	**0.388**	0.412	0.407	1.730	1.042
	Avg	**0.271**	**0.312**	0.281	0.326	0.277	0.323	0.281	0.325	**0.272**	**0.317**	0.288	0.332	0.757	0.610
Electricity	96	**0.133**	**0.223**	0.142	0.239	**0.134**	**0.229**	0.141	0.235	0.141	0.233	0.148	0.240	0.219	0.314
	192	**0.152**	**0.240**	0.161	0.257	**0.154**	**0.247**	0.159	0.252	0.160	0.250	0.162	0.253	0.231	0.322
	336	**0.165**	**0.254**	0.182	0.278	**0.169**	0.264	0.173	0.268	0.173	**0.263**	0.178	0.269	0.246	0.337
	720	**0.192**	**0.281**	0.217	0.309	0.201	0.294	0.201	0.295	**0.197**	**0.284**	0.225	0.317	0.280	0.363
	Avg	**0.161**	**0.250**	0.176	0.271	**0.164**	0.259	0.169	0.263	0.168	**0.258**	0.178	0.270	0.244	0.334
Traffic	96	0.396	**0.230**	**0.391**	0.268	0.413	**0.261**	0.426	0.276	0.419	0.269	**0.395**	0.268	0.522	0.290
	192	0.426	**0.243**	**0.418**	0.276	0.432	**0.271**	0.458	0.289	0.443	0.276	**0.417**	0.276	0.530	0.293
	336	0.448	**0.252**	**0.432**	0.284	0.450	**0.277**	0.486	0.297	0.460	0.283	**0.433**	0.283	0.558	0.305
	720	0.490	**0.271**	**0.464**	0.301	0.486	**0.295**	0.498	0.313	0.490	0.299	**0.467**	0.302	0.589	0.328
	Avg	0.440	**0.249**	**0.426**	0.282	0.445	**0.276**	0.467	0.294	0.453	0.282	**0.428**	0.282	0.550	0.304
Weather	96	**0.153**	**0.192**	0.168	0.211	0.157	0.200	0.156	0.202	**0.150**	**0.188**	0.174	0.214	0.158	0.230
	192	**0.200**	**0.235**	0.214	0.254	0.206	0.245	0.207	0.250	**0.202**	**0.238**	0.221	0.254	0.206	0.277
	336	**0.258**	**0.279**	0.273	0.297	0.262	0.287	0.262	0.291	**0.260**	**0.282**	0.278	0.296	0.272	0.335
	720	**0.337**	**0.331**	0.353	0.347	0.344	**0.342**	**0.343**	0.343	**0.343**	0.353	0.358	0.347	0.398	0.418
	Avg	**0.237**	**0.259**	0.252	0.277	0.242	0.269	0.242	0.272	**0.239**	**0.261**	0.258	0.278	0.259	0.315
Solar-Energy	96	**0.158**	**0.193**	0.192	**0.207**	**0.173**	0.233	0.197	0.241	0.204	0.208	0.203	0.237	0.310	0.331
	192	**0.187**	**0.210**	0.223	**0.231**	**0.199**	0.257	0.231	0.264	0.240	0.232	0.233	0.261	0.734	0.725
	336	**0.205**	**0.220**	0.243	**0.250**	**0.211**	0.263	0.241	0.268	0.264	0.252	0.248	0.273	0.750	0.735
	720	**0.218**	**0.226**	0.250	0.257	**0.209**	0.270	0.250	0.281	0.268	**0.255**	0.249	0.275	0.769	0.765
	Avg	**0.192**	**0.212**	0.227	**0.236**	**0.198**	0.256	0.230	0.264	0.244	0.237	0.233	0.262	0.641	0.639

Table 13: Forecasting performance of Geometric Bending Attention (GBA) when applied to cross-channel and temporal modeling. The lookback window is fixed at $T = 96$, and the prediction horizons are set to $\tau \in \{96, 192, 336, 720\}$. *GBA-Channel* denotes the original design used in Navigator for cross-channel interaction, while *GBA-Time* adopts the temporal patching and embedding scheme of PatchTST (Nie et al., 2023) to construct temporal tokens and replaces the standard attention over these tokens with the proposed geometric attention mechanism.

Attention Variants		GBA-Channel (original)			GBA-Time		
Metrics		MSE	MAE	Parameter	MSE	MAE	Parameter
ETTh1	96	**0.361**	**0.384**	1.729 M	0.367	0.387	1.835 M
	192	**0.414**	**0.413**	1.827 M	0.418	0.415	2.327 M
	336	**0.454**	**0.437**	3.479 M	0.461	**0.437**	3.979 M
	720	**0.460**	**0.460**	2.098 M	0.470	**0.460**	2.598 M
	Avg	**0.422**	**0.424**	2.283 M	0.429	0.425	2.685 M
ETTh2	96	**0.284**	**0.329**	1.263 M	0.286	0.330	1.316 M
	192	0.359	0.378	0.917 M	**0.355**	**0.375**	1.167 M
	336	**0.401**	**0.411**	0.954 M	**0.401**	0.412	1.204 M
	720	**0.410**	**0.430**	1.449 M	0.423	0.436	1.698 M
	Avg	**0.364**	**0.387**	1.146 M	0.366	0.388	1.346 M
ETTm1	96	**0.303**	**0.336**	0.867 M	0.307	0.339	0.920 M
	192	**0.354**	**0.362**	0.917 M	0.358	0.364	1.167 M
	336	**0.385**	**0.384**	0.954 M	0.393	0.389	1.204 M
	720	**0.450**	**0.423**	1.315 M	0.457	0.425	1.565 M
	Avg	**0.373**	**0.376**	1.013 M	0.379	0.379	1.214 M
ETTm2	96	**0.168**	**0.246**	0.867 M	0.171	0.248	0.920 M
	192	**0.233**	**0.289**	0.917 M	0.240	0.294	1.167 M
	336	**0.293**	**0.328**	0.954 M	0.294	0.330	1.204 M
	720	**0.389**	**0.386**	0.657 M	0.391	0.388	0.907 M
	Avg	**0.271**	**0.312**	0.849 M	0.274	0.315	1.050 M
Weather	96	**0.153**	**0.192**	1.731 M	0.165	0.207	1.837 M
	192	**0.200**	**0.235**	3.408 M	0.212	0.245	3.907 M
	336	**0.258**	**0.279**	1.904 M	0.261	0.281	2.403 M
	720	**0.337**	**0.331**	2.101 M	0.340	0.334	2.600 M
	Avg	**0.237**	**0.259**	2.286 M	0.245	0.267	2.687 M
Electricity	96	**0.133**	**0.223**	4.995 M	0.148	0.232	5.494 M
	192	**0.152**	**0.240**	3.466 M	0.165	0.248	3.966 M
	336	**0.165**	**0.254**	3.540 M	0.181	0.265	4.039 M
	720	**0.192**	**0.281**	3.737 M	0.221	0.298	4.236 M
	Avg	**0.161**	**0.250**	3.935 M	0.179	0.261	4.434 M

relatively stable over a wide range of settings. Overall, moderate hidden dimensions are sufficient to achieve strong forecasting performance, indicating that Navigator does not rely on excessively large representations.



(a) ETTh1 (b) ETTh2 (c) ETTm1 (d) ETTm2

Figure 4: Sensitivity analysis of the hidden dimension d_{model} under the setting of lookback window $T = 96$ and prediction horizon $\tau = 96$.

Table 14: Forecasting performance with Predictability Basis computed from varying proportions of the training data, denoted as 'M_ratio'. All experiments are trained on the full training set. The lookback window is fixed at $T = 96$. The prediction horizons are set to $\tau \in \{12, 24, 48, 96\}$ for PEMSD3 and PEMSD4, and $\tau \in \{96, 192, 336, 720\}$ for the remaining datasets.

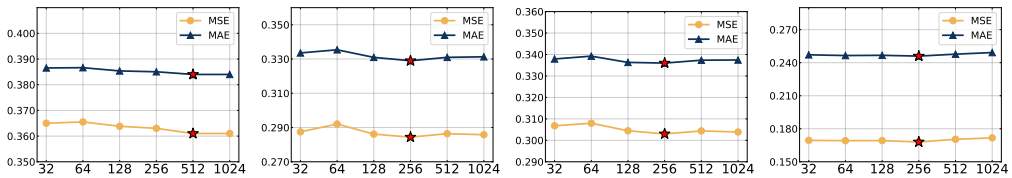
M_ratio		100%		80%		60%		40%		20%		10%	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.361	0.384	0.362	0.383	0.364	0.385	0.364	0.385	0.363	0.385	0.364	0.384
	192	0.414	0.413	0.414	0.413	0.414	0.413	0.414	0.413	0.414	0.413	0.414	0.413
	336	0.454	0.437	0.455	0.437	0.455	0.437	0.453	0.436	0.455	0.437	0.455	0.437
	720	0.460	0.460	0.459	0.460	0.460	0.461	0.461	0.461	0.460	0.460	0.460	0.460
	Avg	0.422	0.424	0.423	0.423	0.423	0.424	0.423	0.424	0.423	0.424	0.423	0.424
ETTh2	96	0.284	0.329	0.285	0.330	0.284	0.329	0.285	0.330	0.286	0.330	0.284	0.330
	192	0.359	0.378	0.360	0.378	0.357	0.377	0.360	0.379	0.360	0.379	0.359	0.378
	336	0.401	0.411	0.407	0.414	0.404	0.413	0.406	0.413	0.403	0.412	0.401	0.411
	720	0.410	0.430	0.417	0.433	0.419	0.434	0.417	0.434	0.414	0.432	0.410	0.430
	Avg	0.364	0.387	0.367	0.389	0.366	0.388	0.367	0.389	0.366	0.388	0.364	0.387
Weather	96	0.153	0.192	0.153	0.192	0.153	0.191	0.152	0.191	0.152	0.191	0.153	0.192
	192	0.200	0.235	0.199	0.233	0.200	0.235	0.200	0.234	0.201	0.237	0.202	0.237
	336	0.258	0.279	0.257	0.279	0.258	0.279	0.259	0.280	0.258	0.280	0.259	0.280
	720	0.337	0.331	0.337	0.331	0.338	0.331	0.337	0.331	0.337	0.331	0.337	0.331
	Avg	0.237	0.259	0.237	0.259	0.237	0.259	0.237	0.259	0.237	0.260	0.238	0.260
Electricity	96	0.133	0.223	0.133	0.224	0.133	0.223	0.133	0.223	0.135	0.225	0.133	0.223
	192	0.152	0.240	0.152	0.240	0.152	0.239	0.152	0.240	0.152	0.240	0.152	0.239
	336	0.165	0.254	0.164	0.253	0.164	0.254	0.164	0.253	0.164	0.254	0.165	0.254
	720	0.192	0.281	0.191	0.280	0.193	0.282	0.191	0.280	0.192	0.281	0.191	0.280
	Avg	0.161	0.250	0.160	0.249	0.161	0.250	0.160	0.249	0.161	0.250	0.160	0.249
PEMSD3	12	0.065	0.162	0.065	0.162	0.064	0.162	0.064	0.162	0.065	0.163	0.065	0.163
	24	0.078	0.179	0.078	0.179	0.078	0.178	0.078	0.178	0.077	0.178	0.077	0.178
	48	0.096	0.201	0.098	0.203	0.098	0.203	0.098	0.203	0.099	0.204	0.098	0.204
	96	0.123	0.228	0.121	0.227	0.122	0.228	0.120	0.227	0.122	0.227	0.123	0.229
	Avg	0.091	0.193	0.091	0.193	0.091	0.193	0.090	0.193	0.091	0.193	0.091	0.194
PEMSD4	12	0.068	0.164	0.068	0.164	0.068	0.163	0.068	0.163	0.068	0.164	0.068	0.164
	24	0.076	0.173	0.076	0.173	0.076	0.173	0.076	0.174	0.076	0.173	0.075	0.172
	48	0.085	0.184	0.085	0.185	0.085	0.184	0.085	0.185	0.085	0.184	0.086	0.185
	96	0.096	0.197	0.097	0.197	0.097	0.197	0.096	0.196	0.095	0.195	0.096	0.196
	Avg	0.081	0.180	0.082	0.180	0.082	0.179	0.081	0.180	0.081	0.179	0.081	0.179

Table 15: Efficiency comparison between Navigator and other models on the ETTm1 dataset with lookback window T = 96 and prediction horizon τ = 96. Training Time denotes the average time required per epoch for the model.

Model	Parameters	MACs	Training Time	Memory Usage	MSE	MAE
PatchLGA	42.13 K	2.3 M	18.8 s	1598 MB	0.327	0.357
DistDF	8.59 M	112.38 M	13.7 s	1816 MB	0.316	0.357
VPNet	3.3 M	81.65 M	5.9 s	2490 MB	0.313	0.342
RAFT	44.26 K	309.79 K	3.8 s	1414 MB	0.325	0.369
FreDF	224.22 K	1.57 M	10.1 s	1878 MB	0.324	0.362
FilterNet	71.79 K	408.48 K	6.5 s	1404 MB	0.321	0.361
FreTS	3.24 M	22.28 M	4.9 s	1794 MB	0.339	0.374
iTransformer	1.24 M	8.65 M	12.9 s	1922 MB	0.334	0.368
PatchTST	1.49 M	102.06 M	13.8 s	2144 MB	0.329	0.367
TimeXer	2.18 M	65.2 M	15.5 s	2048 MB	0.318	0.356
Navigator(ours)	867.48 K	6.14 M	12.2 s	1498 MB	0.303	0.336

Table 16: Efficiency comparison between Navigator and other models on the Weather dataset with lookback window T = 96 and prediction horizon τ = 96. Training Time denotes the average time required per epoch for the model.

Model	Parameters	MACs	Training Time	Memory Usage	MSE	MAE
PatchLGA	42.13 K	6.9 M	33.2 s	1638 MB	0.174	0.213
DistDF	518.84 K	80.1 M	30.8 s	2136 MB	0.164	0.209
VPNet	8.06 M	387.31 M	13.3 s	2296 MB	0.157	0.192
RAFT	44.26 K	929.38 K	6.8 s	1396 MB	0.185	0.241
FreDF	4.83 M	101.44 M	26.1 s	2098 MB	0.164	0.202
FilterNet	71.82 K	1.23 M	11.8 s	1406 MB	0.162	0.207
FreTS	3.24 M	66.85 M	7.9 s	1926 MB	0.184	0.239
iTransformer	4.83 M	101.44 M	18.0 s	2012 MB	0.174	0.214
PatchTST	5.33 M	1.21 G	22.6 s	2506 MB	0.177	0.218
TimeXer	8.3 M	771.04 M	28.3 s	2336 MB	0.157	0.205
Navigator(ours)	1.73 M	36.95 M	10.5 s	1940 MB	0.153	0.192



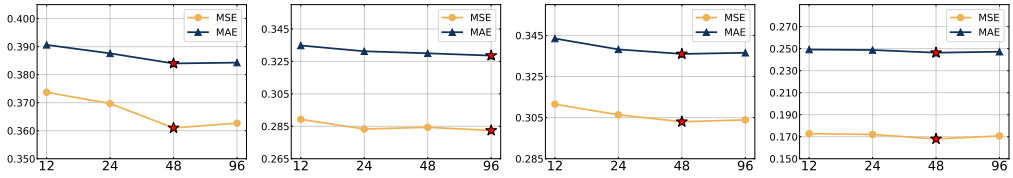
(a) ETTh1 (b) ETTh2 (c) ETTm1 (d) ETTm2

Figure 5: Sensitivity analysis of the hidden dimension d_{ff} under the setting of lookback window $T = 96$ and prediction horizon $\tau = 96$.

D Analysis of the Predictability Basis

D.1 Effect of Predictability Basis Dimension r

The parameter r denotes the dimensionality of the Predictability Basis $\mathbf{M}_r \in \mathbb{R}^{r\times T}$, where $1 \le r \le T$. It determines how many predictive temporal directions are retained to form the compact temporal space used by Foreseer. As shown in Figure 6, increasing r from a small value generally improves forecasting performance at first, indicating that retaining more predictive temporal directions can provide richer forecasting-relevant information. However, the improvement gradually saturates, and excessively large r brings little additional benefit or may slightly degrade performance. This suggests that a compact subset of predictive temporal directions is sufficient, while retaining too many directions may introduce less informative or redundant information. Overall, Navigator shows relatively stable performance over a broad range of r, with intermediate settings providing a favorable balance between information preservation and compactness.



(a) ETTh1 (b) ETTh2 (c) ETTm1 (d) ETTm2

Figure 6: Effect of Predictability Basis Dimension r on forecasting performance across four ETT datasets. The lookback window and prediction horizon are fixed at $T = 96$ and $\tau = 96$, respectively.

D.2 Visualizations of Predictability Basis $\mathbf{M}_r$

Figure 7 visualizes the Predictability Bases computed from different datasets. Clear structural differences can be observed across datasets. For example, Electricity and Traffic exhibit relatively regular and repeated patterns, whereas Solar-Energy, Weather, and the PEMS datasets show more heterogeneous structures. These distinct patterns indicate that the Predictability Basis adapts to the statistical characteristics of each dataset rather than imposing a fixed temporal representation. Within each basis, different columns also exhibit diverse temporal patterns across the historical window, reflecting different temporal directions identified from the training data. This further suggests that the predictive information contained in historical observations is distributed non-uniformly across different directions in the temporal space. Overall, the visualizations provide qualitative evidence that the Predictability Basis captures dataset-specific predictive structure and organizes forecasting-relevant historical information into a compact set of temporal directions.

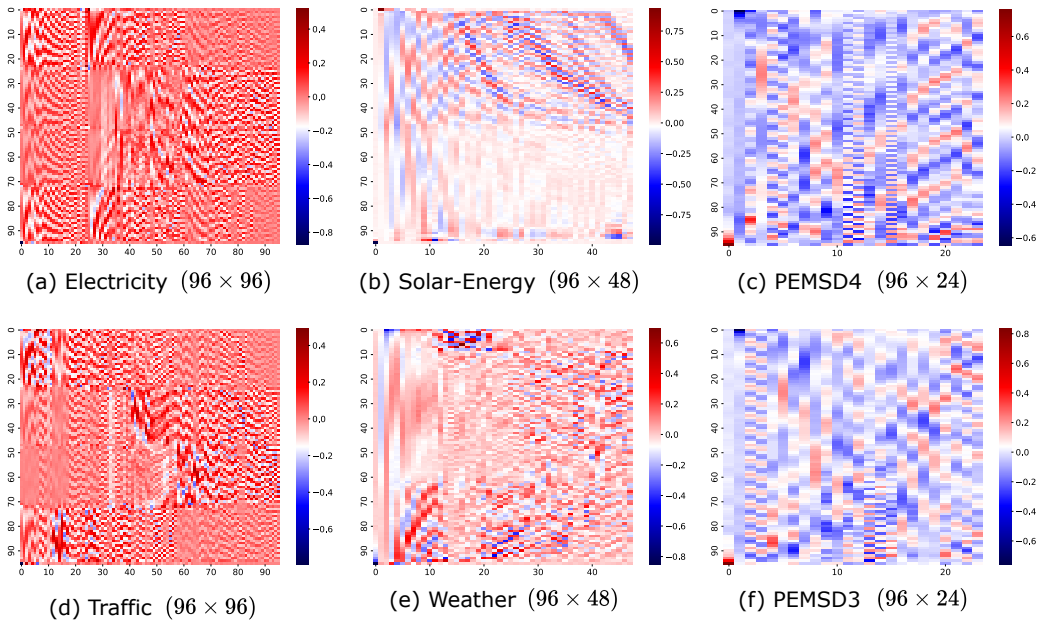
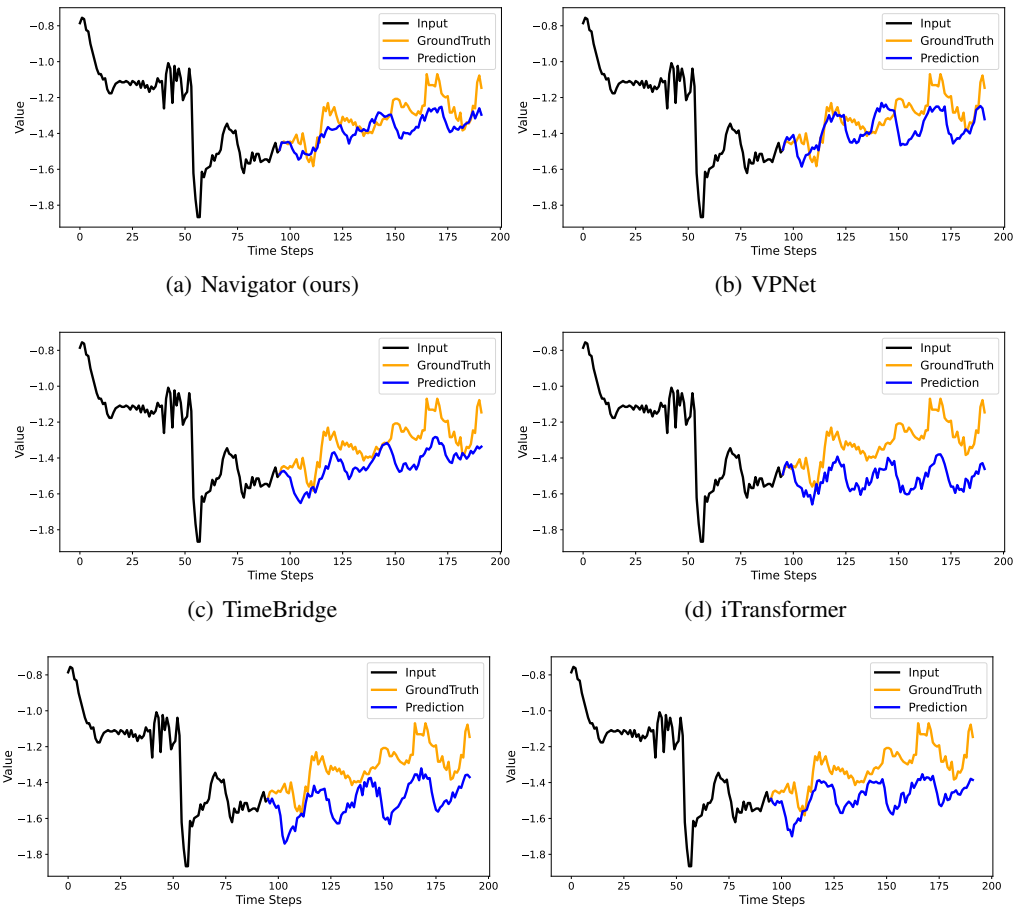


Figure 7: Visualizations of Predictability Bases $\mathbf{M}_r^\top \in \mathbb{R}^{T \times r}$ computed from the training data of different datasets under the input/prediction length setting of $96/96$. Each row corresponds to a historical time step, while each column represents a temporal direction. Here, $T = 96$ denotes the lookback window, and r $(1 \leq r \leq T)$ denotes the number of retained temporal directions, whose value is determined for each dataset under the corresponding forecasting setting.

E Visualizations of Predictions

To provide an intuitive comparison with representative state-of-the-art methods, we visualize forecasting results on the ETTh1 and Weather datasets in Figures 8 and 9. We compare Navigator with VPNet (Liu et al., 2026), TimeBridge (Liu et al., 2025), iTransformer (Liu et al., 2024), PAttn (Tan et al., 2024), and PatchTST (Nie et al., 2023), which represent different forecasting architectures. As shown in the figures, Navigator more closely follows the ground-truth trajectories and better preserves temporal variations, yielding more accurate and stable forecasts across the selected examples.

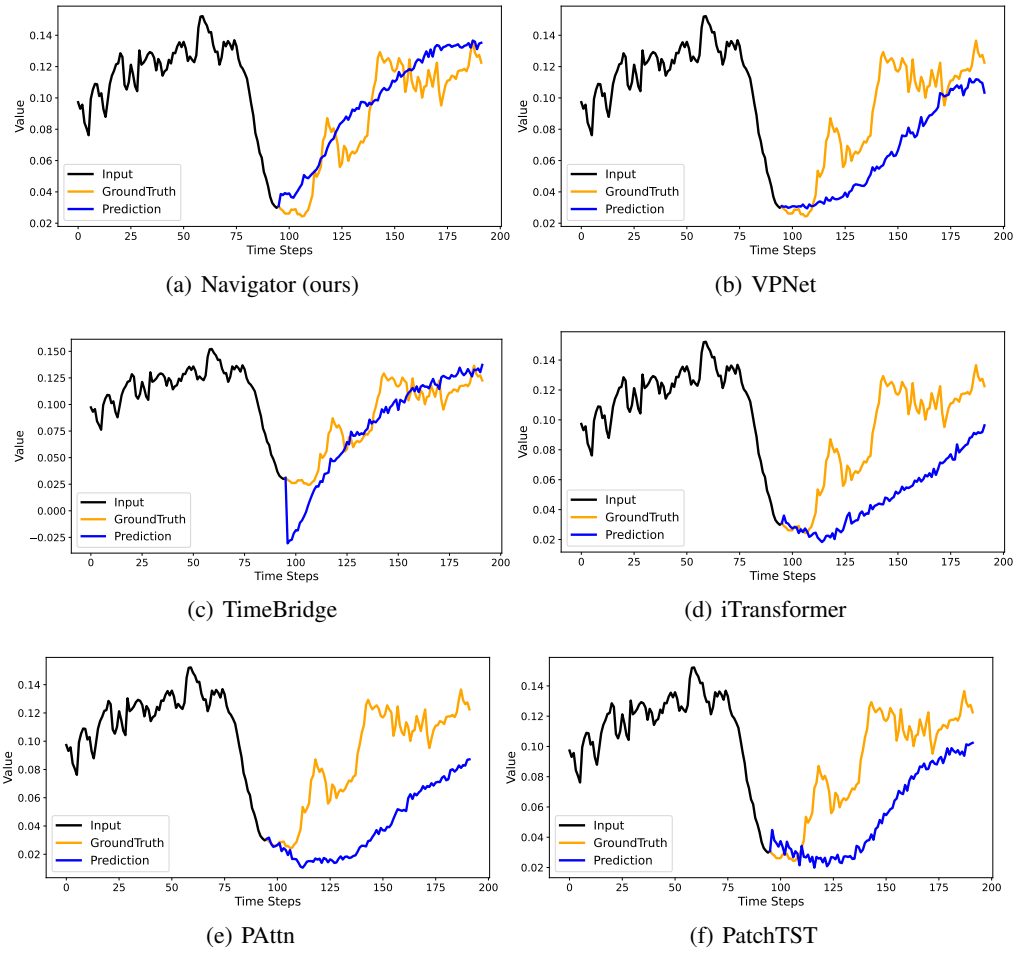


(a) Navigator (ours) (b) VPNet

(c) TimeBridge (d) iTransformer

(e) PAttn (f) PatchTST

Figure 8: Visualization of forecasting results on the ETTh1 dataset with lookback window $T = 96$ and prediction horizon $\tau = 96$.



(a) Navigator (ours) (b) VPNet

(c) TimeBridge (d) iTransformer

(e) PAttn (f) PatchTST

Figure 9: Visualization of forecasting results on the Weather dataset with lookback window $T = 96$ and prediction horizon $\tau = 96$.