QIYAO-I: A MANIFOLD BASED FOUNDATION MODEL FOR IRREGULAR MULTIVARIATE TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Irregular multivariate time series forecasting is a challenging yet important problem in real-world applications, where observations are often irregularly sampled and asynchronously recorded across variables. Existing time series foundation models are mostly built on regularly sampled sequences, making them difficult to generalize to irregular time intervals and asynchronous cross-variable dependencies. To address these challenges, we propose **QiYao-I**, a manifold based foundation model for irregular multivariate time series forecasting. Specifically, we introduce a novel sampling-conditioned temporal manifold attention mechanism that maps real timestamps into a learnable temporal manifold feature space and injects temporal manifold biases into attention layers, enabling the model to capture both irregular time intervals and local sampling structures. Further, we propose a dynamic variable interaction mechanism with frequency awareness. It selectively performs cross-variable message passing under asynchronous observations. Extensive experiments on real-world irregular multivariate forecasting benchmarks demonstrate that QiYao-I achieves superior performance compared with both time series foundation models and end-to-end irregular forecasting models, showing strong generalization ability in zero-shot and few-shot settings.

1 INTRODUCTION

In recent years, foundation models have been reshaping the paradigm of time series modeling. Unlike traditional models designed for a single dataset or a specific task, time series foundation models are typically pretrained on large-scale data to learn transferable and generalizable temporal representations, enabling unified modeling across diverse forecasting tasks.

However, most existing time series foundation models assume regular multivariate time series (MTS), where two pairs of observations are equally spaced and same time intervals Δ_t contain the same number of sampled points, as shown in Fig. 1(a). Similarly, different variables are observed across fixed regular temporal intervals, enabling cross-variable interactions either at the same timestamp or across multiple sampling intervals whose durations are the same, as shown in Fig. 1(c)).

Nevertheless, this assumption does not hold in many real world scenarios. Real-world systems often produce irregular multivariate time series (IMTS). For example, sensor failures may lead to irregular intervals, and different clinical procedures yield variable-specific asynchronous observations. As shown in Fig. 1(b), adjacent observations in IMTS have irregular intervals. Even if two pairs of observations span the same time gap Δ_t, their intervals have different numbers of sampled points. Moreover, Fig. 1(d) shows that different variables are observed at variable-specific irregular timestamps, so that cross-variable effects occur distinctively across time intervals.

Although existing time series foundation models have shown strong generalization ability in regularly sampled settings, they still rely on regular temporal structures and aligned variable observations, making direct transfer to IMTS challenging.

Challenge 1: Limited generalization to irregular time steps. Most mainstream time series foundation models (Khwaja et al., 2026; Ansari et al., 2025; Liu et al., 2025d; Das et al., 2023; Liu et al., 2024b) treat observations as equally spaced tokens and use discrete positional embeddings.

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.

These embeddings capture token order but not actual time intervals: under irregular sampling, two observation pairs with the same token order may correspond to very different real time intervals, leading to incorrect dependency learning and biased generalization. Although MIRA (Li et al., 2026) introduces scalar timestamps into positional embeddings, it does not fully account for local sampling structure. As illustrated in Fig. 1(b), the same time interval can imply different dependencies in densely and sparsely sampled regions. Therefore, simply introducing scalar timestamps is not enough to fully address the generalization problem under irregular time steps.

Challenge 2: Limited cross-variable interaction under asynchronous observations. Channel-independent models (Liu et al., 2025d; Das et al., 2023; Liu et al., 2024b; Wang et al., 2025a;b) process variables separately, failing to capture cross-variable dependencies. Channel-dependent models (Khwaja et al., 2026; Ansari et al., 2025; Woo et al., 2024) enable such exchange but typically assume observations are aligned within shared time intervals, as shown in Fig. 1(c). In IMTS, variables are observed asynchronously, and dependencies can span distinct intervals with different sampling structures, as shown in Fig. 1(d). As a result, existing time series foundation models still lack a transferable interaction mechanism that can capture asynchronous cross-variable dependencies.

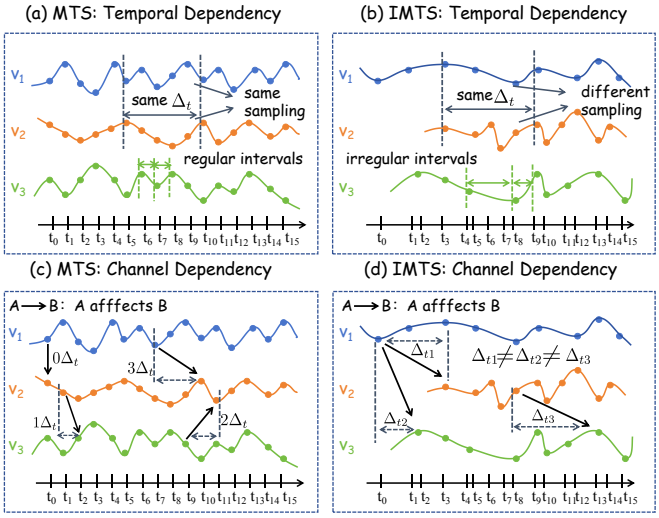

Figure 1: The difference between regular multivariate time series (MTS) and irregular multivariate time series (IMTS).

To address the challenge of irregular temporal modeling in foundation model generalization, we introduce *a novel sampling-conditioned temporal manifold attention mechanism*. Specifically, the temporal manifold refers to a learnable space that embeds scalar timestamps. Instead of relying on dataset-specific interpolation, this mechanism maps scalar timestamps into a learnable temporal manifold. It models temporal relations by learning manifold distances and converts these distances into temporal manifold biases. Furthermore, the model introduces a Gaussian-kernel descriptor to distinguish observation pairs, which span the same time interval but have different numbers of sampled points by adaptively adjusting the bandwidth of these temporal manifold biases. By injecting temporal manifold biases into attention logits, the model can learn transferable temporal dependencies across heterogeneous IMTS data with different sampling structures. **To address the challenge of asynchronous cross-variable modeling for foundation models**, we propose *a frequency-guided dynamic variable interaction mechanism*. Observations from different variables are not aligned, and their dependencies involve multiple points within a local interval. We therefore group observations into patches for cross-variable interaction. To identify related patch intervals across asynchronously observed variables, we compute temporal distances between patches and select the top-K nearest patches from other variables. However, being close in time does not indicate whether their observed values follow similar patterns, so giving all selected patches equal weight may introduce unrelated information. We therefore use learnable spectrum projection and aggregation to extract patch-level and global-level frequency features directly from irregular values and timestamps. Patch-level features capture local patterns within each selected patch, while variable-level features summarize broader patterns. Their similarities jointly determine the aggregation weights to emphasize information from patches with similar local and global temporal patterns.

To this end, we propose **QiYao-I**, a unified foundation model framework designed for irregular multivariate time series. Our main contributions are summarized as follows:

- We propose QiYao-I, which jointly models irregular temporal dependencies and asynchronous cross-variable interactions in a unified framework.
- We propose *a novel sampling-conditioned temporal manifold time attention*, which captures irregular time intervals and distinguishes the same interval under different sampling structure.

- We propose *a frequency-guided dynamic variable interaction mechanism* that selects patches from other variables based on temporal distance and weights their contributions by frequency similarity, enabling selective interactions across asynchronous variables.
- Extensive experiments show that QiYao-I achieves superior forecasting performance and strong generalization ability in both zero-shot and few-shot settings.

2 Related Works

End-to-End Irregular Time Series Modeling. Irregular time series modeling aims to handle observations sampled at uneven and asynchronous timestamps. Existing end-to-end methods mainly include time-aware recurrent models (Che et al., 2016; Baytas et al., 2017), continuous-time models Latent ODE (Rubanova et al., 2019), Neural CDE (Kidger et al., 2020), as well as set-, graph-, attention-, and patch-based models (Horn et al., 2020; Zhang et al., 2021; Shukla & Marlin, 2021; Zhang et al., 2024). These methods improve the modeling of irregular observations from different perspectives and often perform well when trained on a specific target dataset. However, most of them are task-specific models and are not designed for large-scale pretraining or zero-shot transfer. Different from these methods, QiYao-I is built as a foundation model for irregular multivariate time series. It directly models irregular timestamps in a learnable temporal manifold.

Time Series Foundation Models. Recent time series foundation models have achieved promising zero-shot and few-shot forecasting performance through large-scale pretraining. Representative models include Toto2 (Khwaja et al., 2026), Chronos2 (Ansari et al., 2025), TimesFM2.5 (Das et al., 2023), MOIRAI (Woo et al., 2024), and MoE-based models such as Time-MoE and MOIRAI-MoE (Shi et al., 2025; Liu et al., 2024a). However, most existing foundation models are designed for regularly sampled time series. They usually regard a sequence as equally spaced tokens and use discrete positional encodings to model temporal order. When applied to irregular time series, these models often require resampling or interpolation, which may change the original sampling structure and introduce extra bias. Although MIRA (Li et al., 2026) introduces continuous-time modeling for irregular medical time series, how to model local sampling structure remains underexplored. In contrast, QiYao-I supports irregular inputs from pretraining to inference. By injecting manifold-distance biases into attention, QiYao-I can better capture irregular time intervals and transfer across datasets with different sampling structure.

Variable Interaction in Foundation Models. Cross-variable interaction is important for multivariate time series, but it remains insufficiently explored in time series foundation models. Many models, such as TimesFM2.5 (Das et al., 2023), Lag-Llama (Rasul et al., 2023), Timer (Liu et al., 2024b), and Sundial (Liu et al., 2025d), AimTS (Chen et al., 2025) mainly adopt a channel-independent paradigm. This design limits explicit information exchange among variables. Some models, such as Chronos2 (Ansari et al., 2025), Toto2 (Khwaja et al., 2026), and MOIRAI (Woo et al., 2024) support multivariate forecasting, but their interactions are usually based on dense attention or aligned time intervals. These designs are more suitable for regular multivariate time series, where variables share the same temporal grid. In irregular multivariate time series, variables are often observed asynchronously and may have different sampling frequencies, and periodic patterns. Dense or fixed interaction may therefore introduce noisy correlations. QiYao-I addresses this problem by constructing dynamic patch-level variable interactions with frequency awareness. Instead of forcing all variables to interact, QiYao-I selects relevant source patches for each target patch, enabling adaptive message passing across asynchronous variables.

3 QiYao-I

3.1 Framework Overview

We develop **QiYao-I**, a foundation model for irregular multivariate time series forecasting. An irregular multivariate time series (IMTS) contains V variables observed at uneven and asynchronous timestamps. After padding, its historical observations are represented by values $\mathbf{X}^h \in \mathbb{R}^{V\times L_h}$, timestamps $\mathbf{T}^h \in \mathbb{R}^{V\times L_h}$, and validity masks $\mathbf{M}^h \in \{0,1\}^{V\times L_h}$. Forecasting queries are specified by future timestamps $\mathbf{T}^p \in \mathbb{R}^{V\times L_p}$ and masks $\mathbf{M}^p \in \{0,1\}^{V\times L_p}$. Here, L_h and L_p are the

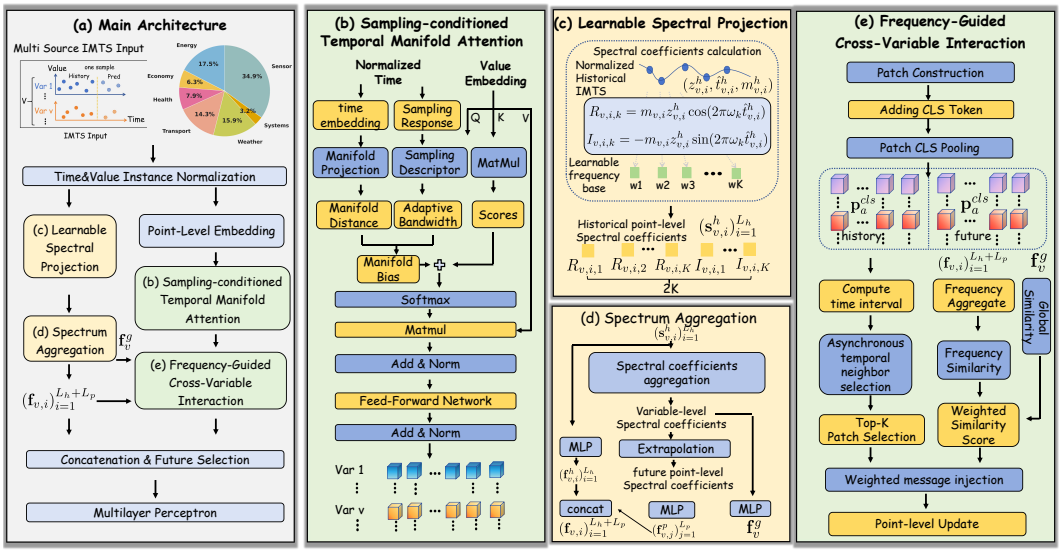

Figure 2: Overall architecture of QiYao-I.

maximum historical and future lengths; the masks indicate valid observations and prediction queries, respectively. The forecasting objective is to learn $\widehat{\mathbf{X}}^p = f_\theta\left(\mathbf{X}^h, \mathbf{T}^h, \mathbf{M}^h, \mathbf{T}^p, \mathbf{M}^p\right)$,

As illustrated in Figure 2, three components model irregular temporal dependencies and asynchronous cross-variable interactions for cross-dataset generalization. First, *Sampling-Conditioned Temporal Manifold Attention* maps timestamps into a learnable temporal manifold and constructs temporal manifold biases from pairwise manifold distances to model irregular temporal dependencies under different sampling structures. Second, to capture periodic patterns and guide selective patch-level interaction, *Learnable Irregular Spectral Projection and Spectrum Aggregation* extracts frequency representations directly from irregular historical observations and extrapolates these representations to future query timestamps. Third, *Frequency-Guided Cross-Variable Interaction* selects temporally relevant patches from other variables and weights them according to local and global frequency similarity to model asynchronous cross-variable interaction. Finally, each future temporal state is fused with its extrapolated frequency representations to produce a forecast.

3.2 Sampling-Conditioned Temporal Manifold Attention

To model irregular temporal dependencies, prior work uses continuous-time RoPE (Su et al., 2024; Li et al., 2026) to model irregular dependency, but its rotary transformations do not describe the local sampling structure. We introduce sampling-conditioned temporal manifold attention, which maps timestamps into a shared learnable space, converts pairwise distances into attention biases, and adapts their bandwidth to each input's local sampling structure. Through multi-source pre-training, the model learns generalizable patterns across different sampling structures, allowing temporal biases to adapt to diverse observation patterns without dataset-specific temporal interpolation.

Bandwidth concept. For a Gaussian kernel $K_\sigma(\delta) = \exp[-\delta^2/(2\sigma^2)]$, its logarithm gives a negative distance penalty. A larger bandwidth σ weakens this penalty at a fixed distance. Our temporal bias follows this log-kernel form, with a bandwidth conditioned on local sampling structure.

Normalization and token embedding. For each variable $v \in [V]$, we compute the historical mean μ_v and standard deviation s_v from valid observations only, and normalize values as $z_{v,i} = (x^h_{v,i} - \mu_v)/s_v$. We use $s_v = 1$ for insufficient history and $\mu_v = 0$ when history is empty. Following the future masked token formulation (Ansari et al., 2025), we represent each future position with a learnable mask token. The historical and future tokens are initialized as:

$$\mathbf{c}^{h,(0)}_{v,i} = \mathbf{W}_x \operatorname{asinh}(z_{v,i}) + \mathbf{b}_x, \qquad \mathbf{c}^{p,(0)}_{v,j} = \mathbf{e}_{\text{mask}}, \tag{1}$$

where $i \in [L_h]$, $j \in [L_p]$, $\mathbf{W}_x \in \mathbb{R}^{d\times 1}$, and $\mathbf{b}_x, \mathbf{e}_{\text{mask}} \in \mathbb{R}^d$.

To place timestamps from heterogeneous datasets on a common scale, each historical or future timestamp is normalized as $\widehat{t} = R(t - t_0)/(t_b - t_0)$, where t_0 and t_b denote the start and historical boundary of the sampled window, respectively, and $R > 0$ is a fixed reference scale. This transformation maps the historical window to $[0, R]$, while allowing future timestamps to exceed R. It preserves relative temporal gaps and is unaffected by missing observations or the forecast horizon.

Learnable temporal manifold distance. Although continuous-time RoPE (Li et al., 2026) injects scalar $\widehat{t}$ into attention queries and keys, this can not reflect the local sampling structures. We first map each normalized timestamp into a learnable temporal manifold. For layer $\ell \in [N_T]$ and component $k \in \{0, \ldots, d_t - 1\}$, the timestamp embedding is:

$$\mathbf{e}_\ell(\widehat{t}_u)[k] = \begin{cases} \omega_{\ell,0}\widehat{t}_u + \alpha_{\ell,0}, & k = 0, \\ \sin(\omega_{\ell,k}\widehat{t}_u + \alpha_{\ell,k}), & 1 \le k < d_t, \end{cases} \qquad \mathbf{e}_\ell(\widehat{t}_u) \in \mathbb{R}^{d_t}, \tag{2}$$

where N_T is the number of temporal layers, d_t is the time-embedding dimension, and all $\omega_{\ell,k}$ and $\alpha_{\ell,k}$ are learnable. For attention head $r \in [H]$, an MLP $\phi_{\ell,r} : \mathbb{R}^{d_t} \to \mathbb{R}^{d_m}$ produces a head-specific manifold representation. For positions $u, w \in [L_h + L_p]$, their manifold distance is:

$$D_{uw}^{\ell,r} = \left\| \phi_{\ell,r}(\mathbf{e}_\ell(\widehat{t}_u)) - \phi_{\ell,r}(\mathbf{e}_\ell(\widehat{t}_w)) \right\|_2^2, \tag{3}$$

Here, H is the number of attention heads and d_m is the representation dimension. Temporal manifold denotes the learned timestamp representation space; $D_{uw}^{\ell,r}$ is its squared Euclidean distance.

Sampling-conditioned bandwidth. The manifold distance $D_{uw}^{\ell,r}$ depends on the timestamp pair, and we condition the bandwidth of the distance-based attention bias on local sampling structure, so that the same distance can receive different penalties under different observation patterns across datasets.

For position u of variable v, let $\bar{t}_u = \min(\widehat{t}_u, R)$, where R is the normalized historical boundary. A historical position examines observations around its own time; a future position examines observations near R. Within a preceding window of width $W > 0$, the valid historical observations are:

$$\mathcal{I}_u = \left\{ i \in [L_h] : m_{v,i}^h = 1,\ 0 \le \bar{t}_u - \widehat{t}_{v,i}^h \le W \right\}. \tag{4}$$

We place A anchors $\{c_a\}_{a=1}^A$ uniformly over $[-1, 0]$. Each anchor measures the concentration of observations in a different part of this window:

$$S_{u,a} = \log\left[1 + \sum_{i \in \mathcal{I}_u} \exp\left(-\frac{1}{2\kappa^2} \left(\frac{\widehat{t}_{v,i}^h - \bar{t}_u}{W} - c_a \right)^2 \right) \right], \quad a \in [A], \tag{5}$$

where $\kappa > 0$ controls the width of the Gaussian kernels. *The sampling responses* $S_{u,1}, \ldots, S_{u,A}$ can therefore distinguish observation pairs that span the same time interval but have different local sampling structures.

Let $\widehat{t}_u^{\,\text{last}}$ be the latest valid historical time not exceeding $\bar{t}_u$, and let $\mathcal{I}_u^{\le}$ contain all such valid historical observations. We form the *sampling descriptor*:

$$\mathbf{s}_u = \left[S_{u,1}, \ldots, S_{u,A}; \log\left(1 + \frac{(\widehat{t}_u - R)_+}{W} \right); \log\left(1 + \frac{(\widehat{t}_u - \widehat{t}_u^{\,\text{last}})_+}{W} \right); \mathbb{I}(|\mathcal{I}_u^{\le}| > 0) \right]. \tag{6}$$

The descriptor $\mathbf{s}_u \in \mathbb{R}^{A+3}$ combines local sampling responses with the query's distance beyond R, its elapsed time since the latest observation, and a history-availability indicator. If no preceding observation exists, the elapsed-time feature and availability indicator are both set to zero.

A layer- and head-specific MLP converts this descriptor into a positive scale:

$$r_u^{\ell,r} = \exp[\lambda \tanh(g_{\ell,r}(\mathbf{s}_u))], \tag{7}$$

where $\lambda > 0$ bounds its adaptation range. The scale adjusts the penalty derived from the manifold distance:

$$b_{uw}^{\ell,r} = -\frac{\eta_{\ell,r} D_{uw}^{\ell,r}}{\beta_{\ell,r}^2 r_u^{\ell,r} r_w^{\ell,r}}, \qquad \eta_{\ell,r} = \text{softplus}(\gamma_{\ell,r}), \quad \beta_{\ell,r} = \text{softplus}(\sigma_{\ell,r}) + \epsilon. \tag{8}$$

The effective bandwidth is $\beta_{\ell,r}\sqrt{r_u^{\ell,r} r_w^{\ell,r}}$. Consequently, two pairs with the same manifold distance can receive different attention biases when their local sampling structures differ.

Temporal manifold attention. Let $\mathbf{q}_u^{\ell,r}, \mathbf{k}_u^{\ell,r} \in \mathbb{R}^{d_r}$ be the query and key of head r, where $d_r = d/H$. Continuous-time RoPE and the manifold bias are combined as:

$$\widetilde{\mathbf{q}}_u^{\ell,r} = \mathcal{R}(\widehat{t}_u)\mathbf{q}_u^{\ell,r}, \qquad \widetilde{\mathbf{k}}_w^{\ell,r} = \mathcal{R}(\widehat{t}_w)\mathbf{k}_w^{\ell,r},$$
$$a_{uw}^{\ell,r} = \frac{(\widetilde{\mathbf{q}}_u^{\ell,r})^\top \widetilde{\mathbf{k}}_w^{\ell,r}}{\sqrt{d_r}} + b_{uw}^{\ell,r} + \mathcal{M}_{uw}, \qquad h_{uw}^{\ell,r} = \mathrm{softmax}_w(a_{uw}^{\ell,r}), \tag{9}$$

Here, $\mathcal{M}_{uw} = 0$ for valid keys and $-\infty$ for padding. Attention operates separately within each variable over its historical tokens and future query tokens, supporting generalization to datasets with different sampling structures.

3.3 Learnable Irregular Spectral Representation

Temporal attention models dependencies between observations within each variable. To capture periodic patterns that transfer across datasets and guide cross-variable interaction, we learn a shared spectral mapping directly from irregular values and timestamps through multi-source pre-training. It provides local and global frequency representations for comparing variables across different inputs. Standard Discrete Fourier Transforms (DFT) (Sundararajan, 2001) assumes uniformly spaced observations. We instead evaluate sinusoidal bases at the actual normalized timestamps and learn their frequencies, obtaining spectral representations without resampling the input.

Learnable spectral projection. Standard DFT use sine and cosine bases on a fixed frequency grid determined by the sampling interval and sequence length. To extract frequency information from irregular observations, we apply sine and cosine bases with learnable frequencies directly to their timestamps. Let K_f be the number of spectral components and $w_k = \exp(\xi_k) > 0$, $k \in [K_f]$, be learnable frequencies. These frequencies are defined on the normalized window-time scale. For variable $v \in [V]$ and historical position $i \in [L_h]$, we compute the real and imaginary coefficients as:

$$[R_{v,i,k}^{\mathrm{re}}, I_{v,i,k}^{\mathrm{im}}] = m_{v,i}^h z_{v,i}[\cos(2\pi w_k \widehat{t}_{v,i}^h), -\sin(2\pi w_k \widehat{t}_{v,i}^h)], \qquad k \in [K_f]. \tag{10}$$

Following the real–imaginary decomposition of Fourier coefficients, we concatenate both components into a real-valued spectral coefficients $\mathbf{s}_{v,i}^h = [\mathbf{R}_{v,i}^{\mathrm{re}}; \mathbf{I}_{v,i}^{\mathrm{im}}] \in \mathbb{R}^{2K_f}$. A residual MLP and a shared projection $\mathcal{E}_f : \mathbb{R}^{2K_f} \to \mathbb{R}^d$ produce spectral features $\mathbf{f}_{v,i}^h = m_{v,i}^h \mathcal{E}_f(\mathbf{s}_{v,i}^h + \mathrm{MLP}_f(\mathbf{s}_{v,i}^h)) \in \mathbb{R}^d$.

Global spectrum aggregation. To summarize periodic patterns across the historical window, we aggregate the point-level coefficients into a variable-level spectrum. This supplies global information for comparing variables and provides the historical spectrum used for future extrapolation. With $n_v = \sum_{i=1}^{L_h} m_{v,i}^h$, we compute $\bar{\mathbf{s}}_v = \frac{1}{\max(n_v,1)} \sum_{i=1}^{L_h} \mathbf{s}_{v,i}^h = [\bar{\mathbf{R}}_v^{\mathrm{re}}; \bar{\mathbf{I}}_v^{\mathrm{im}}] \in \mathbb{R}^{2K_f}$, and obtain the global frequency representation $\mathbf{f}_v^g = \mathcal{E}_f(\bar{\mathbf{s}}_v + \mathrm{MLP}_f(\bar{\mathbf{s}}_v))$.

Future frequency representation extrapolation. Future queries have known timestamps but no observed values, so their frequency features cannot be computed directly using Eq. 10. Assuming that the historical spectral components remain relevant over the forecast horizon, we construct future frequency representations by rotating the aggregated historical spectrum according to the query timestamps. For future position $j \in [L_p]$, define $\psi_{v,j,k} = 2\pi w_k \widehat{t}_{v,j}^p$ and rotate the historical coefficients as:

$$\begin{bmatrix} R_{v,j,k}^{p,\mathrm{re}} \\ I_{v,j,k}^{p,\mathrm{im}} \end{bmatrix} = \begin{bmatrix} \cos\psi_{v,j,k} & -\sin\psi_{v,j,k} \\ \sin\psi_{v,j,k} & \cos\psi_{v,j,k} \end{bmatrix} \begin{bmatrix} \bar{R}_{v,k}^{\mathrm{re}} \\ \bar{I}_{v,k}^{\mathrm{im}} \end{bmatrix}, \qquad k \in [K_f]. \tag{11}$$

Concatenating the rotated coefficients gives $\mathbf{s}_{v,j}^p \in \mathbb{R}^{2K_f}$. The same residual encoder produces $\mathbf{f}_{v,j}^p = \mathcal{E}_f(\mathbf{s}_{v,j}^p + \mathrm{MLP}_f(\mathbf{s}_{v,j}^p)) \in \mathbb{R}^d$. If variable v has no valid history, we set $\mathbf{f}_v^g = \mathbf{0}$ and $\mathbf{f}_{v,j}^p = \mathbf{0}$, so that future queries do not introduce unsupported spectral evidence.

3.4 Frequency-Guided Cross-Variable Interaction

IMTS variables are observed asynchronously, and their dependencies may span multiple observations within local intervals. To support generalizable cross-variable interaction, we group observations into patches, select temporally nearby patches from other variables, and weight their contributions by frequency similarity.

Patch construction and pooling. To summarize local patterns for cross-variable interaction, we pack valid historical and future tokens separately, in timestamp order, into patches containing at most B points. Let $\mathcal{P}_a$ denote patch a and $v(a)$ its variable identity. Within each patch, a learned query applies attention pooling over valid temporal tokens, followed by a residual feed-forward layer, to obtain $\mathbf{p}_a \in \mathbb{R}^d$. The same pooling weights aggregate point-level frequency representations into $\mathbf{f}_a \in \mathbb{R}^d$, which describes the patch's local frequency patterns.

Asynchronous temporal neighbor selection. To identify related patches across asynchronously observed variables, we compute distances between patch time intervals. Let $[s_a, e_a]$ denote the valid time interval of patch a. For each patch a, we compute $\delta_{ab} = \max\{(s_a - e_b)_+, (s_b - e_a)_+\}$ and select the top-K_a nearest patches from other variables: $\mathcal{N}_{K_a}(a) = \mathrm{TopK}_{b:\, v(b)\neq v(a)}(-\delta_{ab})$, where $(x)_+ = \max(x, 0)$. This distance is zero for overlapping intervals and equals the gap between disjoint intervals. The number of neighbors K_a is bounded by $K_{\max}$ and the available cross-variable candidates.

Frequency-aware routing and gated message injection. Being close in time does not imply similar value patterns, so giving all selected patches equal weight may introduce unrelated information. We therefore use the frequency representations obtained through learnable spectrum projection and aggregation to determine their contributions. Patch-level representations $\mathbf{f}_a$ describe local patterns, while variable-level representations $\mathbf{f}^g_{v(a)}$ summarize broader historical patterns. Their similarities jointly determine the aggregation weights:

$$
\begin{aligned}
r_{ab} &= \frac{\rho \cos(\mathbf{f}^g_{v(a)}, \mathbf{f}^g_{v(b)}) + (1-\rho)\cos(\mathbf{f}_a, \mathbf{f}_b)}{\tau_f}, \qquad w_{ab} = \mathrm{softmax}_{b\in\mathcal{N}_{K_a}(a)}(r_{ab}), \\
\mathbf{m}_a &= \tanh(\zeta)\mathbf{W}^o_m \sum_{b\in\mathcal{N}_{K_a}(a)} w_{ab}\mathbf{W}^v_m \mathbf{p}_b.
\end{aligned} \tag{12}
$$

Here, $\rho = \mathrm{sigmoid}(\bar{\rho}) \in (0, 1)$ balances local and global similarities, $\tau_f > 0$ is a learned temperature, and ζ is a learned residual parameter initialized to zero. Thus, the message $\mathbf{m}_a$ emphasizes temporal information from patches with similar local and global frequency patterns.

To control how much of this message each future token receives, we compute $g_i = \mathrm{sigmoid}(\mathbf{W}_g[\mathrm{LN}(\mathbf{h}_i); \mathrm{LN}(\mathbf{m}_a)])$ and update $\mathbf{h}'_i = \mathbf{h}_i + g_i\mathbf{W}_{\mathrm{inj}}\mathbf{m}_a$, where $i \in \mathcal{P}^p_a$ and $g_i \in (0, 1)$.

3.5 Forecasting Objective and Pre-training

For each valid future query (v, j), we fuse its temporal and spectral representations as $\mathbf{o}_{v,j} = \mathrm{MLP}_{\mathrm{pred}}([\mathbf{h}'^p_{v,j}; \mathbf{f}^p_{v,j}])$ and predict $(\widehat{u}^{p,q}_{v,j})_{q\in\mathcal{A}} = \mathbf{W}_{\mathrm{head}}\mathbf{o}_{v,j} + \mathbf{b}_{\mathrm{head}}$, where $\mathcal{A} = \{q_1, \ldots, q_Q\} \subset (0, 1)$ includes 0.5. Using historical statistics μ_v, s_v, define $u^p_{v,j} = \mathrm{asinh}((x^p_{v,j} - \mu_v)/s_v)$ and $e^q_{v,j} = u^p_{v,j} - \widehat{u}^{p,q}_{v,j}$. For valid future targets $\Omega = \{(v, j) : v \in [V],\, j \in [L_p],\, m^p_{v,j} = 1\}$, the masked multi-quantile pinball objective is

$$
\mathcal{L}(\theta) = \frac{1}{|\Omega||\mathcal{A}|} \sum_{(v,j)\in\Omega} \sum_{q\in\mathcal{A}} \max\big\{q\, e^q_{v,j}, (q-1)e^q_{v,j}\big\}. \tag{13}
$$

At inference, the median prediction is mapped back to the original scale as $\widehat{x}^p_{v,j} = \mu_v + s_v \sinh(\widehat{u}^{p,0.5}_{v,j})$.

4 Experiments

4.1 Experiments Setup

Pre-training and Downstream datasets We pretrain on four sources: GIFT-Eval Pretrain, GIFT-Eval (Aksu et al., 2024) training dataset, the Chronos corpus, and our processed irregular corpus. There is no test data leakage. We collect eight open-source datasets spanning seven distinct domains for irregular time series forecasting. More details are showed in Appendix A.1 and A.2.

Baselines To demonstrate the generalization ability of QiYao-I, we compare it with both irregular-input and regular-input *time series foundation models*. To evaluate the few-shot learning ability of QiYao-I, we further compare it with a wide range of *end-to-end irregular time series forecasting models*. More details are showed in Appendix A.3.

Implementation details For pre-training, we implement QiYao-I in PyTorch. The pre-training and zero-shot experiments are conducted on six NVIDIA A800 GPUs with 80GB memory. The few-shot experiments are conducted on four NVIDIA 3090 GPUs with 24GB memory. More details are shown in Appendix A.4 and A.5.

Table 1: Average zero-shot forecasting results on downstream datasets. Lower MSE and MAE values indicate better performance. The best and second-best results are highlighted in purple and blue, respectively.

Dataset	CESNET		APTC		ClusterTrace		Seabirds		Human Activity		FNSPID		GDELT		USHCN		1st Count
Methods	(10^{-3})		(10^{-5})		(10^{-3})		(10^{-6})		(10^{-4})		(10^{-5})		(10^{-4})		(10^{-3})		
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
Regular Time Series Foundation Models																	
Toto-2	22.41	92.64	**386.52**	**4595.26**	**41.58**	**149.01**	313.40	9030.82	42.95	**379.01**	**280.88**	**3027.27**	198.90	**841.81**	1064.11	437.90	**2**
Chronos-2	**22.00**	**90.94**	401.85	4681.75	45.99	157.43	**295.50**	9384.00	**42.43**	**386.46**	311.00	3253.60	198.44	843.06	1123.39	**428.12**	**2**
TimesFM-2.5	22.38	91.79	391.60	4618.95	46.33	159.83	358.00	9740.50	55.64	434.63	382.25	3601.35	204.37	847.71	1507.32	524.45	0
MOIRAI-2	22.45	91.02	400.30	4669.20	53.87	177.82	306.00	**8895.50**	48.06	403.01	381.45	3592.50	203.93	859.96	1202.80	485.03	0
Sundial	22.50	92.06	410.55	4735.05	47.70	163.08	341.00	11192.50	44.57	400.52	421.35	3887.70	204.05	867.22	**968.92**	**429.15**	0
TIME-MOE	22.86	95.63	403.10	4692.85	50.82	166.79	522.00	15041.00	57.61	468.22	385.70	3704.25	**196.51**	867.53	992.84	468.72	**1**
Timer	33.97	137.50	506.45	5357.75	146.54	280.55	55018.50	132425.00	352.65	1125.54	13941.95	23862.85	242.93	1059.99	1336.50	525.53	0
Irregular Time Series Foundation Models																	
MIRA	22.36	92.78	410.21	4700.54	48.86	166.95	469.61	14857.72	917.34	2667.64	447.74	4528.26	417.40	1321.87	1138.82	493.80	0
QiYao-I	**22.01**	**90.24**	**385.30**	**4597.09**	**37.22**	**142.67**	**165.67**	**7411.77**	**41.45**	392.10	**249.80**	**3016.12**	**197.47**	**841.66**	**907.77**	457.05	**11**

Table 2: Average 10% few-shot forecasting results of QiYao-I and full-shot results of end-to-end irregular time-series models. Lower MSE and MAE values indicate better performance. The best and second-best results are highlighted in purple and blue, respectively.

Dataset	CESNET		APTC		ClusterTrace		Seabirds		Human Activity		FNSPID		GDELT		USHCN		1st Count
Methods	(10^{-3})		(10^{-5})		(10^{-3})		(10^{-6})		(10^{-4})		(10^{-5})		(10^{-4})		(10^{-3})		
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
Full-Shot End-to-End Irregular Models																	
TFMixer	23.45	101.45	**385.00**	**4630.00**	**41.40**	**150.15**	400.00	14600.00	**36.50**	**362.00**	**420.00**	**4340.00**	191.50	845.50	**514.70**	**264.70**	**3**
ASTGI	**21.60**	**93.20**	**385.00**	4655.00	41.55	154.90	**300.00**	**12500.00**	39.50	400.50	860.00	7845.00	**188.00**	**840.50**	540.30	353.35	**1**
APN	23.30	101.35	**385.00**	4640.00	43.60	153.60	1050.00	26100.00	**37.00**	375.00	1375.00	8145.00	190.00	855.50	**518.60**	333.35	0
KAFNet	23.55	103.90	**385.00**	**4630.00**	44.95	154.90	1000.00	27800.00	39.00	394.00	**375.00**	4665.00	189.00	860.00	640.55	414.85	**1**
Hi-Patch	24.80	104.05	**385.00**	**4630.00**	43.25	154.80	108900.00	281500.00	39.00	411.50	2295.00	10405.00	192.50	879.50	556.75	386.70	0
HyperIMTS	28.40	128.10	**385.00**	**4630.00**	43.95	158.05	450.00	15050.00	40.00	409.00	545.00	5385.00	189.50	849.50	692.05	486.35	0
GNeuralFlow	23.45	103.75	**385.00**	4650.00	45.95	159.95	14950.00	100750.00	182.00	1052.50	1990.00	10640.00	190.50	860.00	628.80	410.55	0
tPatchGNN	23.45	103.10	590.00	5910.00	45.40	155.70	1200.00	24350.00	37.50	387.50	915.00	7420.00	191.50	**835.50**	665.60	437.40	**1**
GraFITi	23.05	104.05	**385.00**	**4630.00**	44.70	155.30	600.00	20100.00	38.00	387.50	1060.00	7580.00	199.50	919.00	638.80	428.15	0
Warpformer	32.20	118.80	600.00	5970.00	91.55	229.35	142100.00	297150.00	124.00	784.00	3730.00	15865.00	1539.50	3253.00	544.95	390.90	0
CRU	23.55	105.35	**385.00**	**4630.00**	48.25	165.90	6050.00	54500.00	91.00	689.50	1040.00	7935.00	193.00	863.50	645.90	418.75	0
NeuralFlow	23.45	103.45	**385.00**	4650.00	44.70	155.55	1550.00	29950.00	100.00	762.00	685.00	6150.00	190.00	853.00	571.45	372.15	0
mTAN	23.35	103.75	**385.00**	**4630.00**	49.90	168.70	2600.00	31450.00	61.00	569.00	845.00	7225.00	198.50	877.50	828.15	509.10	0
SeFT	50.80	154.20	435.00	5025.00	99.65	244.70	8500.00	86000.00	166.00	1022.00	3670.00	16155.00	1576.50	3067.00	775.35	504.55	0
GRU-D	23.30	103.20	**385.00**	**4630.00**	45.45	155.70	28150.00	119950.00	48.50	492.00	1585.00	9960.00	189.00	851.50	664.40	426.45	0
QiYao-I (10%)	**21.59**	**87.49**	**370.92**	**4507.52**	**36.57**	**142.81**	**171.39**	**7810.57**	39.52	**368.79**	476.13	**3916.12**	**188.08**	848.31	538.82	**253.34**	**10**

4.2 Main Results

4.2.1 Compared to other foundation models

To evaluate the generalization ability of QiYao-I on different irregular multivariate time series forecasting tasks, we compare it with various time series foundation models, including Toto2, Chronos2, TimesFM2.5, MOIRAI2, Sundial, TIME-MOE, and Timer, which are originally designed for regular time series inputs, as well as MIRA, which is originally designed for irregular time series inputs. The experimental results are shown in Table 1. Overall, QiYao-I achieves the best or second-best average results on most datasets, demonstrating cross-dataset generalization ability. In contrast, lower-ranked models show less stable performance across different irregular datasets. These results indicate that QiYao-I exhibits better performance and robustness under different data scales, prediction ranges, and time series distributions. Moreover, the evaluated learned interpolation method for regular-input models can lead to larger forecasting errors than linear interpolation, as shown in Appendix B.3.

4.2.2 COMPARED TO OTHER END TO END IRREGULAR MODELS

To evaluate the adaptation ability of QiYao-I, we compare its 10% few-shot performance with 15 end-to-end irregular time series models trained on the full downstream training sets, including TFMixer, ASTGI, and etc.. As shown in Table 2, QiYao-I achieves 10 best results among the 16 metrics, substantially more than all competing methods. It obtains the best MSE and MAE simultaneously on CESNET, APTC, ClusterTrace, and Seabirds, and also achieves the best MAE on FNSPID and USHCN. Moreover, it ranks second in MAE on Human Activity. These results demonstrate that, despite using only 10% of the downstream training data, QiYao-I can effectively transfer its pretrained irregular temporal and cross-variable representations to heterogeneous target domains, providing stable few-shot performance across different sampling structure, and data scales. We further examine the contribution of pre-training by comparing pre-trained and random initialization under same few-shot settings, with detailed results provided in Appendix B.6.

4.3 MODEL ANALYSIS

Ablation Studies. To assess the proposed components, we conduct ablation studies on FNSPID, Seabirds, and Human Activity. As shown in Table 3, adding the sampling-conditioned temporal manifold bias consistently outperforms standard RoPE and CT-RoPE, demonstrating the benefit of adapting temporal relations to local sampling structures. Removing cross-variable interaction or replacing frequency-guided selection with dense interaction also degrades performance, confirming the importance of selective information exchange across asynchronous variables. More analysis are shown in Appendix B.4 and B.5.

Table 3: Ablation studies of QiYao-I. "Full" denotes dense interaction among all variables, while "Aware" denotes the proposed selective frequency-guided variable interaction. ✓ and ✗ indicate enabled and disabled components, respectively. Bold values indicate the best results.

Variant	Temporal Modeling			Channel Modeling		Forecasting Results					
						FNSPID (10^{-5})		Seabirds (10^{-6})		Human Activity (10^{-4})	
	RoPE	CT-RoPE	Manifold Bias	Full	Aware	MSE	MAE	MSE	MAE	MSE	MAE
1	✓	✗	✗	✗	✓	305.27	3305.51	365.64	11787.29	50.65	449.02
2	✗	✓	✗	✗	✓	259.99	3245.83	576.66	15191.28	42.81	406.64
3	✗	✓	✓	✗	✗	350.97	3651.22	558.11	14269.22	41.48	399.50
4	✗	✓	✓	✓	✗	271.05	3062.03	177.10	8100.20	42.03	399.36
5	✗	✓	✓	✗	✓	**249.80**	**3016.12**	**165.67**	**7411.77**	**41.45**	**392.10**

Visualization of attention. To examine whether the proposed attention distinguishes same timestamp pairs under different local sampling structures, we construct two histories from the same ClusterTrace segment. They have same endpoint timestamps, the same temporal range, and the same number of valid observations, but their have different local sampling structure. As shown in Figure 3, the same endpoint pair, $q = t_L = 544.8$ and $k = t_R = 575.3$, receives clearly different attention in the most sampling-responsive head. This result illustrates that the learned attention relation depends on the sampling structure rather than solely on the two endpoint timestamps.

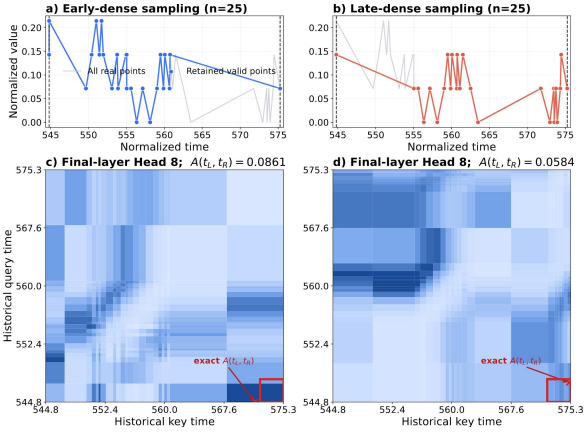

Figure 3: Effect of local sampling structure on temporal attention.

More model analysis are shown in Appendix B.1 and B.2.

5 CONCLUSION

We present QiYao-I, a manifold-based foundation model for irregular multivariate time series forecasting. It learns transferable representations through sampling-conditioned temporal manifold attention and frequency-guided patch interaction, capturing irregular temporal dependencies and asynchronous cross-variable relations. Extensive experiments demonstrate strong zero-shot generalization and few-shot adaptation. Our code is available at `https://anonymous.4open.science/r/QiYao-I-0053/`.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing. All AI-assisted text were carefully reviewed and verified by the authors. The research questions, methodology, experimental design, analysis, and scientific claims were developed independently by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on methodological research for irregular multivariate time series forecasting. Our experiments use existing pretraining datasets and downstream datasets, without involving human subjects or the intentional collection of personally identifiable information. The proposed method is intended for general-purpose forecasting research, and we are not aware of any direct ethical risks arising from the methodology itself.

REPRODUCIBILITY STATEMENT

The model architecture of QiYao-I are described in Section 3, with experimental settings, dataset details, and implementation provided in Appendix A. We report the configurations and evaluation protocols required to reproduce the main experiments. The source code is available at `https://anonymous.4open.science/r/QiYao-I-0053/` and necessary scripts will be released.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74, 2017.

Marin Bilos, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. In *NeurIPS*, pp. 21325–21337, 2021.

Ella Browning, Mark Bolton, Ellie Owen, Akiko Shoji, Tim Guilford, and Robin Freeman. Predicting animal behaviour using deep learning: Gps data alone accurately predict diving in seabirds. *Methods in Ecology and Evolution*, 9(3):681–692, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David A. Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *CoRR*, abs/1606.01865, 2016.

Yuxuan Chen, Shanshan Huang, Yunyao Cheng, Peng Chen, Zhongwen Rao, Yang Shu, Bin Yang, Lujia Pan, and Chenjuan Guo. Aimts: Augmented series and image contrastive learning for time series classification. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 1952–1965. IEEE, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Vinícius M. A. de Souza. Asphalt pavement classification using smartphone accelerometer and complexity invariant distance. *Eng. Appl. Artif. Intell.*, 74:198–211, 2018.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona,*

Spain, August 25-29, 2024, pp. 4918–4927. ACM, 2024. doi: 10.1145/3637528.3671629. URL https://doi.org/10.1145/3637528.3671629.

Shikai Fang, Qingsong Wen, Yingtao Luo, Shandian Zhe, and Liang Sun. Bayotide: Bayesian online multivariate time series imputation with functional decomposition. *arXiv preprint arXiv:2308.14906*, 2023.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten M. Borgwardt. Set functions for time series. In *ICML*, volume 119, pp. 4353–4363, 2020.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, et al. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33:6696–6707, 2020.

Josef Koumar, Karel Hynek, Tomáš Čejka, and Pavel Šiška. Cesnet-timeseries24: Time series dataset for network traffic anomaly detection and forecasting. *Scientific Data*, 12(1):338, 2025.

Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pp. 1–49, 2013.

Boyuan Li, Yicheng Luo, Zhen Liu, Junhao Zheng, Jianming Lv, and Qianli Ma. Hyperimts: Hypergraph neural network for irregular multivariate time series forecasting. *arXiv preprint arXiv:2505.17431*, 2025.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, et al. Mira: Medical time series foundation model for real-world health data. *Advances in Neural Information Processing Systems*, 38:98657–98685, 2026.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024a.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Chenjuan Guo, Bin Yang, and Jilin Hu. ASTGI: adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. *CoRR*, abs/2509.23313, 2025b. doi: 10.48550/ARXIV.2509.23313. URL https://doi.org/10.48550/arXiv.2509.23313.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. *CoRR*, abs/2505.11250, 2025c.

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

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025d.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-patch: Hierarchical patch GNN for irregular multivariate time series. In *ICML*, 2025.

Matthew J Menne, Claude N Williams Jr, and Russell S Vose. Long-term daily and monthly climate records from stations across the contiguous united states (us historical climatology network). Technical report, Environmental System Science Data Infrastructure for a Virtual Ecosystem; CDIAC, 2015.

Giangiacomo Mercatali, Andre Freitas, and Jie Chen. Graph neural flows for unveiling systemic interactions among irregularly sampled time series. *Advances in Neural Information Processing Systems*, 37:57183–57206, 2024.

Xiangfei Qiu, Kangjia Yan, Xvyuan Liu, Xingjian Wu, and Jilin Hu. Bridging time and frequency: A joint modeling framework for irregular multivariate time series forecasting. *arXiv preprint arXiv:2602.00582*, 2026.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, et al. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *ICML*, volume 162, pp. 19388–19405, 2022.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International conference on learning representations*, volume 2025, pp. 34635–34667, 2025.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *ICLR*, 2021.

Francesco Spinnato and Cristiano Landi. Pyrregular: A unified framework for irregular time series, with classification benchmarks. *arXiv preprint arXiv:2505.06047*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Duraisamy Sundararajan. *The discrete Fourier transform: theory, algorithms and applications*. World Scientific, 2001.

Yihang Wang, Yuying Qiu, Peng Chen, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Lightgts: A lightweight general time series forecasting model. *arXiv preprint arXiv:2506.06005*, 2025a.

Yihang Wang, Yuying Qiu, Peng Chen, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series forecasting model with unified representation and adaptive transfer. In *International Conference on Machine Learning.*, 2025b.

Qizhen Weng, Wencong Xiao, Yinghao Yu, Wei Wang, Cheng Wang, Jian He, Yong Li, Liping Zhang, Wei Lin, and Yu Ding. Mlaas in the wild: Workload analysis and scheduling in large-scale heterogeneous GPU clusters. In *NSDI*, pp. 945–960, 2022.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. Grafiti: Graphs for forecasting irregularly sampled time series. In *AAAI*, pp. 16255–16263, 2024.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *SIGKDD*, pp. 3273–3285, 2023.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *ICML*, 2024.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. *arXiv preprint arXiv:2110.05357*, 2021.

Ziyu Zhou, Yiming Huang, Yanyun Wang, Yuankai Wu, James Kwok, and Yuxuan Liang. Revitalizing canonical pre-alignment for irregular multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29115–29123, 2026.

A Experimental Details

A.1 Pre-training Corpus and Sampling

Pretraining data sources. We pre-train QiYao-I on four sources: GIFT-Eval Pretrain, the training portion of GIFT-Eval (Aksu et al., 2024), the Chronos corpus (Ansari et al., 2025), and the our collected irregular corpus from PyRegular (Spinnato & Landi, 2025). The sampling probabilities of these sources are 0.30, 0.25, 0.20, 0.25, respectively. A streaming reader samples source records and windows without loading the entire corpus into GPU memory. Samples from different datasets appear in the same batch, but variables are never assembled across unrelated records. For a native multivariate record, we select at most twenty of its original variables; records containing fewer variables retain their original dimensionality. A univariate sample follows the same temporal and spectral path but has no cross-variable message.

Historical sampling augmentation. We retain original windows alongside four ways of modifying their historical sampling pattern for MTS. Table 4 reports the mixture used in pre-training. Point and block deletion update observation masks without changing the future targets. Sparse subsampling retains a subset of historical observations. Timestamp resampling draws new times inside the historical window and linearly interpolates values only between adjacent, originally valid historical observations. It neither interpolates across an originally missing value nor uses future targets. Each variable in a native multivariate record is augmented independently, producing asynchronous views while preserving the original variable grouping.

Table 4: Historical sampling modes and their probabilities in the pre-training stream. Original windows retain their native sampling.

Mode	Probability	Historical-view construction
Original	0.40	Preserve the original observations
Point deletion	0.20	Independently remove valid points
Block deletion	0.20	Remove one to three contiguous intervals
Timestamp resampling	0.05	Resample times within valid historical intervals
Sparse subsampling	0.15	Retain a strided subset of valid points

A.2 Downstream Benchmarks

We evaluate QiYao-I on eight open-source datasets from distinct domains: USHCN (Menne et al., 2015) from the climate domain, HumanActivity for activity, CESNET (Koumar et al., 2025) for cybersecurity, APTC (de Souza, 2018) and ClusterTrace (**?**) for engineering, FNSPID (Dong et al., 2024) for finance, GDELT (Leetaru & Schrodt, 2013) for social events, and Seabirds (Browning et al., 2018) for nature. Dataset details are as followed. For these datasets, we select useful instances and variables, where we ensure that the time span and number of time points of all used samples are as similar as possible. Then we process them into a unified standard format. We summarize the statistics of these datasets in Table 5. The collection is comprehensive, covering a wide range of domains, sampling patterns and missing rate. The "Samples" reported in Table 5 denotes the number of original continuous records following existing irregular works, rather than the number of forecasting instances used for training and evaluation. For datasets with sufficient records, we split the data at the record level. For datasets with only a few but sufficiently long records, such as CESNET, we perform a chronological split along the time axis and construct rolling forecasting windows within each split.

Table 5: The statistics of evaluation datasets.

Dataset	Domain	# Variables	# Samples	Avg # obs.	Max Length	Missing Rate	Missing Pattern Complexity	Sampling Irregularity	Skewness
HumanActivity	Activity	12	25	6593	304861.22	0.750	0.165	0.396	0.116
USHCN	Climate	5	1114	314	48.00	0.780	0.567	0.999	2.001
CESNET	Cybersecurity	5	11	1700	1999.87	0.000	0.000	0.637	1.497
ClusterTrace	Engineering	3	3	5005	1525.58	0.057	0.119	4.111	0.795
APTC	Engineering	3	2109	399	1232.00	0.001	0.001	0.000	-0.060
FNSPID	Finance	6	10	3495	5303.00	0.001	0.001	0.606	1.100
GDELT	Social	4	6	1684	1460.69	0.042	0.188	1.279	3.140
Seabirds	Nature	3	108	2442	10078.00	0.001	0.001	0.283	0.691

Following existing settings, USHCN and APTC are randomly split into training, validation, and test sets with a ratio of 60%/20%/20%, while Clustertrace, HumanActivity, CESNET, FNSPID, Seabirds, and GDELT use chronological splits within each instance before aggregating the corresponding segments. We use dataset-specific lookback and prediction lengths, which correspond to the real-time ranges of historical observations and future queries in each IMTS dataset, as shown in 6. The test loader follows the original sample-wise or time-wise split shown in the table. We apply the same split and valid-target mask to every compared model. The dataset-level normalization statistics are computed from the training and validation portions, excluding the test portion. USHCN uses z-score normalization; the other datasets use masked min–max normalization. We evaluate all eligible test windows and exclude only invalid future targets from metric calculations.

Table 6: Downstream evaluation configurations. L_h is the historical length and L_p lists the two prediction lengths reported in the main tables.

Dataset	L_h	L_p	Split	Normalization
CESNET	672	168, 336	Time	Min–max
APTC	400	50, 100	Sample	Min–max
ClusterTrace	168	24, 48	Time	Min–max
Seabirds	1440	360, 720	Time	Min–max
Human Activity	6000	200, 3000	Time	Min–max
FNSPID	730	180, 365	Time	Min–max
GDELT	365	30, 90	Time	Min–max
USHCN	24	1, 12	Sample	Z-score

The following is dataset details:

Human Activity. This dataset consists of 12 irregularly measured 3D positional variables from sensors worn on the ankles, belts, and chests of five individuals performing various activities. We chunk the original time series into 25 IMTS. The first 6000 milliseconds are used as observed data to predict the sensor positions for the next 200/3000 milliseconds.

USHCN (Menne et al., 2015). The USHCN dataset includes over 150 years of climate data from multiple U.S. stations, covering 5 climate variables. We focus on data from 1114 stations between 1996 and 2000, resulting in 1114 IMTS. Each instance uses data from the previous 24 months to predict the next 1/12 month's climate conditions.

CESNET (Koumar et al., 2025). This dataset comprises network traffic data collected from the CESNET academic network, representing a large-scale collection of flow-level statistics. It contains 11 IMTS derived from millions of network flows, where each instance includes 5 features such as packet counts, byte counts, and inter-arrival times recorded over varying observation windows. Following standard protocol for network traffic analysis, the flows are processed into discrete intervals. We use the first 672 hours as the observed sequence to predict the traffic characteristics and flow types in the subsequent 168/336 hours.

ClusterTrace (Weng et al., 2022). The ClusterTrace dataset provides a detailed trace of a production cluster containing multiple machines over a period of time. It captures complex workloads and resource utilization metrics, including CPU, memory, and disk usage for various tasks and containers. Due to the dynamic scheduling and heterogeneous nature of cloud workloads, the resource metrics are recorded as IMTS across thousands of unique job instances. Following common practices in cluster management research, the data is processed to focus on temporal resource demands. We select 3 IMTS and 3 variables for irregular time series forecasting. We use the resource usage data from the first 168 hours as the observed sequence to predict the workload intensity and resource requirements for the subsequent 24/48 hours.

APTC (de Souza, 2018). This dataset is a large-scale collection of traffic speed data obtained from multiple dual-loop detectors on an urban expressway. After processed, it contains over 2109 IMTS instances, with each instance capturing 3 traffic-related variables. Due to detector malfunctions and data transmission issues, the dataset exhibits significant irregularity and missing values. We use the traffic observations from the previous 400 milliseconds as the observed sequence to predict the traffic conditions and flow patterns in the subsequent 50/100 milliseconds.

FNSPID (Dong et al., 2024). The FNSPID dataset is a comprehensive Financial News and Stock Price Integration Dataset, combining stock price movements with corresponding financial news headlines. It covers thousands of stocks from major exchanges, featuring irregularly timed news events alongside stock price variables such as open, close, and volume. The alignment of sparse news signals with continuous price data results in multiple multi-modal IMTS. We select 10 IMTS and 6 variables for the irregular time series forecasting. To facilitate stock trend and volatility forecasting, each instance is organized into windows of sequential market activities. We use the integrated data from the previous 730 days as observed data to predict the stock price fluctuations and market sentiment for the following 180/365 days.

GDELT (Leetaru & Schrodt, 2013). This dataset is sourced from the Global Database of Events, Language, and Tone, which monitors world broadcast, print, and web news in multiple languages. It captures physical activities and socio-political events globally, resulting in a massive collection of IMTS with varying temporal resolutions based on event occurrences. The dataset includes multiple variables such as the Goldstein scale, average tone, and event counts across thousands of distinct themes and locations. To adapt it for the irregular time series forecasting, we select 6 IMTS instances and 4 variables. We use the event dynamics from the previous 365 days as the observed data to predict the event intensity and regional stability in the subsequent 30/90 days.

Seabirds (Browning et al., 2018). This dataset tracks the movement and behavioral patterns of seabirds using high-resolution GPS and tri-axial accelerometer sensors. It contains IMTS representing multiple distinct movement variables, such as altitude, speed, and body orientation, collected from multiple species during foraging trips. Given the irregular nature of wildlife tracking due to varying signal quality and nesting cycles, the data is segmented into multiple IMTS instances, each spanning a specific duration of flight or diving activity. We select 108 IMTS and 3 variables for the irregular time series forecasting. We use the sensor measurements from the first 1440 minutes as the observed data to predict the bird's trajectory and behavioral states in the subsequent 360/720 minutes.

A.3 Baselines and Input Adaptation

Foundation models. The zero-shot comparison includes Toto-2 (Khwaja et al., 2026), Chronos-2 (Ansari et al., 2025), TimesFM-2.5 (Das et al., 2023), Moirai-2 (Liu et al., 2025a), Sundial (Liu et al., 2025d), Time-MoE (Shi et al., 2025), and Timer (Liu et al., 2024b), which accept regularly sampled inputs, as well as MIRA (Li et al., 2026), which supports irregular timestamps. For regular-input models, irregular historical observations are converted to a regular input grid by interpolation, as shown in 1. QiYao-I and MIRA receive irregular timestamps and observation masks directly. We use each model's median forecast as its point prediction when probabilistic outputs are available. Every model is scored against the same original valid future targets.

End-to-end irregular models. The few-shot comparison includes TFMixer (Qiu et al., 2026), ASTGI (Liu et al., 2025b), APN (Liu et al., 2025c), KAFNet (Zhou et al., 2026), Hi-Patch (Luo et al., 2025), HyperIMTS (Li et al., 2025), GNeuralFlow (Mercatali et al., 2024), tPatchGNN Zhang et al. (2024), GraFITi Yalavarthi et al. (2024), Warpformer Zhang et al. (2023), CRU Schirmer et al. (2022), NeuralFlow Bilos et al. (2021), mTAN Shukla & Marlin (2021), SeFT Horn et al. (2020), and GRU-D Che et al. (2016). As indicated in the main table, these baselines use their full training splits, whereas QiYao-I is adapted using only 10% of the corresponding training data. This difference in training-data access is retained explicitly in the table caption and should be considered when interpreting rankings.

A.4 Model and Optimization Settings

Table 7 lists the configuration used by the reported QiYao-I checkpoint. The sampling-measure kernel width κ controls how each anchor counts observations; it is distinct from the learned effective bandwidth of the manifold attention bias. During pre-training, the number of time-nearest cross-variable patches is sampled from 1 to 16 when interaction is active. History length is set 576 and prediction length is set 192.

Algorithm 1 Linear Resampling for Irregular Historical Observations

Require: Irregular time points T, observations X, mask M, number of regular samples L
Ensure: Regularly sampled sequence X^{reg}

for each sample i **do**
  Find the valid historical time range according to M_i
  Generate a regular grid T_i^{reg} within this range
  for each variable d **do**
    Select valid observations according to $M_{i,d}$
    if there is no valid observation **then**
      Set $X_{i,d}^{\mathrm{reg}} \leftarrow 0$
    else if there is only one valid observation **then**
      Set $X_{i,d}^{\mathrm{reg}}$ as a constant sequence
    else
      Sort observations by time
      Average values with duplicated timestamps
      Linearly interpolate observations onto T_i^{reg}
    end if
  end for
end for
return X^{reg}

Table 7: QiYao-I architecture used in the main experiments.

Configuration	Value	Configuration	Value
Transformer layers	3	Hidden dimension	256
Attention heads	8	Feed-forward dimension	1024
Time embedding dimension	32	Manifold dimension / head	16
Sampling anchors A	9	Sampling window W	32
Descriptor kernel width κ	0.2	Time reference R	576
Learnable frequencies	16	Patch size	8
Maximum native variables	20	Forecast quantiles	21
Maximum neighbors K	16	Dropout	0.1

Pre-training stages. QiYao-I is first trained on univariate windows to learn stable sampling-conditioned temporal manifold attention. The cross-variable module remains in the architecture but has no eligible neighbors for these samples. Then, the multivariate stage starts from the best univariate checkpoint, initializes a new AdamW optimizer. This stage uses native multivariate records and also include univariate samples for joint training. Both stages use FP32. The univariate stage uses six NVIDIA A800 GPUs, and the multivariate stage also uses six NVIDIA A800 GPUs. Both use AdamW with a peak learning rate of 6×10^{-5}, 1000 warm-up steps, cosine decay to 6×10^{-6}, weight decay 0.01, and gradient clipping at norm 1.0. The training seed is 260914.

Ten-percent few-shot adaptation. For each downstream dataset and prediction length, we start from the same pre-trained checkpoint and train on a fixed 10% subset of its training split. USHCN retains z-score normalization during both adaptation and evaluation.

A.5 Evaluation Metrics and Aggregation

All metrics use the median forecast after the model's inverse value transformation. For dataset variable v, let $\mathcal{Q}_v$ be its valid future targets across test samples, and let $n_v = |\mathcal{Q}_v|$. We first compute variable-level errors

$$E_v^{\mathrm{MSE}} = \frac{1}{n_v} \sum_{j \in \mathcal{Q}_v} (y_{v,j} - \widehat{y}_{v,j})^2, \qquad E_v^{\mathrm{MAE}} = \frac{1}{n_v} \sum_{j \in \mathcal{Q}_v} |y_{v,j} - \widehat{y}_{v,j}|. \tag{14}$$

Only variables with $n_v > 0$ enter the reported macro average: $\text{Macro-MSE} = |\mathcal{V}|^{-1} \sum_{v \in \mathcal{V}} E_v^{\mathrm{MSE}}$ and $\text{Macro-MAE} = |\mathcal{V}|^{-1} \sum_{v \in \mathcal{V}} E_v^{\mathrm{MAE}}$, where $\mathcal{V} = \{v : n_v > 0\}$. For each dataset, the Avg

row is the arithmetic mean of its two reported prediction lengths, calculated separately for MSE and MAE.

B ADDITIONAL MODEL ANALYSIS

B.1 EFFICIENCY ANALYSIS

We evaluate inference efficiency on ClusterTrace using 37 test windows in 10 batches of up to four. Figure 4 compares inference time, peak allocated GPU memory (bubble area), and forecasting error. QiYao-I achieves the lowest macro MSE, with 8.07 s of recorded inference time and 4.08 GiB of peak GPU allocation. Toto-2 is faster and uses less memory, but its MSE is higher; all other profiled baselines also yield larger errors.

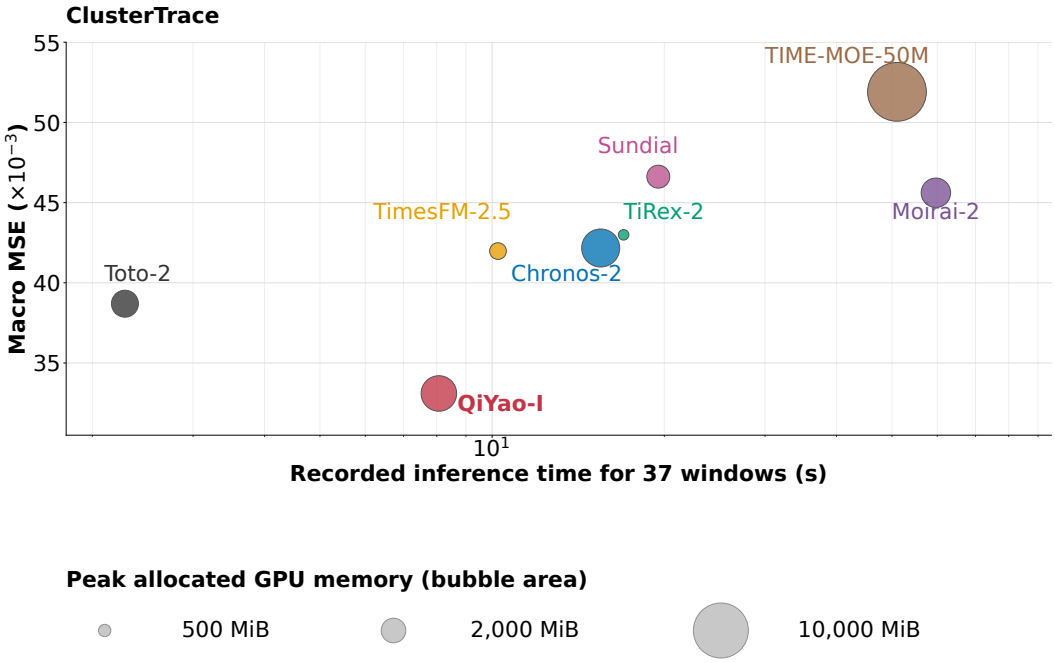

Figure 4: Efficiency analysis.

B.2 ANALYSIS OF TOPK SELECTION

Figure 5 examines the number of cross-variable patches selected for interaction. We vary K at inference while keeping the pre-trained model fixed, and report the average macro MSE and MAE over two prediction lengths for each dataset. On Human Activity, both errors decrease as K increases from 1 to 16, suggesting that additional neighboring patches provide useful information. On USHCN, both metrics improve up to $K = 8$ but rise slightly at $K = 16$. Thus, a larger neighbor set is not always beneficial: its optimal size depends on how much relevant information the selected patches provide. These results support using selective rather than unrestricted cross-variable interaction.

B.3 ANALYSIS OF INTERPOLATION METHOD

Following prior irregular time series work, we use linear interpolation (Li et al., 2026) as a simple and effective way to convert irregular observations into regular inputs for regular-input models. We construct a regular grid over the original time range and linearly estimate each variable from neighboring observations. This is used for all regular-input baselines and does not require any extra training. In contrast, learned imputation would introduce another trained modules and extra hyperparameters, making the comparison less direct. Nevertheless, we additionally evaluated Chronos-2, Toto2, and MOIRAI2 with irregular inputs first interpolated by BayOTIDE (Fang et al., 2023), a SOTA interpolation method. The details are shown in the Table 8.QiYao-I still performs well.

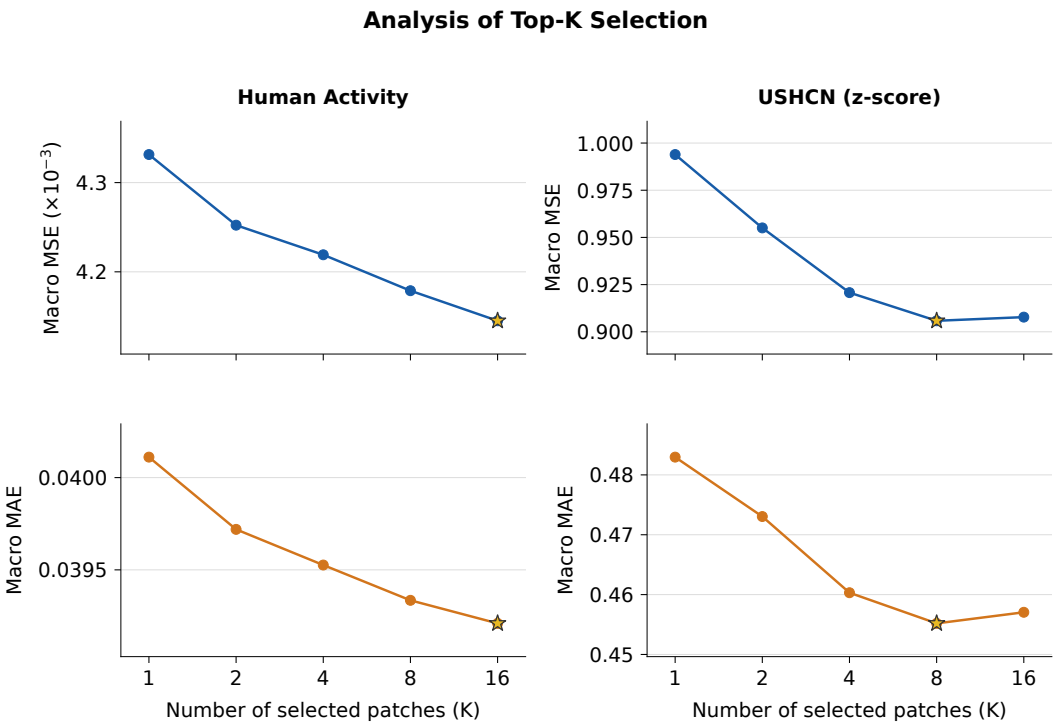

Figure 5: TopK analysis.

Table 8: Zero-shot forecasting results with linear and BayOTIDE interpolation. Results are averaged over two prediction lengths: CESNET (168, 336), FNSPID (180, 365), and Human Activity (200, 3000). All baseline point forecasts use the median ($q = 0.5$). QiYao-I directly models irregular observations. Lower values are better; bold indicates the best result.

Method	**CESNET** (10^{-3})		**FNSPID** (10^{-5})		**Human Activity** (10^{-4})	
	MSE	MAE	MSE	MAE	MSE	MAE
Linear interpolation						
Chronos-2	**22.00**	90.94	311.00	3253.60	42.43	386.46
Toto-2	22.41	92.64	280.88	3027.27	42.95	**379.01**
Moirai-2	22.45	91.02	381.45	3592.50	48.06	403.01
BayOTIDE interpolation						
Chronos-2	22.67	91.96	1377.25	8382.05	240.41	1242.97
Toto-2	28.12	99.81	2877.98	12947.19	203.70	1162.71
Moirai-2	22.28	91.84	783.18	6950.05	114.40	874.15
Direct irregular input						
QiYao-I	22.01	**90.24**	**249.80**	**3016.12**	**41.45**	392.10

B.4 Sampling-Conditioned Bandwidth

Experimental setup. We evaluate ClusterTrace with prediction lengths 24 and 48, and USHCN with prediction lengths 1 and 12. We compare three settings: *Full*, which retains the learned sampling-conditioned bias; *Fixed*, which sets $r_u^{\ell,r} = 1$ while preserving the learned manifold distances and the parameters $\beta_{\ell,r}$ and $\eta_{\ell,r}$; and *No bias*, which sets $b_{uw}^{\ell,r} = 0$. All remaining model parameters and operations are unchanged.

Table 9: Bandwidth interventions. Original denotes the complete test set. AVG gives the equally weighted average over the two prediction lengths. Lower values are better.

Dataset	Pred.	Input	Full		Fixed ($r = 1$)		No bias ($b = 0$)	
			MSE	MAE	MSE	MAE	MSE	MAE
ClusterTrace	24	Original	0.031973	0.132937	0.033591	0.136933	0.040467	0.151741
	48	Original	0.042463	0.152406	0.046573	0.159275	0.044720	0.159278
	AVG	Original	0.037218	0.142671	0.040082	0.148104	0.042594	0.155509
USHCN	1	Original	0.928796	0.459120	0.940851	0.463556	0.949627	0.465101
	12	Original	0.886734	0.454973	0.888887	0.455911	0.895354	0.468892
	AVG	Original	0.907765	0.457046	0.914861	0.459734	0.922491	0.466997

Table 10: Endpoint responses to different local sampling structures. The paired inputs have same endpoints and point counts. ΔB is the mean absolute relative bandwidth change; Δb is the mean absolute change in the adaptive bias. The last column reports the maximum absolute change in the fixed-bandwidth bias.

Dataset	Pred.	Windows	Variable-window pairs	ΔB (%)	Δb	Fixed-bias change
ClusterTrace	24	37	111	5.51	2382.79	0
ClusterTrace	48	19	57	5.39	2456.90	0
USHCN	1	128	640	1.96	1305.74	0
USHCN	12	128	640	1.98	1315.91	0

Results. As shown in Table 9, the full model achieves lower MSE and MAE than both intervention settings across the evaluated prediction lengths on ClusterTrace and USHCN. Compared with fixed bandwidths, sampling-conditioned bandwidths provide larger improvements on ClusterTrace and smaller but consistent improvements on USHCN. Removing the manifold bias also increases forecasting errors on both datasets. Under this frozen-checkpoint evaluation, these results suggest that the learned temporal bias contributes to forecasting and that adapting its bandwidth to local sampling structures provides additional benefits.

Response to local sampling structure. We additionally examine the shared left-to-right endpoint pair in every eligible variable and selected window, covering all three temporal layers and eight attention heads. For effective bandwidth $B_{uw}^{\ell,r} = \beta_{\ell,r}\sqrt{r_u^{\ell,r} r_w^{\ell,r}}$, we measure the absolute relative change $100|B_{uw}^{\ell,r,\mathrm{late}}/B_{uw}^{\ell,r,\mathrm{early}} - 1|$. We also measure the absolute change in the endpoint bias. Table 10 averages these quantities over all eligible variable-window and layer-head combinations.

The fixed-bandwidth endpoint bias remains unchanged, whereas the adaptive bandwidth and bias respond to the changed sampling structure in all 24 layer-head combinations. This confirms that the mechanism distinguishes sampling patterns beyond the endpoint interval and observation count.

B.5 Frequency-Based Routing with Fixed Neighbors

Experimental setup. To isolate the contribution of frequency-based aggregation weights, we evaluate Human Activity with prediction lengths 200 and 3000, and FNSPID with prediction lengths 180 and 365. For each input batch, we verify that all compared settings use same temporal Top-K neighbor indices and validity masks. We compare *Local + Global*, *Uniform*, *Local only*, and *Global only* routing.

Results. As shown in Table 11, combined frequency routing reduces average MSE relative to uniform aggregation by 0.39% on Human Activity and 1.09% on FNSPID. The corresponding MAE is nearly unchanged on Human Activity and decreases by 0.28% on FNSPID. However, local-only routing yields slightly better average results on Human Activity, while global-only routing performs

Table 11: Frozen-checkpoint routing interventions with identical temporal neighbors. AVG denotes the equally weighted average over the two prediction lengths. All values are unscaled.

Dataset	Pred.	Local + Global		Uniform		Local only		Global only	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Human Activity	200	0.00291967	0.03096386	0.00293623	0.03101229	0.00291467	0.03093724	0.00293426	0.03104303
	3000	0.00537056	0.04745604	0.00538634	0.04739915	0.00537137	0.04746498	0.00536983	0.04744481
	AVG	0.00414511	0.03920995	0.00416129	0.03920572	0.00414302	0.03920111	0.00415204	0.03924392
FNSPID	180	0.00148230	0.02358637	0.00150630	0.02361973	0.00148254	0.02358576	0.00148155	0.02358435
	365	0.00351367	0.03673599	0.00354458	0.03687244	0.00351467	0.03674194	0.00351207	0.03672738
	AVG	0.00249798	0.03016118	0.00252544	0.03024608	0.00249860	0.03016385	0.00249681	0.03015587

Table 12: Comparison of pre-trained and random initialization using the same fixed 10% training subsets and training settings. Results are reported as mean ± sample standard deviation over three optimization seeds. AVG first averages the two prediction lengths within each repetition and then summarizes the three repetitions. The lower mean in each comparison is shown in bold.

Dataset / Pred.	Initialization	Macro-MSE	Macro-MAE
CESNET / 168	Pre-trained	$\mathbf{0.021285} \pm 0.000063$	$\mathbf{0.087062} \pm 0.000023$
	Random	0.021836 ± 0.000754	0.088425 ± 0.002280
CESNET / 336	Pre-trained	$\mathbf{0.021522} \pm 0.000275$	$\mathbf{0.087548} \pm 0.000318$
	Random	0.022771 ± 0.000503	0.090517 ± 0.000634
CESNET / AVG	Pre-trained	$\mathbf{0.021403} \pm 0.000162$	$\mathbf{0.087305} \pm 0.000164$
	Random	0.022304 ± 0.000353	0.089471 ± 0.001335
ClusterTrace / 24	Pre-trained	$\mathbf{0.032888} \pm 0.000656$	$\mathbf{0.133225} \pm 0.001104$
	Random	0.039602 ± 0.000496	0.149163 ± 0.000718
ClusterTrace / 48	Pre-trained	$\mathbf{0.041636} \pm 0.001027$	$\mathbf{0.153120} \pm 0.000885$
	Random	0.043186 ± 0.001579	0.155469 ± 0.002915
ClusterTrace / AVG	Pre-trained	$\mathbf{0.037262} \pm 0.000633$	$\mathbf{0.143173} \pm 0.000343$
	Random	0.041394 ± 0.000562	0.152316 ± 0.001213
USHCN / 1	Pre-trained	0.564125 ± 0.003276	0.262441 ± 0.002382
	Random	$\mathbf{0.562252} \pm 0.003004$	$\mathbf{0.261325} \pm 0.002290$
USHCN / 12	Pre-trained	$\mathbf{0.517573} \pm 0.000536$	$\mathbf{0.246203} \pm 0.000415$
	Random	0.527935 ± 0.010005	0.246396 ± 0.001192
USHCN / AVG	Pre-trained	$\mathbf{0.540849} \pm 0.001812$	0.254322 ± 0.001002
	Random	0.545094 ± 0.004568	$\mathbf{0.253861} \pm 0.000550$

slightly better on FNSPID. These findings support modest, task-dependent benefits of frequency weighting over equal weights.

B.6 Pre-trained versus Random Initialization

Experimental setup. We compare pre-trained and random initialization of the same complete architecture on CESNET (168/336), ClusterTrace (24/48), and USHCN (1/12). For each prediction length, both initializations use the same fixed 10% training subset across three optimization seeds.

We retain the existing preprocessing protocol, with normalization statistics estimated from the training and validation splits, excluding the test split. Thus, 10% refers to the training windows used for gradient updates. USHCN uses z-score normalization, while the other datasets use min–max normalization.

Results. As shown in Table 12, pre-trained initialization achieves lower mean MSE and MAE at both prediction lengths on CESNET and ClusterTrace. Compared with random initialization, it reduces average MSE by 4.04% and 9.98%, and average MAE by 2.42% and 6.00%, respectively. Results on USHCN are mixed: random initialization performs slightly better at prediction length 1, whereas pre-training yields lower mean errors at length 12, with only a marginal MAE difference. Across its two lengths, pre-training reduces MSE by 0.78% but increases MAE by 0.18%. These results support the benefit of pre-training for adaptation with limited training data, while showing that its gains depend on the dataset and prediction length.

B.7 LIMITATIONS

Although QiYao-I is trained with a multi-quantile objective, our evaluation focuses on median point forecasts using MSE and MAE, following prior work. These metrics do not establish the calibration of its predictive quantiles. Our pre-training corpus is underexplored. A controlled comparison of pre-training on naturally irregular data, synthetically generated data, and their combination would further clarify their respective contributions to cross-dataset generalization.

Table 13: Per-horizon zero-shot forecasting results on CESNET and APTC. MSE and MAE are reported on the same scales as Table 1; lower is better. Best and second-best values are marked in purple and blue, respectively.

Method	**CESNET** (10^{-3})				**APTC** (10^{-5})			
	$P = 168$		$P = 336$		$P = 50$		$P = 100$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Toto-2	22.39	92.69	22.44	92.58	**389.76**	**4615.61**	**383.29**	**4574.92**
Chronos-2	**22.08**	91.23	**21.93**	**90.65**	406.40	4708.10	397.30	4655.40
TimesFM-2.5	22.44	91.90	22.31	91.67	395.50	4640.80	387.70	4597.10
MOIRAI-2	22.18	**90.68**	22.72	91.35	405.70	4701.10	394.90	4637.30
Sundial	22.68	92.00	22.32	92.11	414.00	4753.70	407.10	4716.40
TIME-MOE	22.14	93.40	23.58	97.87	406.80	4714.20	399.40	4671.50
Timer	31.33	130.08	36.61	144.91	434.80	4961.40	578.10	5754.10
MIRA	22.49	93.28	22.24	92.28	405.20	4673.66	415.23	4727.43
QiYao-I	**21.84**	**89.99**	**22.17**	**90.49**	**386.72**	**4608.78**	**383.88**	**4585.40**

Table 14: Per-horizon zero-shot forecasting results on ClusterTrace and Seabirds. MSE and MAE are reported on the same scales as Table 1; lower is better. Best and second-best values are marked in purple and blue, respectively.

Method	**ClusterTrace** (10^{-3})				**Seabirds** (10^{-6})			
	$P = 24$		$P = 48$		$P = 360$		$P = 720$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Toto-2	**38.70**	**141.20**	**44.46**	**156.83**	184.55	**6618.25**	442.25	11443.38
Chronos-2	42.17	149.08	49.81	165.78	**175.00**	6940.00	**416.00**	11828.00
TimesFM-2.5	41.98	149.71	50.68	169.95	233.00	7565.00	483.00	11916.00
MOIRAI-2	45.61	163.02	62.12	192.63	195.00	7061.00	417.00	**10730.00**
Sundial	46.62	159.84	48.79	166.31	249.00	9435.00	433.00	12950.00
TIME-MOE	51.91	167.87	49.72	165.72	416.00	13015.00	628.00	17067.00
Timer	118.33	240.93	174.75	320.17	26911.00	91971.00	83126.00	172879.00
MIRA	45.75	159.39	51.96	174.51	396.57	13308.31	542.64	16407.14
QiYao-I	**31.97**	**132.94**	**42.46**	**152.41**	**124.58**	**6450.75**	**206.77**	**8372.78**

C FULL RESULTS

Table 15: Per-horizon zero-shot forecasting results on Human Activity and FNSPID. MSE and MAE are reported on the same scales as Table 1; lower is better. Best and second-best values are marked in purple and blue, respectively.

Method	**Human Activity** (10^{-4})				**FNSPID** (10^{-5})			
	$P = 200$		$P = 3000$		$P = 180$		$P = 365$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Toto-2	31.82	**301.85**	54.08	**456.17**	**166.14**	**2365.01**	**395.63**	**3689.53**
Chronos-2	31.14	306.45	**53.72**	**466.47**	191.20	2543.40	430.80	3963.80
TimesFM-2.5	39.80	334.90	71.47	534.36	220.90	2826.60	543.60	4376.10
MOIRAI-2	32.94	309.99	63.17	496.02	201.60	2653.10	561.30	4531.90
Sundial	**30.21**	**303.81**	58.92	497.23	257.10	3111.80	585.60	4663.60
TIME-MOE	32.78	331.21	82.44	605.23	236.80	2990.30	534.60	4418.20
Timer	54.90	450.58	650.40	1800.49	7772.30	18312.50	20111.60	29413.20
MIRA	707.47	2222.64	1127.21	3112.64	465.51	4568.12	429.98	4488.40
QiYao-I	**29.20**	309.64	**53.71**	474.56	**148.23**	**2358.64**	**351.37**	**3673.60**

Table 16: Per-horizon zero-shot forecasting results on GDELT and USHCN. MSE and MAE are reported on the same scales as Table 1; lower is better. Best and second-best values are marked in purple and blue, respectively. USHCN uses z-score normalization; its ranking includes only models with verified per-horizon results.

Method	**GDELT** (10^{-4})				**USHCN** (10^{-3})			
	$P = 30$		$P = 90$		$P = 1$		$P = 12$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Toto-2	201.56	**847.14**	196.25	**836.48**	1166.67	432.62	961.55	443.17
Chronos-2	202.21	851.26	**194.66**	**834.86**	1095.07	419.62	1151.71	**436.62**
TimesFM-2.5	205.71	853.00	203.03	842.42	1302.10	461.30	1712.54	587.61
MOIRAI-2	204.85	861.93	203.00	857.98	1130.38	440.32	1275.23	529.75
Sundial	209.39	881.07	198.71	853.37	1042.24	**415.81**	**895.61**	**442.50**
TIME-MOE	**198.94**	869.98	**194.07**	865.08	**859.57**	**403.62**	1126.11	533.82
Timer	232.91	1030.10	252.94	1089.89	1442.00	558.05	1231.00	493.00
MIRA	402.70	1328.72	432.10	1315.03	1052.04	485.50	1225.61	502.11
QiYao-I	**196.71**	**844.94**	198.23	838.37	**928.80**	459.12	**886.73**	454.97

Table 17: Per-horizon forecasting results on CESNET and APTC. QiYao-I uses 10% few-shot training; the end-to-end baselines use full-shot training. MSE and MAE use the scales of the main few-shot table. The best and second-best reported values are highlighted in purple and blue, respectively; ties at the reported precision are highlighted equally.

Method	**CESNET** (10^{-3})				**APTC** (10^{-5})			
	$P = 168$		$P = 336$		$P = 50$		$P = 100$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TFMixer	23.30	100.30	23.60	102.60	**390.00**	**4680.00**	**380.00**	**4580.00**
ASTGI	**21.60**	**92.10**	**21.60**	**94.30**	**390.00**	**4680.00**	**380.00**	4630.00
APN	23.10	101.70	23.50	101.00	**390.00**	4690.00	**380.00**	4590.00
KAFNet	23.60	103.70	23.50	104.10	**390.00**	**4680.00**	**380.00**	**4580.00**
Hi-Patch	24.40	103.70	25.20	104.40	**390.00**	**4680.00**	**380.00**	**4580.00**
HyperIMTS	27.00	122.80	29.80	133.40	**390.00**	**4680.00**	**380.00**	**4580.00**
GNeuralFlow	23.40	103.90	23.50	103.60	**390.00**	4710.00	**380.00**	4590.00
tPatchGNN	23.40	102.70	23.50	103.50	560.00	5740.00	620.00	6080.00
GraFITi	23.10	104.20	23.00	103.90	**390.00**	**4680.00**	**380.00**	**4580.00**
Warpformer	23.40	101.60	41.00	136.00	570.00	5790.00	630.00	6150.00
CRU	23.30	103.80	23.80	106.90	**390.00**	**4680.00**	**380.00**	**4580.00**
NeuralFlow	23.50	104.00	23.40	102.90	**390.00**	4700.00	**380.00**	4600.00
mTAN	23.50	103.80	23.20	103.70	**390.00**	**4680.00**	**380.00**	**4580.00**
SeFT	50.80	154.00	50.80	154.40	430.00	4970.00	440.00	5080.00
GRU-D	23.30	103.10	23.30	103.30	**390.00**	**4680.00**	**380.00**	**4580.00**
QiYao-I (10%)	**21.34**	**87.08**	**21.83**	**87.90**	**369.84**	**4504.04**	**371.99**	**4510.99**

Table 18: Per-horizon forecasting results on ClusterTrace and Seabirds. QiYao-I uses 10% few-shot training; the end-to-end baselines use full-shot training. MSE and MAE use the scales of the main few-shot table. The best and second-best reported values are highlighted in purple and blue, respectively; ties at the reported precision are highlighted equally.

Method	**ClusterTrace** (10^{-3})				**Seabirds** (10^{-6})			
	$P = 24$		$P = 48$		$P = 360$		$P = 720$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TFMixer	**39.20**	**147.30**	43.60	153.00	300.00	13400.00	500.00	15800.00
ASTGI	40.60	157.50	**42.50**	**152.30**	**200.00**	**11700.00**	**400.00**	**13300.00**
APN	42.50	153.10	44.70	154.10	700.00	19400.00	1400.00	32800.00
KAFNet	45.00	155.00	44.90	154.80	900.00	28300.00	1100.00	27300.00
Hi-Patch	43.70	156.20	42.80	153.40	24800.00	132900.00	193000.00	430100.00
HyperIMTS	44.70	157.50	43.20	158.60	500.00	16300.00	**400.00**	13800.00
GNeuralFlow	47.20	163.20	44.70	156.70	13700.00	96000.00	16200.00	105500.00
tPatchGNN	46.20	157.30	44.60	154.10	1000.00	22500.00	1400.00	26200.00
GraFITi	44.80	155.90	44.60	154.70	500.00	18000.00	700.00	22200.00
Warpformer	86.00	220.00	97.10	238.70	32300.00	151900.00	251900.00	442400.00
CRU	48.50	171.70	48.00	160.10	6300.00	55200.00	5800.00	53800.00
NeuralFlow	44.70	155.50	44.70	155.60	1700.00	30200.00	1400.00	29700.00
mTAN	50.60	176.20	49.20	161.20	4500.00	42900.00	700.00	20000.00
SeFT	99.50	244.60	99.80	244.80	8400.00	88500.00	8600.00	83500.00
GRU-D	46.30	157.20	44.60	154.20	8800.00	59600.00	47500.00	180300.00
QiYao-I (10%)	**32.67**	**132.90**	**40.48**	**152.71**	**162.70**	**7674.32**	**180.09**	**7946.82**

Table 19: Per-horizon forecasting results on Human Activity and FNSPID. QiYao-I uses 10% few-shot training; the end-to-end baselines use full-shot training. MSE and MAE use the scales of the main few-shot table. The best and second-best reported values are highlighted in purple and blue, respectively; ties at the reported precision are highlighted equally.

Method	**Human Activity** (10^{-4})				**FNSPID** (10^{-5})			
	$P = 200$		$P = 3000$		$P = 180$		$P = 365$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TFMixer	**25.00**	**281.00**	**48.00**	**443.00**	**200.00**	**3210.00**	640.00	5470.00
ASTGI	28.00	317.00	51.00	484.00	720.00	7240.00	1000.00	8450.00
APN	26.00	297.00	**48.00**	453.00	320.00	4080.00	2430.00	12210.00
KAFNet	27.00	305.00	51.00	483.00	350.00	4430.00	**400.00**	**4900.00**
Hi-Patch	29.00	337.00	**49.00**	486.00	660.00	5490.00	3930.00	15320.00
HyperIMTS	26.00	303.00	54.00	515.00	560.00	5470.00	**530.00**	5300.00
GNeuralFlow	185.00	1064.00	179.00	1041.00	2910.00	13690.00	1070.00	7590.00
tPatchGNN	27.00	303.00	**48.00**	472.00	640.00	5600.00	1190.00	9240.00
GraFITi	26.00	296.00	50.00	479.00	660.00	5990.00	1460.00	9170.00
Warpformer	40.00	436.00	208.00	1132.00	3770.00	16140.00	3690.00	15590.00
CRU	82.00	645.00	100.00	734.00	1010.00	7980.00	1070.00	7890.00
NeuralFlow	98.00	752.00	102.00	772.00	760.00	6520.00	610.00	5780.00
mTAN	52.00	519.00	70.00	619.00	890.00	7130.00	800.00	7320.00
SeFT	164.00	1019.00	168.00	1025.00	3690.00	16480.00	3650.00	15830.00
GRU-D	40.00	434.00	57.00	550.00	1290.00	9480.00	1880.00	10440.00
QiYao-I (10%)	**25.75**	**287.07**	53.28	**450.51**	**258.26**	**2905.11**	694.00	**4927.13**

Table 20: Per-horizon forecasting results on GDELT and USHCN. QiYao-I uses 10% few-shot training; the end-to-end baselines use full-shot training. MSE and MAE use the scales of the main few-shot table. The best and second-best reported values are highlighted in purple and blue, respectively; ties at the reported precision are highlighted equally. USHCN uses z-score normalization.

Method	**GDELT** (10^{-4})				**USHCN** (10^{-3})			
	$P = 30$		$P = 90$		$P = 1$		$P = 12$	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TFMixer	191.00	**843.00**	192.00	848.00	**537.60**	**251.50**	**491.80**	**277.90**
ASTGI	**188.00**	**840.00**	**188.00**	**841.00**	582.80	370.90	**497.80**	335.80
APN	190.00	851.00	190.00	860.00	**523.00**	309.70	514.20	357.00
KAFNet	189.00	859.00	189.00	861.00	583.80	351.60	697.30	478.10
Hi-Patch	195.00	886.00	190.00	873.00	571.90	369.70	541.60	403.70
HyperIMTS	**188.00**	848.00	191.00	851.00	662.40	505.90	721.70	466.80
GNeuralFlow	190.00	858.00	191.00	862.00	547.30	354.90	710.30	466.20
tPatchGNN	191.00	**840.00**	192.00	**831.00**	577.60	363.90	753.60	510.90
GraFITi	**188.00**	848.00	211.00	990.00	559.30	333.40	718.30	522.90
Warpformer	1610.00	3332.00	1469.00	3174.00	552.50	409.00	537.40	372.80
CRU	190.00	844.00	196.00	883.00	674.10	428.10	617.70	409.40
NeuralFlow	190.00	848.00	190.00	858.00	564.30	350.60	578.60	393.70
mTAN	205.00	897.00	192.00	858.00	1069.30	651.50	587.00	366.70
SeFT	1388.00	2989.00	1765.00	3145.00	799.50	503.70	751.20	505.40
GRU-D	189.00	849.00	189.00	854.00	578.70	365.20	750.10	487.70
QiYao-I (10%)	**188.23**	848.59	**187.93**	848.03	560.59	**260.23**	517.05	**246.45**