MIRA: Medical Time Series Foundation Model for Real-World Health Data

Published in Neurips 2025

Cited by 3

Motivation & The Irregularity

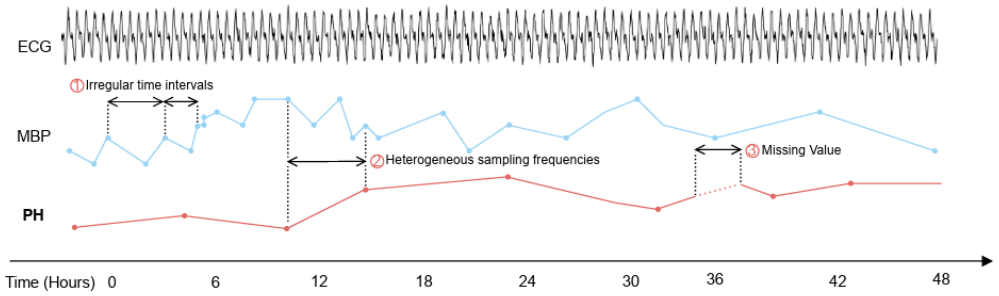

Figure 1: Medical time series exhibit ① irregular intervals, ② heterogeneous sampling rates, and ③ frequent missingness driven by clinical workflows.

- Research question:
  - How can we design a scalable and generalizable foundation model that captures the irregularity and multi-frequency nature of medical time series, while enabling robust transfer across diverse medical tasks?

- MIRA is specifically designed to address the challenges of medical time series:
  - A Continuous-Time Rotary Positional Encoding (CT-RoPE), which extends the standard RoPE [1] to handle time values using learnable frequency adjustments, allowing MIRA to accurately model the varying frequency intervals of irregularly sampled data.
  - MIRA uses a frequency-specific mixture-of-experts (MoE) layer that learns to route frequency-related information through specialized components.
  - Finally, it integrates a Continuous Dynamics Extrapolation Block, using Neural Ordinary Differential Equations (Neural ODEs).

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

Novelties

- They propose MIRA, a new foundation model specially designed for medical time series forecasting. MIRA directly tackles the challenges of irregular sampling and varied temporal dynamics within a single, well-structured system.
- They preprocess and release an extensive and diverse corpus of medical time series, containing over 454 billion observations, to be used for model pre-training. This collection, built from multiple public datasets, aims to cover diverse types of medical time series.
- They establish a comprehensive benchmark suite that covers a wide range of clinical forecasting tasks. This development MIRA into robust and generalizable medical time series models.

Architecture

- MIRA utilizes a decoder-only architecture with three specific adaptations for irregularity:
  - Continuous-Time Rotary Positional Encoding (CT-ROPE): Generalizes standard ROPE to operate directly on real-valued, continuous timestamps.
  - Frequency-Specific Mixture-of-Experts (MoE): Routes tokens to specialized experts based on their temporal dynamics (latent frequency regimes).
  - Continuous Dynamics (CD) Extrapolation Block: Uses Neural ODEs to model the continuous trajectory of latent states, allowing forecasting at any arbitrary future timestamp.

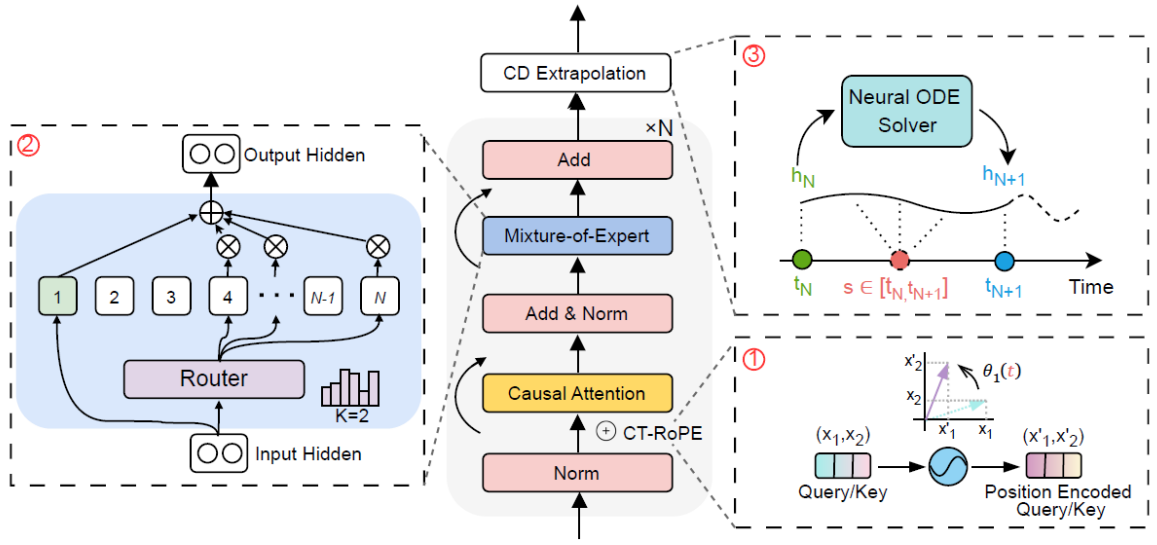

Figure 2: Architecture of **MIRA**. ① Takes irregular medical time series and timestamps as input, applying **CT-RoPE** for continuous temporal encoding. ② A **Sparse Temporal Mixture-of-Experts** layer routes tokens to specialized experts based on frequency. ③ A **Continuous Dynamics Extrapolation Block** evolves latent states toward arbitrary target timestamps for flexible time-aware forecasting.

Technical Focus: CT-ROPE

- Standard ROPE Limitation: Assumes discrete, uniform token indices.
- MIRA Solution: Discretizes continuous timestamps $t \geq 0$ into rotation angles without assuming fixed intervals.
- Key Property: The inner product between the rotated query and key depends only on the relative time difference: $\left(R_\Theta(t_m)\right)^\top R_\Theta(t_n) = R_\Theta(t_n - t_m)$.
- Benefit: Enables the attention mechanism to capture continuous-time relational structures efficiently.

Technical Focus: Sparse Temporal MoE

- Objective: Handle the heterogeneity of medical signals (e.g., millisecond-level ECG vs. hour-level lab tests).
- The Mechanism:
  - Replaces standard feedforward layers with a pool of N lightweight experts.
  - A Router selects the top K (typically K=2) experts for each token.
  - A Shared Expert acts as a global residual pathway for all tokens.
- Outcome: Improved temporal specialization and computational efficiency.

$$\mathrm{MoE}\left(\bar{\mathbf{u}}_t^l\right) = g_{N+1,t} \cdot \mathrm{FFN}_{N+1}\left(\bar{\mathbf{u}}_t^l\right) + \sum_{i=1}^{N} g_{i,t} \cdot \mathrm{FFN}_i\left(\bar{\mathbf{u}}_t^l\right), \tag{6}$$

where $\mathrm{FFN}_i(\cdot)$ denotes the i-th expert network, $\mathrm{FFN}_{N+1}(\cdot)$ the shared expert, and $g_{i,t}$, $g_{N+1,t}$ the corresponding routing weights. The non-shared expert weights $g_{i,t}$ are obtained via a softmax gating mechanism followed by top-K selection:

$$s_{i,t} = \frac{\exp\left((\mathbf{W}_i^l)^\top \bar{\mathbf{u}}_t^l\right)}{\sum_{j=1}^{N} \exp\left((\mathbf{W}_j^l)^\top \bar{\mathbf{u}}_t^l\right)}, \quad g_{i,t} = \begin{cases} s_{i,t}, & \text{if } s_{i,t} \in \mathrm{TopK}(\{s_{j,t}\}_{j=1}^{N}, K), \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\mathbf{W}_i^l \in \mathbb{R}^D$ are the trainable gating vectors. The shared expert weight $g_{N+1,t}$ is computed independently using a sigmoid gate:

$$g_{N+1,t} = \sigma\left((\mathbf{W}_{N+1}^l)^\top \bar{\mathbf{u}}_t^l\right), \tag{8}$$

where $\sigma(\cdot)$ denotes the element-wise sigmoid function and $\mathbf{W}_{N+1}^l \in \mathbb{R}^D$ is the shared expert gating vector.

Continuous Dynamics via Neural ODEs

- The Extrapolation Challenge: Standard transformers cannot incorporate a target timestamp during inference if it wasn't part of the original grid.
- The ODE Approach: MIRA evolves the hidden state h(t_N) to a target time t_{N+1} by integrating the dynamics function f:

$$h(t_{N+1}) = h(t_N) + \int_{t_N}^{t_{N+1}} f(s - t_N, h(s); \theta_{ODE}) ds \tag{10}$$

- Solver: Uses adaptive step-size solvers like Dormand-Prince (RK45) for numerical stability.
- Result: Accurate forecasting at arbitrary, unseen, or irregular time points.

Large-Scale Medical Pretraining

- MIRA was pretrained on a curated corpus of 454 billion observations
- Training Strategy: Resumed from Time-MoE checkpoints and trained using Huber loss for robustness against clinical outliers.

$$\mathcal{L}_{\text{Huber}}(x, \hat{x}) = \begin{cases} \frac{1}{2}(x - \hat{x})^2, & \text{if } |x - \hat{x}| \leq \delta, \\ \delta \cdot (|x - \hat{x}| - \frac{1}{2}\delta), & \text{otherwise,} \end{cases} \tag{11}$$

- They adopt a similar model configuration strategy following Time-MoE by providing three model variants of increasing scale

Table 1: A high-level summary of MIRA model configurations.

	Layers	Experts	K	d_{model}	d_{ff}	d_{expert}	Activated Params	Total Params
$MIRA_{small}$	8	8	2	288	1152	144	30 M	73 M
$MIRA_{base}$	12	8	2	384	1536	192	50 M	114 M
$MIRA_{large}$	12	8	2	768	3072	384	200 M	455 M

Result

- Out-of-Distribution (OOD) Forecasting MIRA outperformed all 13 baselines (including TimesFM and Chronos) on unseen datasets like MIT-BIH and CDC-IHA.
- MIRA large reduced forecasting error (RMSE) by an average of 8% in OOD scenarios.

Table 2: Zero-shot forecasting performance on out-of-distribution datasets that are regularly sampled but contain missing values. Reported values are averaged across all prediction lengths. Lower RMSE and MAE indicate better predictions. **Red**: the best; Blue: the second best compared to zero-shot baselines; Underline: the best performance compared to full-shot baselines.

Models	Zero-shot Ours						Full-shot Time Series Models									
	MIRA$_{small}$		MIRA$_{base}$		MIRA$_{large}$		Contiformer		T-PatchGNN		ODE-RNN		Neural-CDE		TimesFM	
Metrics	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Cinc 2012 (10^1)	7.136	6.984	6.762	6.734	6.221	6.115	5.987	5.985	6.247	6.246	6.997	6.995	7.498	7.497	-	-
Heart Rate (10^{-1})	1.795	1.511	1.723	1.431	1.392	1.310	0.774	0.633	0.627	0.497	0.945	0.683	0.671	0.587	1.753	0.832
MIT-BIH	0.293	0.198	0.199	0.141	0.173	0.130	0.453	0.354	0.705	0.627	0.882	0.623	0.242	0.196	0.335	0.141
CDC-IHA (10^1)	5.976	4.684	5.729	4.502	5.534	4.401	5.211	4.103	9.522	7.974	10.068	9.052	7.892	6.766	15.633	4.408
JH COVID-19 (10^2)	0.407	0.355	0.504	0.349	0.478	0.336	0.323	0.297	0.350	0.291	0.424	0.331	0.545	0.503	2.329	0.322
ILI	1.294	1.077	1.218	1.024	1.154	1.041	0.391	0.224	0.195	0.143	0.410	0.264	0.423	0.314	2.034	1.333
1st Count	0	0	0	0	4	3	0	0	0	0	0	0	0	0	0	0

Baseline	Zero-shot Foundation Models Pre-trained on General-domain Corpora															
	Lag-Llama		TimeGPT		Timer		Moment$_{small}$		Moment$_{base}$		Moment$_{large}$		Moirai-MoE$_{small}$		Moirai-MoE$_{base}$	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Heart Rate (10^{-1})	1.764	1.488	2.258	1.915	1.901	1.704	2.966	1.852	2.939	1.835	2.917	1.735	1.982	1.600	2.144	1.652
MIT-BIH	0.217	0.169	0.231	0.185	0.255	0.213	0.417	0.229	0.416	0.228	0.413	0.224	0.208	0.167	0.208	0.167
CDC-IHA (10^1)	6.531	4.846	6.654	4.860	6.424	4.857	17.803	5.307	17.631	5.288	17.689	5.260	7.099	5.405	7.639	5.862
JH COVID-19 (10^2)	3.596	1.432	1.879	1.580	2.647	2.328	3.077	0.553	3.097	0.554	3.064	0.549	1.452	0.725	19.391	3.190
ILI	1.780	1.366	2.011	1.077	1.882	1.492	1.570	1.056	1.566	1.054	1.566	1.052	2.001	1.678	1.983	1.664
1st Count	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Baseline	Zero-shot Foundation Models Pre-trained on General-domain Corpora															
	Moirai$_{small}$		Moirai$_{base}$		Moirai$_{large}$		Time-MoE$_{base}$		Time-MoE$_{large}$		Chronos$_{small}$		Chronos$_{base}$		Chronos$_{large}$	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Heart Rate (10^{-1})	2.359	1.965	2.156	1.767	2.098	1.644	0.850	0.650	0.833	0.639	1.357	0.587	1.189	0.489	1.218	0.506
MIT-BIH	0.343	0.249	0.421	0.302	0.593	0.149	0.171	0.135	0.172	0.135	0.353	0.147	0.361	0.149	0.350	0.147
CDC-IHA (10^1)	6.835	5.271	7.328	5.526	6.788	5.302	6.311	4.748	6.312	4.715	15.502	4.421	15.825	4.438	15.986	4.517
JH COVID-19 (10^2)	1.917	0.695	0.991	0.474	0.614	0.402	0.596	0.402	0.512	0.371	4.826	1.031	3.835	0.551	3.478	0.521
ILI	1.995	1.671	1.871	1.561	1.808	1.499	1.288	0.951	1.366	1.015	2.054	1.400	1.940	1.308	1.870	1.252
1st Count	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Baseline	Zero-shot Foundation Models Continue Pre-trained on Medical Corpora															
	Moirai$_{small}$		Moirai$_{base}$		Moirai$_{large}$		Time-MoE$_{base}$		Time-MoE$_{large}$		Chronos$_{small}$		Chronos$_{base}$		Chronos$_{large}$	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Heart Rate (10^{-1})	2.047	1.685	1.907	1.536	1.601	1.263	0.648	0.524	0.603	0.488	1.049	0.555	0.965	0.488	0.902	0.458
MIT-BIH	0.239	0.204	0.274	0.190	0.219	0.166	0.201	0.155	0.185	0.143	0.311	0.137	0.320	0.139	0.306	0.127
CDC-IHA (10^1)	6.690	5.310	6.934	5.376	6.696	5.114	6.327	4.698	6.299	4.666	14.502	4.321	14.825	4.338	14.986	4.417
JH COVID-19 (10^2)	0.579	0.448	0.619	0.478	0.812	0.337	0.509	0.353	0.517	0.362	4.225	0.947	3.535	0.510	3.328	0.469
ILI	1.528	1.289	1.435	1.191	1.501	1.229	1.188	0.915	1.201	0.927	1.705	1.127	1.639	1.081	1.547	1.028
1st Count	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0

Result

- Evaluated on the WHO FluNet dataset with missing rates from 10% to 90%.
- MIRA showed minimal performance degradation even at severe (90%+) information loss, while competitor models dropped quickly.

Table 4: Zero-shot forecasting performance on different missing rates. Reported values are averaged across all prediction lengths. Lower RMSE and MAE indicate better predictions. **Red**: the best; Blue: the second best.

	Missing Rate	10%+		20%+		30%+		40%+		50%+		60%+		70%+		80%+		90%+		1st Count	
	Metrics (10^2)	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Ours	MIRA$_{small}$	3.15	2.82	3.28	2.95	3.41	3.08	3.54	3.20	3.67	3.34	3.80	3.47	3.93	3.61	4.07	3.75	4.22	3.90	0	0
	MIRA$_{base}$	2.98	2.68	3.11	2.81	3.25	2.94	3.38	3.07	3.51	3.20	3.64	3.34	3.78	3.48	3.93	3.63	4.10	3.79	0	1
	MIRA$_{large}$	2.85	2.56	2.95	2.65	3.08	2.78	3.21	2.90	3.35	3.04	3.49	3.18	3.63	3.31	3.99	3.86	3.97	3.64	8	8
General Setting	Time-MoE$_{base}$	3.05	2.78	3.18	2.92	3.32	3.05	3.45	3.18	3.58	3.31	3.71	3.43	3.85	3.57	3.99	3.71	4.15	3.86	0	0
	Time-MoE$_{large}$	3.20	2.91	3.33	3.04	3.47	3.17	3.60	3.29	3.73	3.41	3.86	3.54	3.99	3.67	4.13	3.81	4.28	3.96	0	0
	Moirai$_{large}$	4.80	4.18	5.39	4.64	5.68	4.55	6.02	4.83	6.09	5.06	6.18	5.02	6.31	5.09	6.51	5.24	6.71	5.36	0	0
	Moirai$_{base}$	4.84	4.15	4.93	4.10	5.20	4.23	5.64	4.53	5.78	4.61	5.80	4.60	5.95	4.97	5.95	4.73	6.06	4.87	0	0
	Moirai$_{small}$	4.91	4.22	5.04	4.20	6.12	4.93	6.23	4.98	6.44	5.49	6.47	5.18	6.59	5.60	6.66	5.20	6.86	5.71	0	0
	Chronos$_{large}$	4.71	4.10	4.85	4.22	5.02	4.31	5.31	4.52	5.58	4.70	5.82	4.91	5.97	5.10	6.15	5.32	6.43	5.49	0	0
	Chronos$_{base}$	5.02	4.40	5.21	4.51	5.43	4.68	5.62	4.81	5.85	4.93	6.02	5.12	6.21	5.31	6.40	5.49	6.67	5.62	0	0
	Chronos$_{small}$	5.32	4.71	5.49	4.82	5.68	4.93	5.81	5.02	6.01	5.20	6.21	5.32	6.40	5.49	6.58	5.61	6.83	5.72	0	0
Medical Setting	Time-MoE$_{large}$	2.95	2.70	3.08	2.83	3.22	2.96	3.35	3.09	3.48	3.22	3.61	3.35	3.75	3.49	3.89	3.63	4.04	3.78	1	0
	Time-MoE$_{base}$	3.05	2.78	3.18	2.91	3.32	3.04	3.45	3.17	3.58	3.30	3.71	3.43	3.85	3.57	3.99	3.71	4.15	3.86	0	0
	Moirai$_{large}$	4.41	3.76	4.92	4.20	5.35	4.41	5.69	4.76	5.93	4.80	6.09	4.92	6.23	4.98	6.48	5.29	6.62	5.30	0	0
	Moirai$_{base}$	4.55	3.92	4.78	4.07	5.18	4.26	5.49	4.46	5.78	4.64	5.70	4.51	5.87	4.97	5.92	4.75	5.96	4.71	0	0
	Moirai$_{small}$	5.02	4.21	5.36	4.40	5.55	4.98	5.97	4.92	6.25	5.34	6.43	5.11	6.44	5.35	6.57	5.41	6.58	5.30	0	0
	Chronos$_{large}$	4.30	3.89	4.58	4.02	4.83	4.20	5.02	4.31	5.32	4.51	5.61	4.73	5.83	4.91	6.02	5.10	6.31	5.32	0	0
	Chronos$_{base}$	4.61	4.10	4.83	4.20	5.02	4.31	5.32	4.51	5.58	4.70	5.82	4.91	5.97	5.10	6.15	5.32	6.43	5.49	0	0
	Chronos$_{small}$	4.83	4.31	5.02	4.43	5.32	4.51	5.58	4.70	5.82	4.91	5.97	5.10	6.15	5.32	6.43	5.49	6.71	5.61	0	0

Ablation Study

Component Impact: Removing the CD-Extrapolation Block caused the largest performance drop (5-8% increase in error), highlighting its critical role in forecasting.

Table 5: Ablation studies. (**Left**) Evaluated with different model components. (**Right**) Analysis performance and inference speed across different top$_k$ setups. Lower values indicate better performance.

Component	**RMSE**	**MAE**
$MIRA_{base}$	0.154	0.118
w/o CT-RoPE	0.158	0.125
w/o MoE Block	0.157	0.122
w/o CT-Extrapolation Block	0.162	0.128

Top$_k$ Setup	**RMSE**	**MAE**	**Speed (s/iter)**
w/ {Top$_1$}	0.160	0.121	0.097
w/ {Top$_2$}	0.154	0.118	0.101
w/ {Top$_4$}	0.154	0.117	0.112
w/ {Top$_6$}	0.156	0.120	0.124
w/ {Top$_8$}	0.159	0.122	0.127

Conclusion & Limitations

- Summary: MIRA is the first large-scale foundation model to successfully integrate continuous-time modeling with sparse MoE for medical time series.
- Clinical Potential: Reduces the need for task-specific tuning and helps share knowledge across diverse institutions and data modalities.
- Limitations: While pretrained on large public datasets, real-world deployment complexity and residual privacy risks remain areas for future work.

Related Work

Table 6: Comparison between time series models.

Method	MIRA (Ours)	Sundial (2025)	Time-MoE (2024)	Moirai-MoE (2024)	Moirai (2024)	TimesFM (2024)	Moment (2024)	Chronos (2024)	Timer (2024)	Lag-Llama (2023)	TimeGPT (2023)
Architecture	Decoder	Decoder	Decoder	Decoder	Encoder	Decoder	Encoder	EncDec	Decoder	Decoder	EncDec
(Max) Model Size	455M	444M	2.4B	1.1B	311M	500M	385M	710M	67M	200M	Unknown
Input Token	Point	Patch	Point	Patch	Patch	Patch	Patch	Point	Patch	Point	Patch
Dataset Scale	454B	1TB	309B	27B	27B/231B*	100B	1.13B	84B	29B	0.36B	100B
Max Length	512	2880	4096	5000	5000	512	512	512	1440	1024	Unknown
FFN	Sparse	Dense	Sparse	Sparse	Dense	Dense	Dense	Dense	Dense	Dense	Dense

* Depend on the way of calculation according to the original paper.

Dataflow:

- From Raw Input to Prediction
  - Raw Input Layer Data Collection:
    - The model receives paired sequences of timestamps and observations $\{(t_i, x_i)\}_{\{i=1\}}^{N}$.
    - Observation Handling: x_i can be real values or missing entries (NaN), representing either regular-grid missingness or irregular sampling.
  - Preprocessing & Normalization
    - Time Normalization: Raw timestamps are rescaled into a canonical range using min-max normalization to ensure stable phase computation.
    - Embedding: Observations are mapped into a $d_{\{model\}}$ dimensionality through initial input hidden layers.
  - Stacked Decoder Layers (x N)
    - Normalization: Tokens pass through a LayerNorm.
    - Continuous-Time Rotary Positional Encoding (CT-RoPE): Linear rotation angles $\theta_i(t)$ are applied to the latent features based on their continuous timestamps, allowing the model to capture relative temporal differences.
    - Causal Attention: Standard dot-product attention with causal masking ensures the model only processes historical information.
    - Frequency-Specific Mixture-of-Experts (MoE): Latent tokens are routed to specialized experts (Top-2) based on their temporal frequency, with a shared expert acting as a global residual pathway.
    - Continuous Dynamics (CD) Extrapolation BlockState Integration: The final processed latent state $h(t_N)$ and the target timestamp $t_{\{N+1\}}$ are fed into a Neural ODE solver.
    - Neural ODE Solver (RK45): The solver integrates the dynamics function f over the interval $[t_N, t_{\{N+1\}}]$ to "evolve" the hidden state to the future target point.
  - Output Head
    - Value Prediction: The evolved latent state $h(t_{N+1})$ is projected back to the observation space to produce the predicted value $\hat{x}_{\{L+1:L+H\}}$.