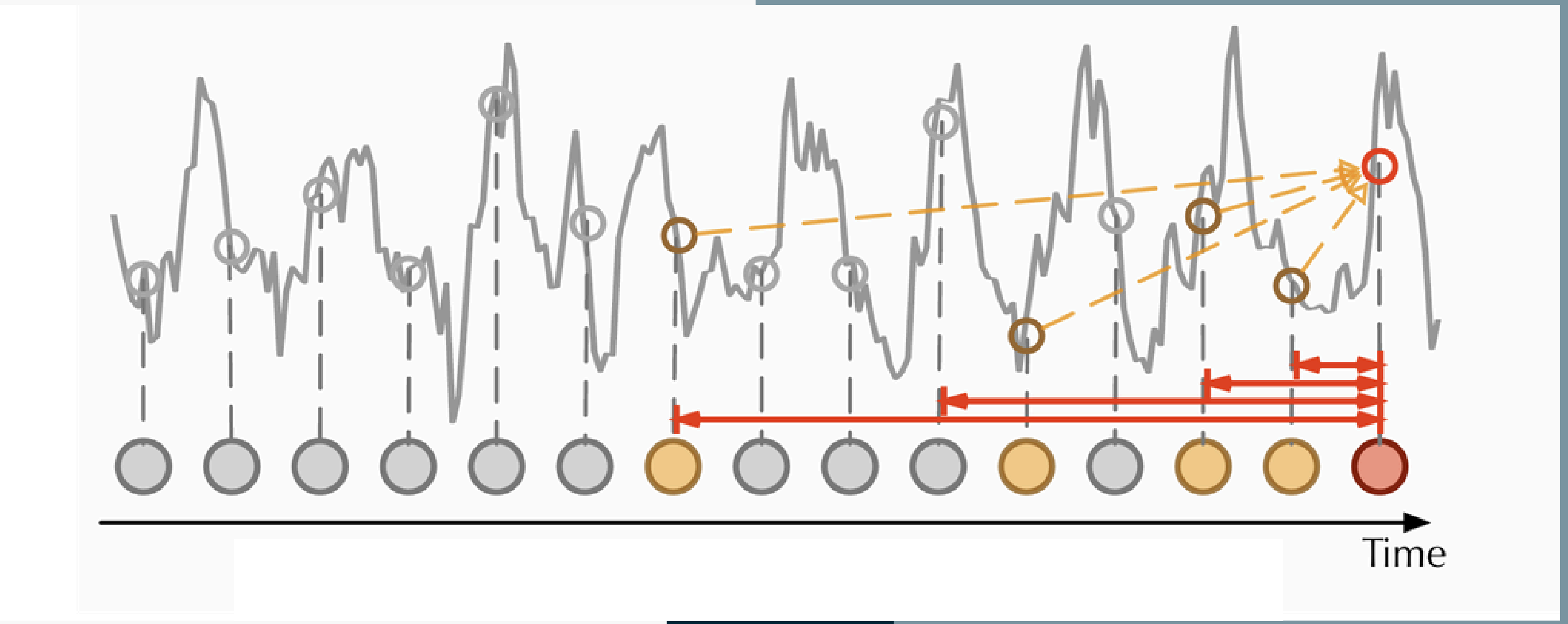
Autoformer

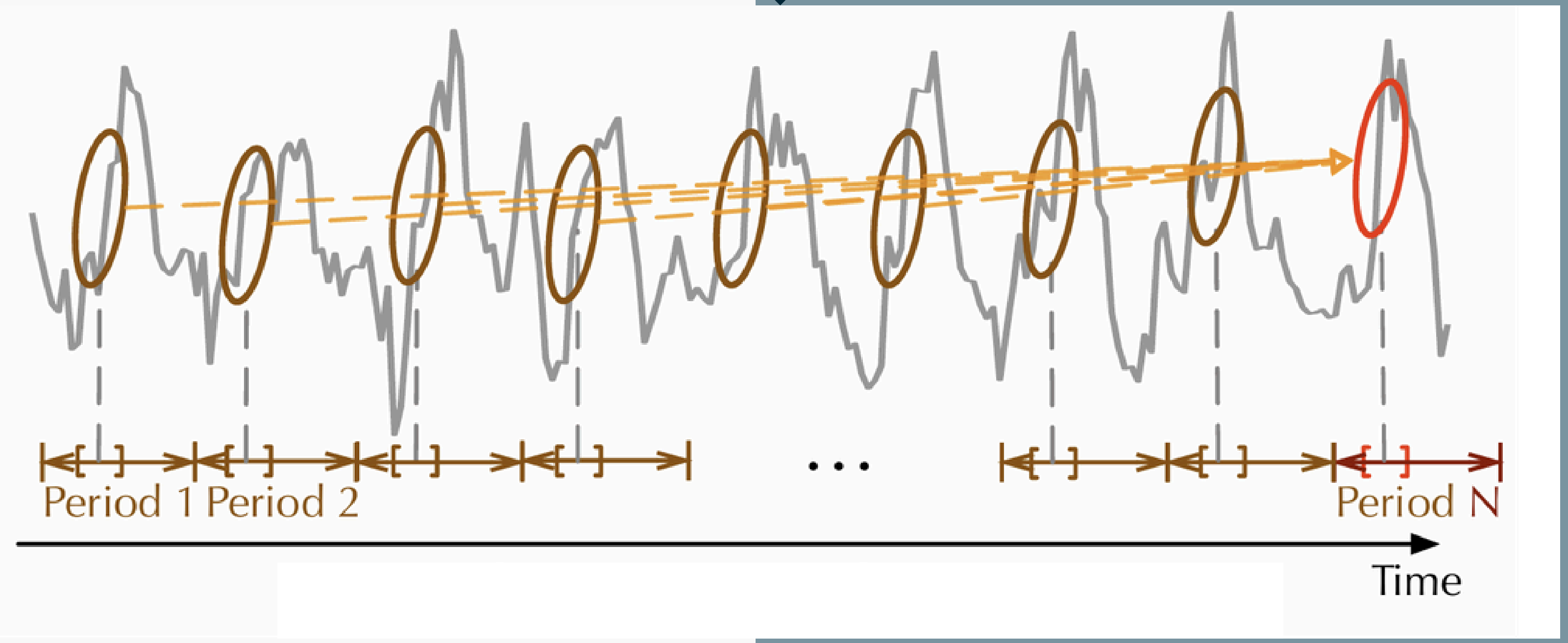
Mohamed Yassir Ousdid

April 16th , 2025

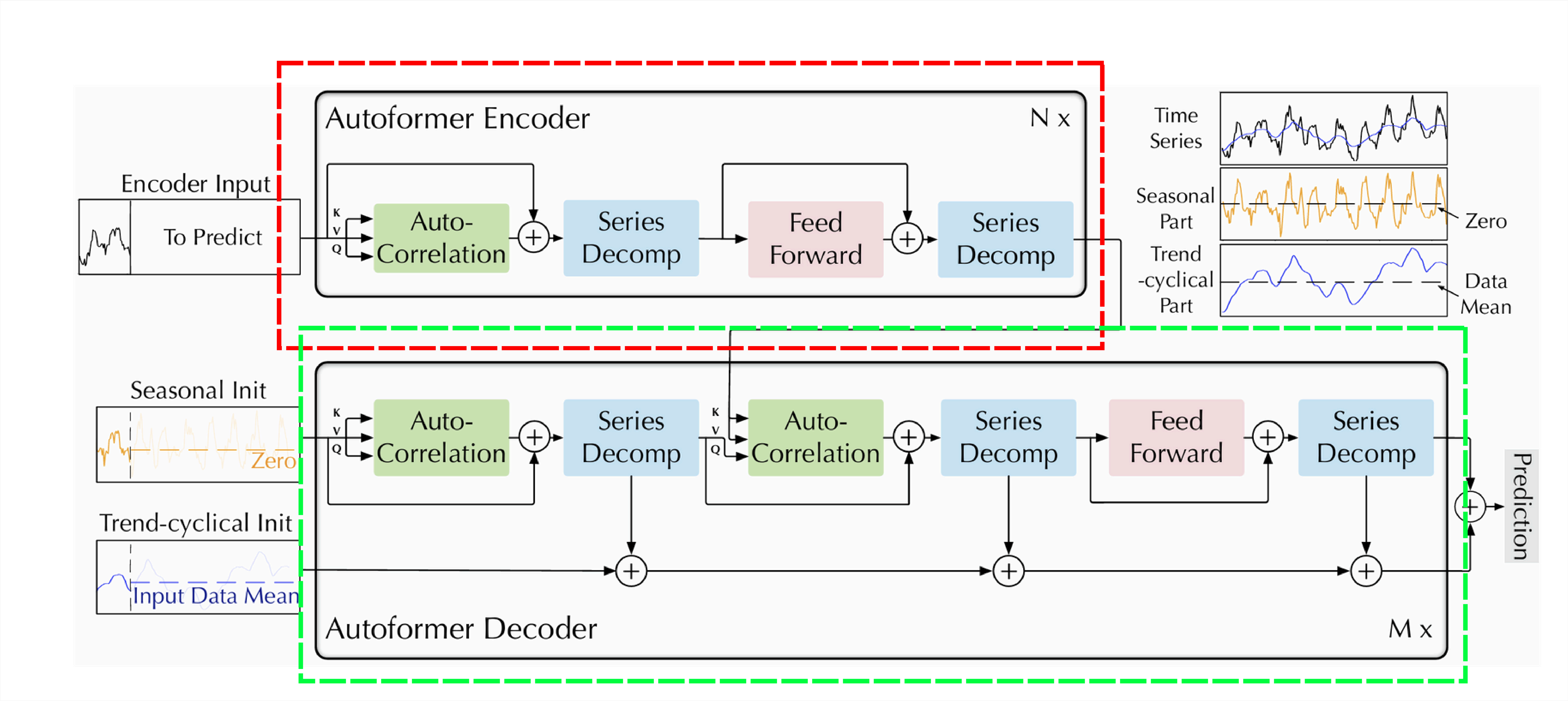
Introduction & Motivation

Previous Transformer Based Model	**Autoformer Model [1] (2021)**
Patterns get mixed up and hard to understand in long-term data	Uses decomposition to split the data into trend and seasonality
Point-wise attention is inefficient and causes a bottleneck in capturing long-term patterns	Uses auto-correlation to find and compare repeating patterns across the time series

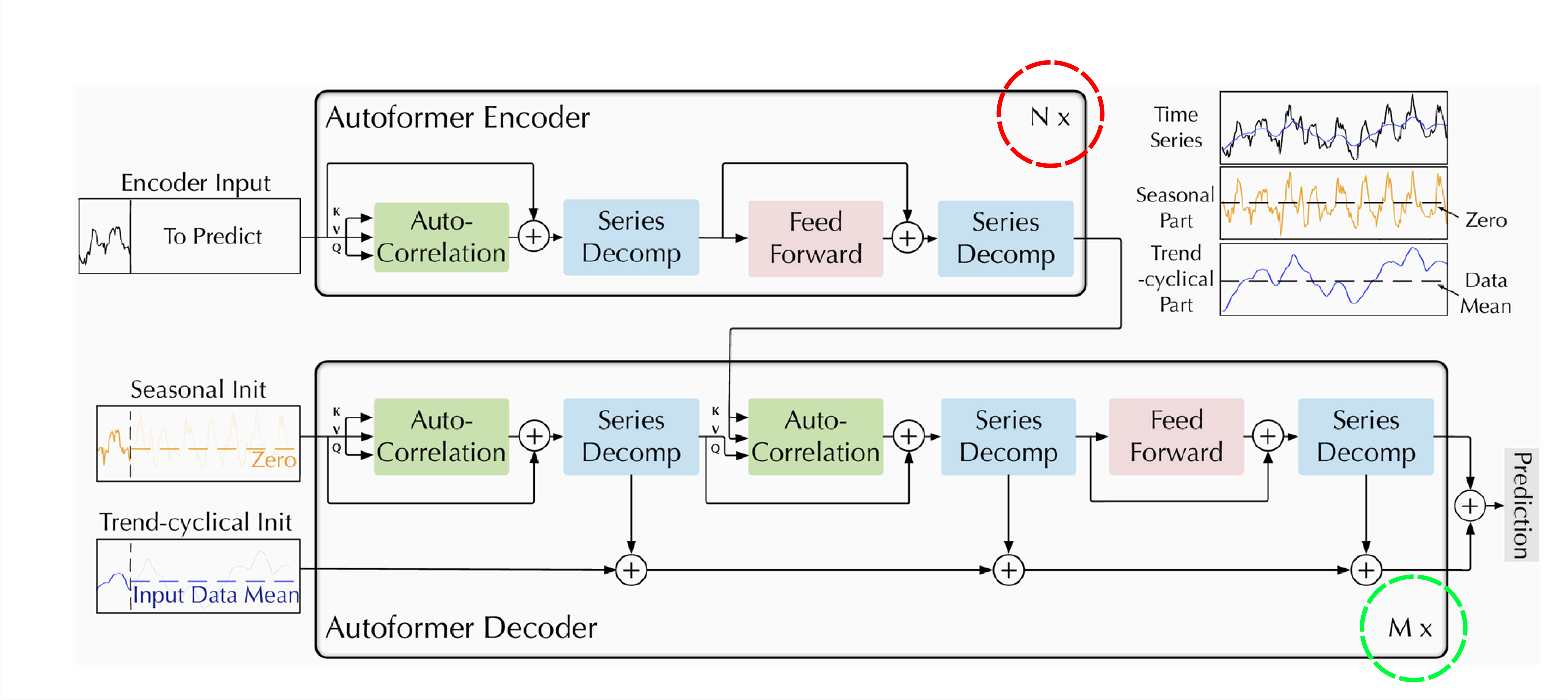




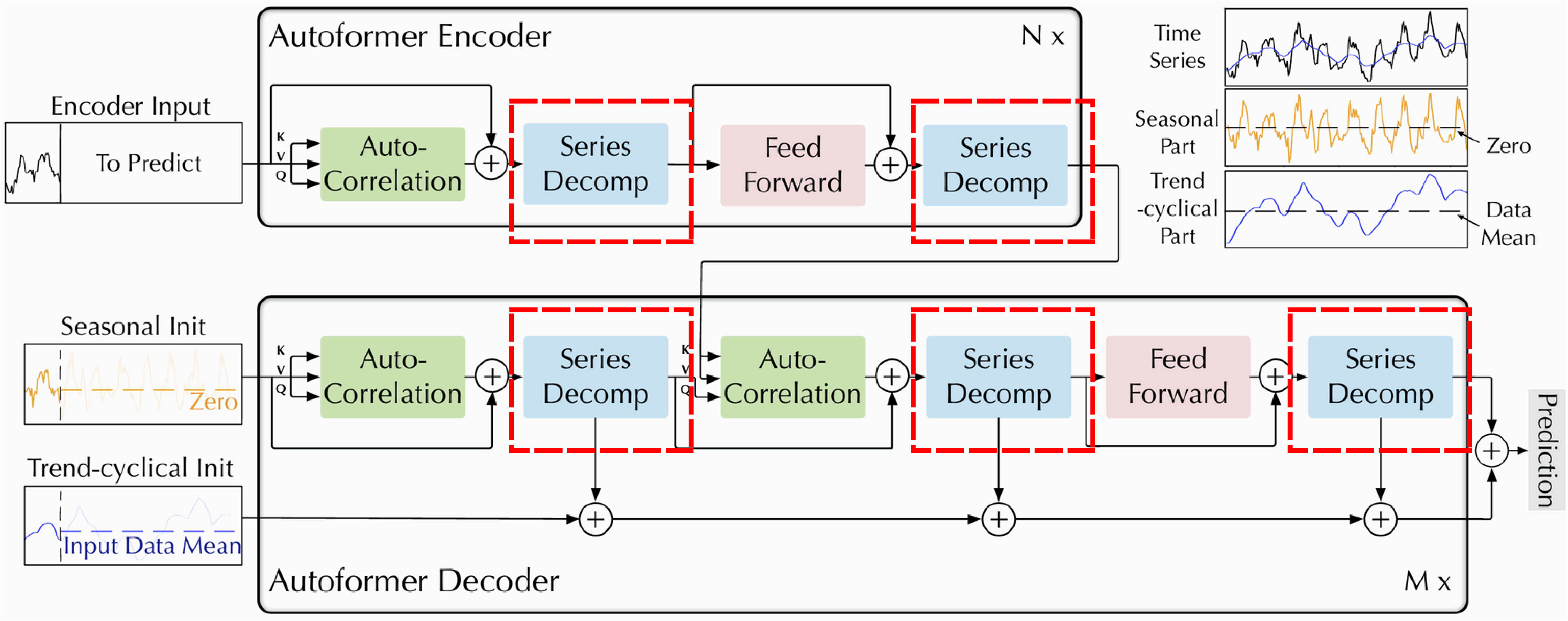
Architecture



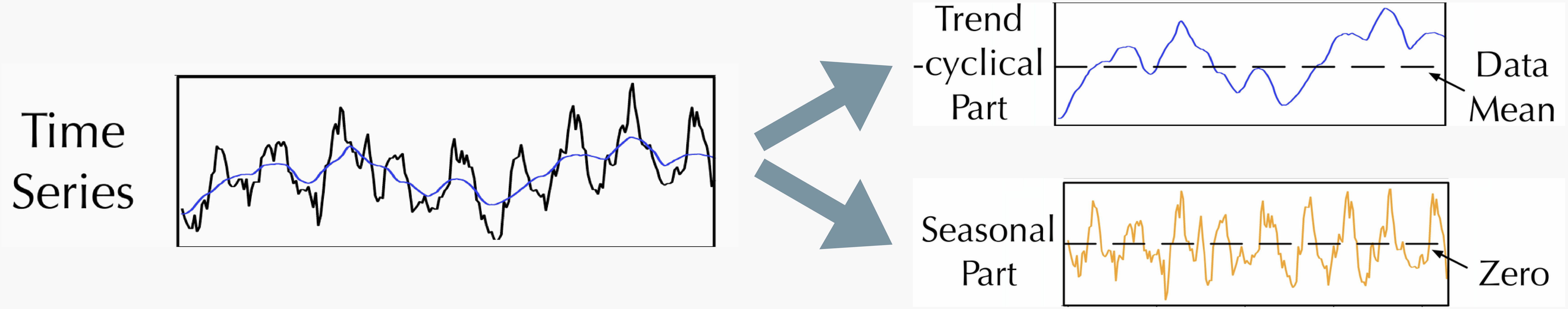
Architecture



Architecture

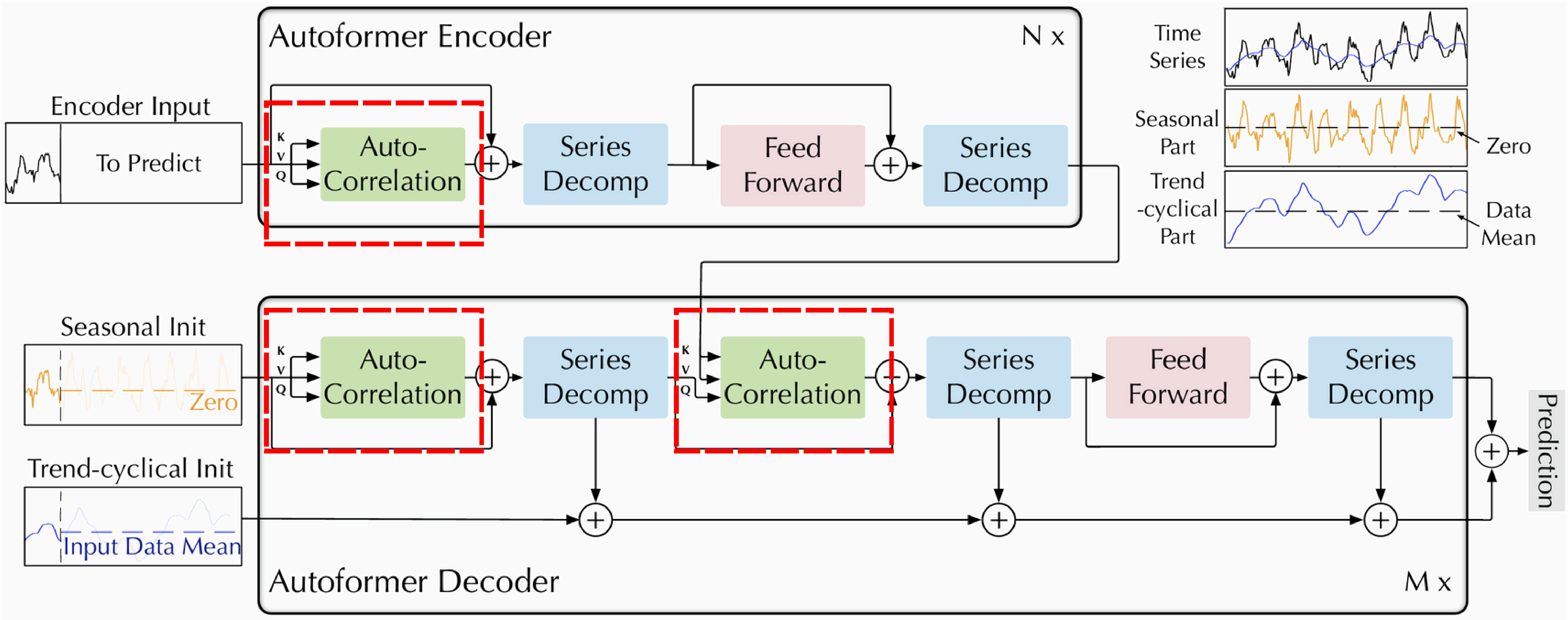


Architecture : Decomposition



$$\mathcal{X}_{\mathrm{t}} = \mathrm{AvgPool}(\mathrm{Padding}(\mathcal{X}))$$
$$\mathcal{X}_{\mathrm{s}} = \mathcal{X} - \mathcal{X}_{\mathrm{t}},$$

Architecture



Architecture : AutoCorrelation

Autocorrelation:

$$\mathcal{R}_{\mathcal{X}\mathcal{X}}(\tau) = \lim_{L\to\infty} \frac{1}{L} \sum_{t=1}^{L} \mathcal{X}_t \mathcal{X}_{t-\tau}.$$

Wiener–Khinchin theorem :

$$\mathcal{S}_{\mathcal{X}\mathcal{X}}(f) = \mathcal{F}(\mathcal{X}_t)\,\mathcal{F}^*(\mathcal{X}_t) = \int_{-\infty}^{\infty} \mathcal{X}_t e^{-i2\pi tf}\mathrm{d}t\,\overline{\int_{-\infty}^{\infty} \mathcal{X}_t e^{-i2\pi tf}\mathrm{d}t}$$

$$\mathcal{R}_{\mathcal{X}\mathcal{X}}(\tau) = \mathcal{F}^{-1}(\mathcal{S}_{\mathcal{X}\mathcal{X}}(f)) = \int_{-\infty}^{\infty} \mathcal{S}_{\mathcal{X}\mathcal{X}}(f) e^{i2\pi f\tau}\mathrm{d}f,$$

$\rightarrow$ $\mathcal{O}(L \log L)$

Architecture : AutoCorrelation

$$\tau_1, \cdots, \tau_k = \underset{\tau \in \{1, \cdots, L\}}{\arg \operatorname{Topk}} \left(\mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau) \right)$$

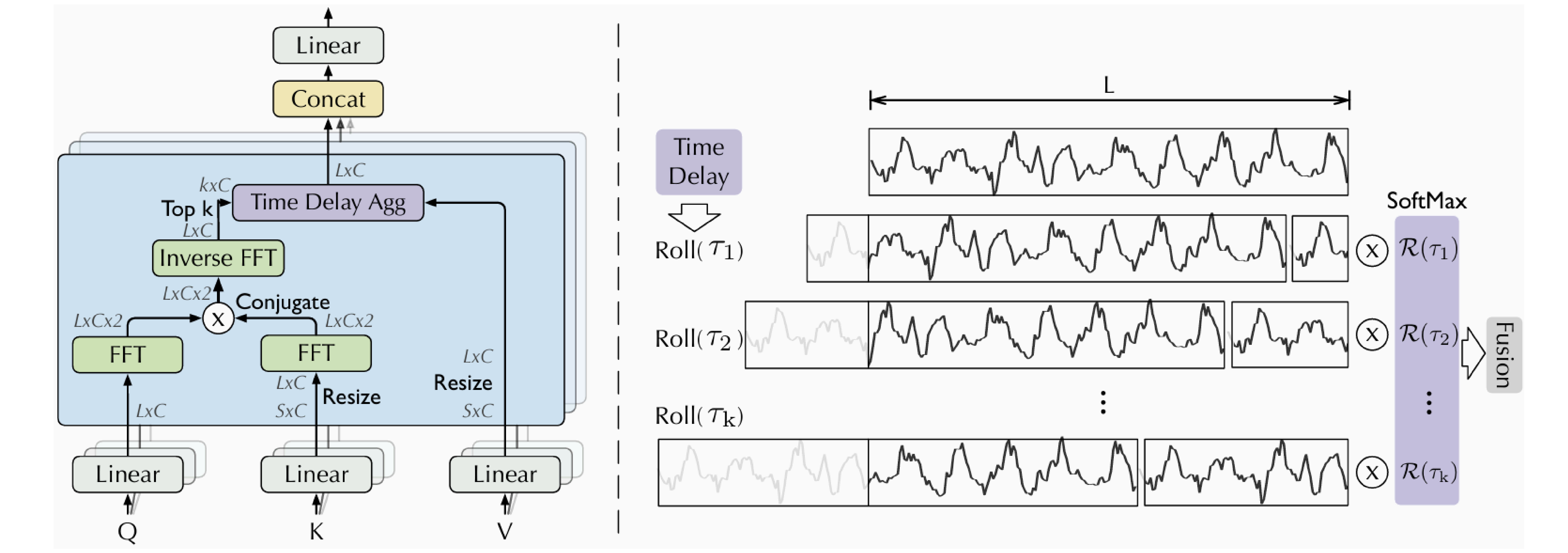
$$\widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_1), \cdots, \widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_k) = \operatorname{SoftMax}\left(\mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau_1), \cdots, \mathcal{R}_{\mathcal{Q},\mathcal{K}}(\tau_k) \right)$$

$$\text{Auto-Correlation}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \sum_{i=1}^{k} \operatorname{Roll}(\mathcal{V}, \tau_i) \widehat{\mathcal{R}}_{\mathcal{Q},\mathcal{K}}(\tau_i),$$

- **Pick top-k most correlated repeating patterns**
- **Use a softmax to give probabilities**
- **Roll and combine with a weighted sum**

→ $\mathcal{O}(L \log L)$

Architecture : AutoCorrelation



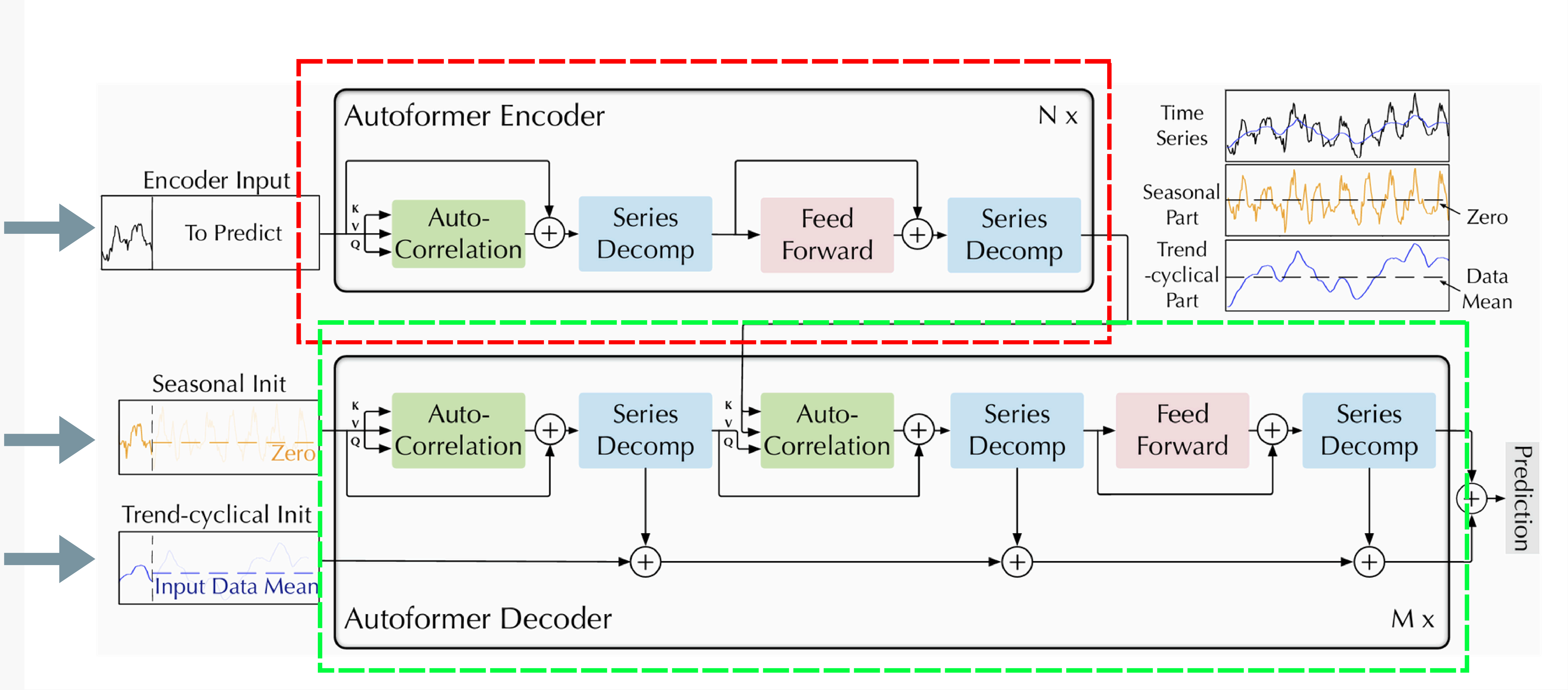
$$\mathrm{MultiHead}(\mathcal{Q},\mathcal{K},\mathcal{V}) = \mathcal{W}_{\mathrm{output}} * \mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)$$
$$\text{where } \mathrm{head}_i = \text{Auto-Correlation}(\mathcal{Q}_i,\mathcal{K}_i,\mathcal{V}_i).$$

Architecture : Initialization

$\mathcal{X}_{\text{en}}$

$\mathcal{X}_{\text{des}}$

$\mathcal{X}_{\text{det}}$



$$
\begin{aligned}
\mathcal{X}_{\text{ens}}, \mathcal{X}_{\text{ent}} &= \text{SeriesDecomp}(\mathcal{X}_{\text{en}\,\frac{I}{2}:I}) \\
\mathcal{X}_{\text{des}} &= \text{Concat}(\mathcal{X}_{\text{ens}}, \mathcal{X}_0) \\
\mathcal{X}_{\text{det}} &= \text{Concat}(\mathcal{X}_{\text{ent}}, \mathcal{X}_{\text{Mean}}),
\end{aligned}
$$

Architecture : Encoder Decoder

Encoder :

$$\mathcal{S}_{\text{en}}^{l,1}, \_ = \text{SeriesDecomp}\Big(\text{Auto-Correlation}(\mathcal{X}_{\text{en}}^{l-1}) + \mathcal{X}_{\text{en}}^{l-1}\Big)$$

$$\mathcal{S}_{\text{en}}^{l,2}, \_ = \text{SeriesDecomp}\Big(\text{FeedForward}(\mathcal{S}_{\text{en}}^{l,1}) + \mathcal{S}_{\text{en}}^{l,1}\Big),$$

Decoder :

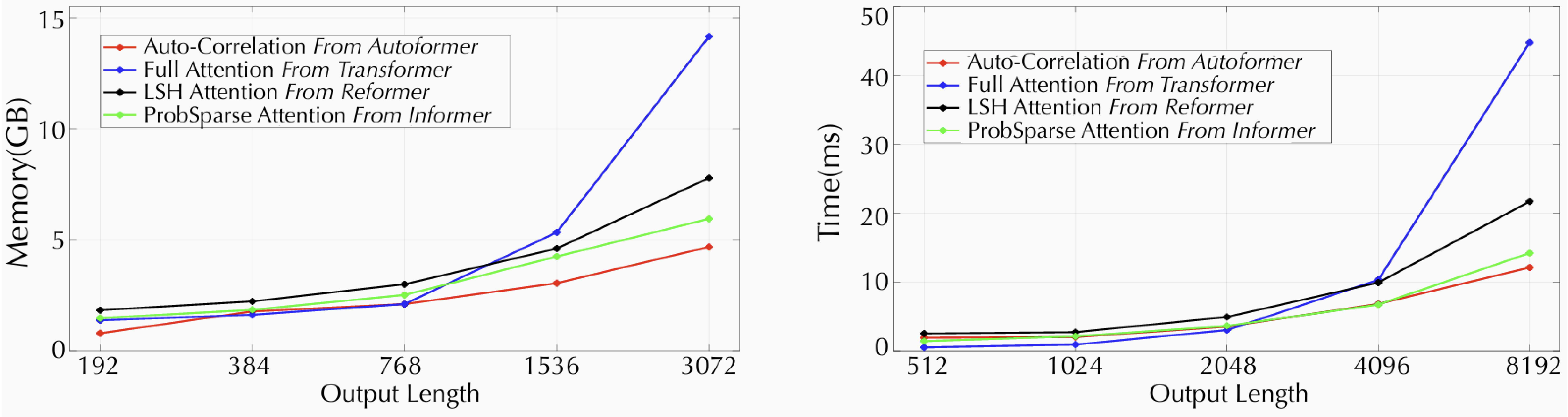
$$\mathcal{S}_{\text{de}}^{l,1}, \mathcal{T}_{\text{de}}^{l,1} = \text{SeriesDecomp}\Big(\text{Auto-Correlation}(\mathcal{X}_{\text{de}}^{l-1}) + \mathcal{X}_{\text{de}}^{l-1}\Big)$$

$$\mathcal{S}_{\text{de}}^{l,2}, \mathcal{T}_{\text{de}}^{l,2} = \text{SeriesDecomp}\Big(\text{Auto-Correlation}(\mathcal{S}_{\text{de}}^{l,1}, \mathcal{X}_{\text{en}}^{N}) + \mathcal{S}_{\text{de}}^{l,1}\Big)$$

$$\mathcal{S}_{\text{de}}^{l,3}, \mathcal{T}_{\text{de}}^{l,3} = \text{SeriesDecomp}\Big(\text{FeedForward}(\mathcal{S}_{\text{de}}^{l,2}) + \mathcal{S}_{\text{de}}^{l,2}\Big)$$

$$\mathcal{T}_{\text{de}}^{l} = \mathcal{T}_{\text{de}}^{l-1} + \mathcal{W}_{l,1} * \mathcal{T}_{\text{de}}^{l,1} + \mathcal{W}_{l,2} * \mathcal{T}_{\text{de}}^{l,2} + \mathcal{W}_{l,3} * \mathcal{T}_{\text{de}}^{l,3},$$

Effeciency Analysis



(a) Memory Efficiency Analysis (b) Running Time Efficiency Analysis

$$\rightarrow \mathcal{O}(L \log L)$$

Experiments : Multivariate

Models		**Autoformer**		Informer[48]		LogTrans[26]		Reformer[23]		LSTNet[25]		LSTM[17]		TCN[4]	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT*	96	**0.255**	**0.339**	0.365	0.453	0.768	0.642	0.658	0.619	3.142	1.365	2.041	1.073	3.041	1.330
	192	**0.281**	**0.340**	0.533	0.563	0.989	0.757	1.078	0.827	3.154	1.369	2.249	1.112	3.072	1.339
	336	**0.339**	**0.372**	1.363	0.887	1.334	0.872	1.549	0.972	3.160	1.369	2.568	1.238	3.105	1.348
	720	**0.422**	**0.419**	3.379	1.388	3.048	1.328	2.631	1.242	3.171	1.368	2.720	1.287	3.135	1.354
Electricity	96	**0.201**	**0.317**	0.274	0.368	0.258	0.357	0.312	0.402	0.680	0.645	0.375	0.437	0.985	0.813
	192	**0.222**	**0.334**	0.296	0.386	0.266	0.368	0.348	0.433	0.725	0.676	0.442	0.473	0.996	0.821
	336	**0.231**	**0.338**	0.300	0.394	0.280	0.380	0.350	0.433	0.828	0.727	0.439	0.473	1.000	0.824
	720	**0.254**	**0.361**	0.373	0.439	0.283	0.376	0.340	0.420	0.957	0.811	0.980	0.814	1.438	0.784
Exchange	96	**0.197**	**0.323**	0.847	0.752	0.968	0.812	1.065	0.829	1.551	1.058	1.453	1.049	3.004	1.432
	192	**0.300**	**0.369**	1.204	0.895	1.040	0.851	1.188	0.906	1.477	1.028	1.846	1.179	3.048	1.444
	336	**0.509**	**0.524**	1.672	1.036	1.659	1.081	1.357	0.976	1.507	1.031	2.136	1.231	3.113	1.459
	720	**1.447**	**0.941**	2.478	1.310	1.941	1.127	1.510	1.016	2.285	1.243	2.984	1.427	3.150	1.458
Traffic	96	**0.613**	**0.388**	0.719	0.391	0.684	0.384	0.732	0.423	1.107	0.685	0.843	0.453	1.438	0.784
	192	**0.616**	**0.382**	0.696	0.379	0.685	0.390	0.733	0.420	1.157	0.706	0.847	0.453	1.463	0.794
	336	**0.622**	**0.337**	0.777	0.420	0.733	0.408	0.742	0.420	1.216	0.730	0.853	0.455	1.479	0.799
	720	**0.660**	**0.408**	0.864	0.472	0.717	0.396	0.755	0.423	1.481	0.805	1.500	0.805	1.499	0.804
Weather	96	**0.266**	**0.336**	0.300	0.384	0.458	0.490	0.689	0.596	0.594	0.587	0.369	0.406	0.615	0.589
	192	**0.307**	**0.367**	0.598	0.544	0.658	0.589	0.752	0.638	0.560	0.565	0.416	0.435	0.629	0.600
	336	**0.359**	**0.395**	0.578	0.523	0.797	0.652	0.639	0.596	0.597	0.587	0.455	0.454	0.639	0.608
	720	**0.419**	**0.428**	1.059	0.741	0.869	0.675	1.130	0.792	0.618	0.599	0.535	0.520	0.639	0.610
ILI	24	**3.483**	**1.287**	5.764	1.677	4.480	1.444	4.400	1.382	6.026	1.770	5.914	1.734	6.624	1.830
	36	**3.103**	**1.148**	4.755	1.467	4.799	1.467	4.783	1.448	5.340	1.668	6.631	1.845	6.858	1.879
	48	**2.669**	**1.085**	4.763	1.469	4.800	1.468	4.832	1.465	6.080	1.787	6.736	1.857	6.968	1.892
	60	**2.770**	**1.125**	5.264	1.564	5.278	1.560	4.882	1.483	5.548	1.720	6.870	1.879	7.127	1.918

* *ETT* means the ETTm2. See Appendix A for the **full benchmark** of ETTh1, ETTh2, ETTm1.

input-96-predict-336 setting
MSE Reduction

- *ETT : 74%*
- *Electricity : 18%*
- *Exchange : 61%*
- *Traffic : 15%*
- *Weather : 21%*

38% Overal settings

Experiments : Univariate

Models		**Autoformer**		N-BEATS[29]		Informer[48]		LogTrans[26]		Reformer[23]		DeepAR[34]		Prophet[39]		ARIMA[1]	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT	96	**0.065**	**0.189**	0.082	0.219	0.088	0.225	0.082	0.217	0.131	0.288	0.099	0.237	0.287	0.456	0.211	0.362
	192	**0.118**	**0.256**	0.120	0.268	0.132	0.283	0.133	0.284	0.186	0.354	0.154	0.310	0.312	0.483	0.261	0.406
	336	**0.154**	**0.305**	0.226	0.370	0.180	0.336	0.201	0.361	0.220	0.381	0.277	0.428	0.331	0.474	0.317	0.448
	720	**0.182**	**0.335**	0.188	0.338	0.300	0.435	0.268	0.407	0.267	0.430	0.332	0.468	0.534	0.593	0.366	0.487
Exchange	96	0.241	0.387	0.156	0.299	0.591	0.615	0.279	0.441	1.327	0.944	0.417	0.515	0.828	0.762	**0.112**	**0.245**
	192	**0.273**	**0.403**	0.669	0.665	1.183	0.912	1.950	1.048	1.258	0.924	0.813	0.735	0.909	0.974	0.304	0.404
	336	**0.508**	**0.539**	0.611	0.605	1.367	0.984	2.438	1.262	2.179	1.296	1.331	0.962	1.304	0.988	0.736	0.598
	720	**0.991**	**0.768**	1.111	0.860	1.872	1.072	2.010	1.247	1.280	0.953	1.894	1.181	3.238	1.566	1.871	0.935

input-96-predict-336 setting

MSE Reduction

- *ETT : 14%*
- *Exchange : 17%*

Experiments : Case Study

- **Dataset :** UCI household electricity data[4]
- **Time Period :** January 2008 – June 2008 (~4,320 hourly samples)
- **Preprocessing :** Dropping missing values, Scaling values with MinMaxScaler, Focusing on Global Active Power
- **Model Architecture :** Autocorrelation via FFT, Time Delay Aggregation, Decomposition Layer
- **Training and Evaluation** : 80% training / 20% test Split, Adam Optimizer, MSE, Loss Function

Conclusion

Pros	Cons
Captures both short- and long-term dependencies efficiently	More complex architecture thus higher training cost

Thank you

Q & A

[1] Wu, H., Zhang, Y., Chen, J., & Zhou, Y. (2021). Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In Advances in Neural Information Processing Systems (NeurIPS 2021)

[2] Simhayev, E., Rasul, K., & Rogge, N. (2023, June 16). Yes, transformers are effective for time series forecasting (+ Autoformer). Hugging Face Blog.

[3] Wang : Literature Review 22: A Paper on Long-Term Time Series Prediction [Video]

[4] Individual Household Electric Power Consumption - UCI Machine Learning Repository