TIMER-XL: LONG-CONTEXT TRANSFORMERS FOR UNIFIED TIME SERIES FORECASTING

Yong Liu,* Guo Qin,* Xiangdong Huang, Jianmin Wang, Mingsheng Long✉
School of Software, BNRist, Tsinghua University, Beijing 100084, China
{liuyong21,qinguo24}@mails.tsinghua.edu.cn
{huangxdong,jimwang,mingsheng}@tsinghua.edu.cn

Yong Liu

Guo Qin

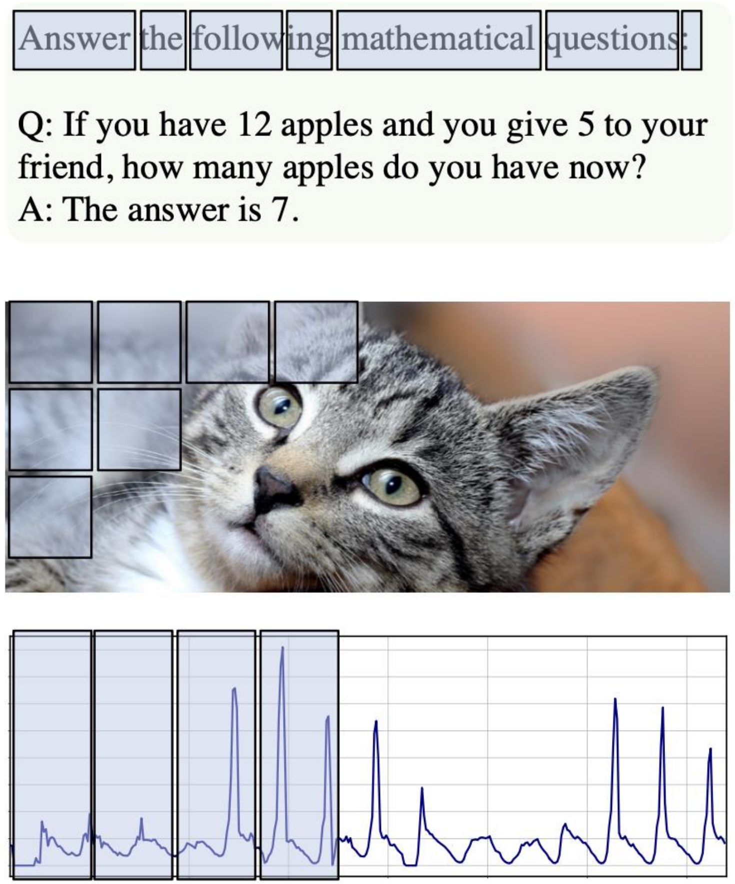
Xiangdong Huang

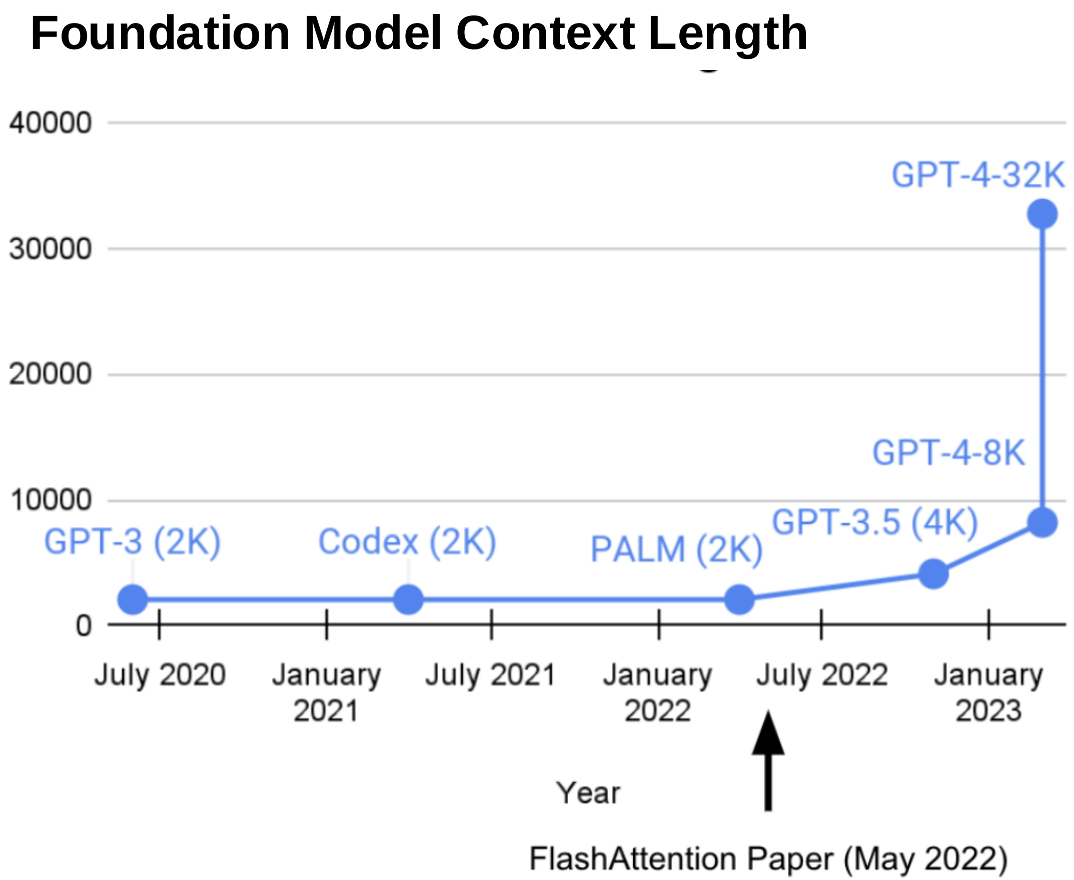
Jianmin Wang

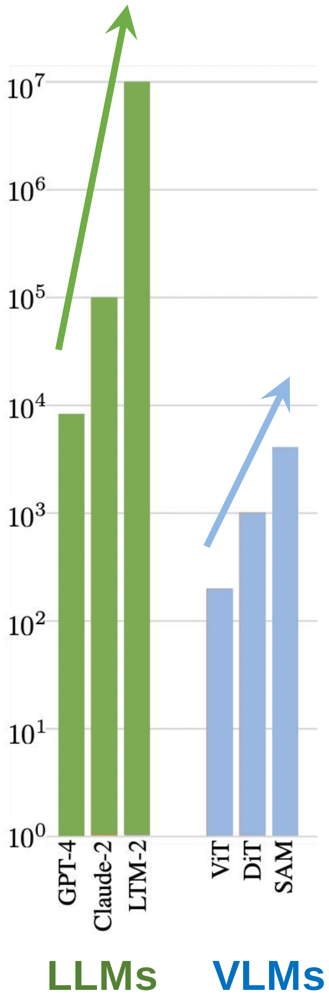
Mingsheng Long

Context Length Matters

Context Length of Foundation Models is Scaling

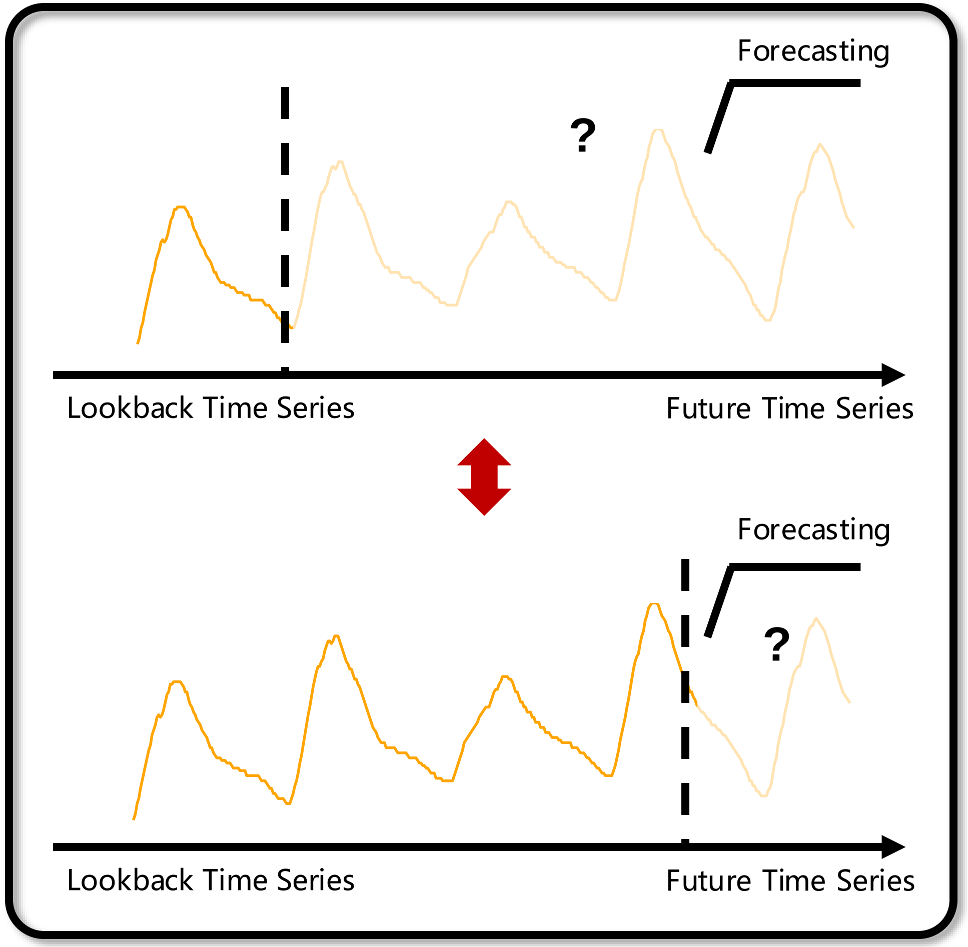






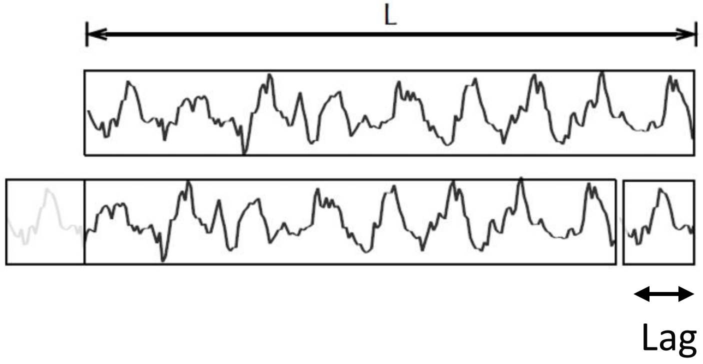
Long-Context Forecasting

Long-Term Forecasting -> Long-Context Forecasting

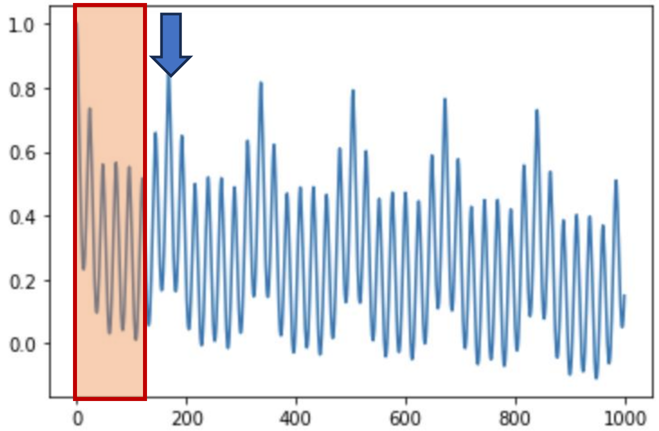


$$\mathcal{R}_{\mathcal{XX}}(\tau) = \lim_{L\to\infty} \frac{1}{L}\sum_{t=1}^{L}\mathcal{X}_t\mathcal{X}_{t-\tau}.$$

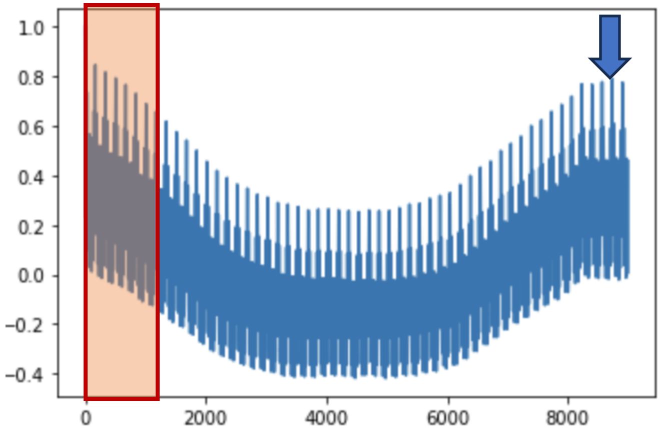
ACF indicates Periodicity



Weekly Period



Yearly Period

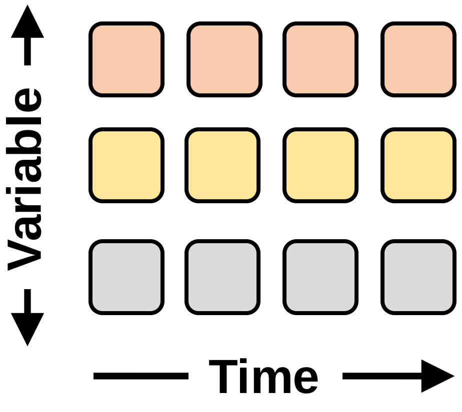


INFORMATION INCOMPLETE

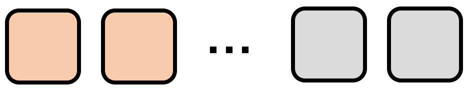
Unified Time Series Forecasting

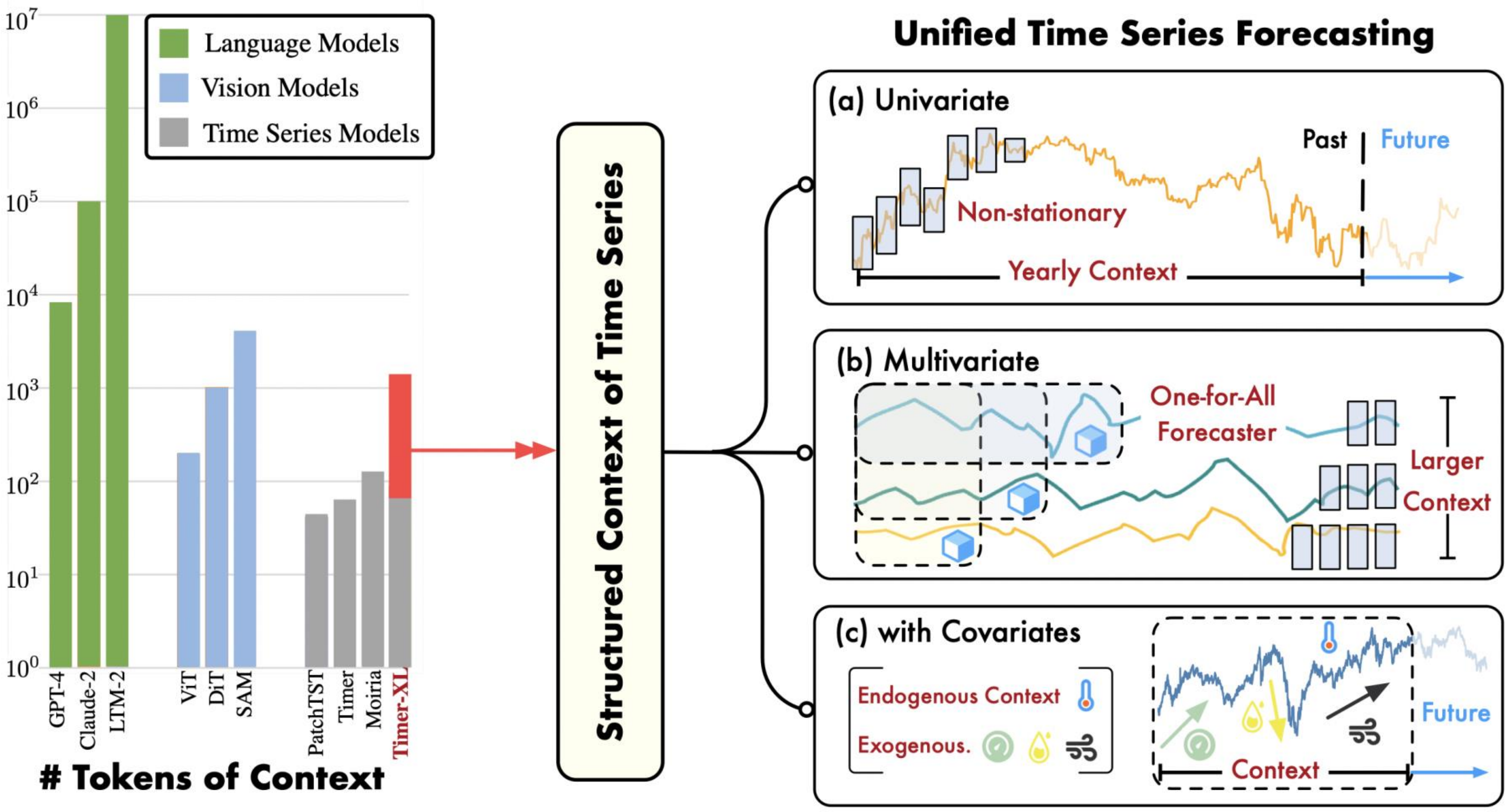
Long-Context Forecasting -> Unified Time Series Forecasting

2D Time Series



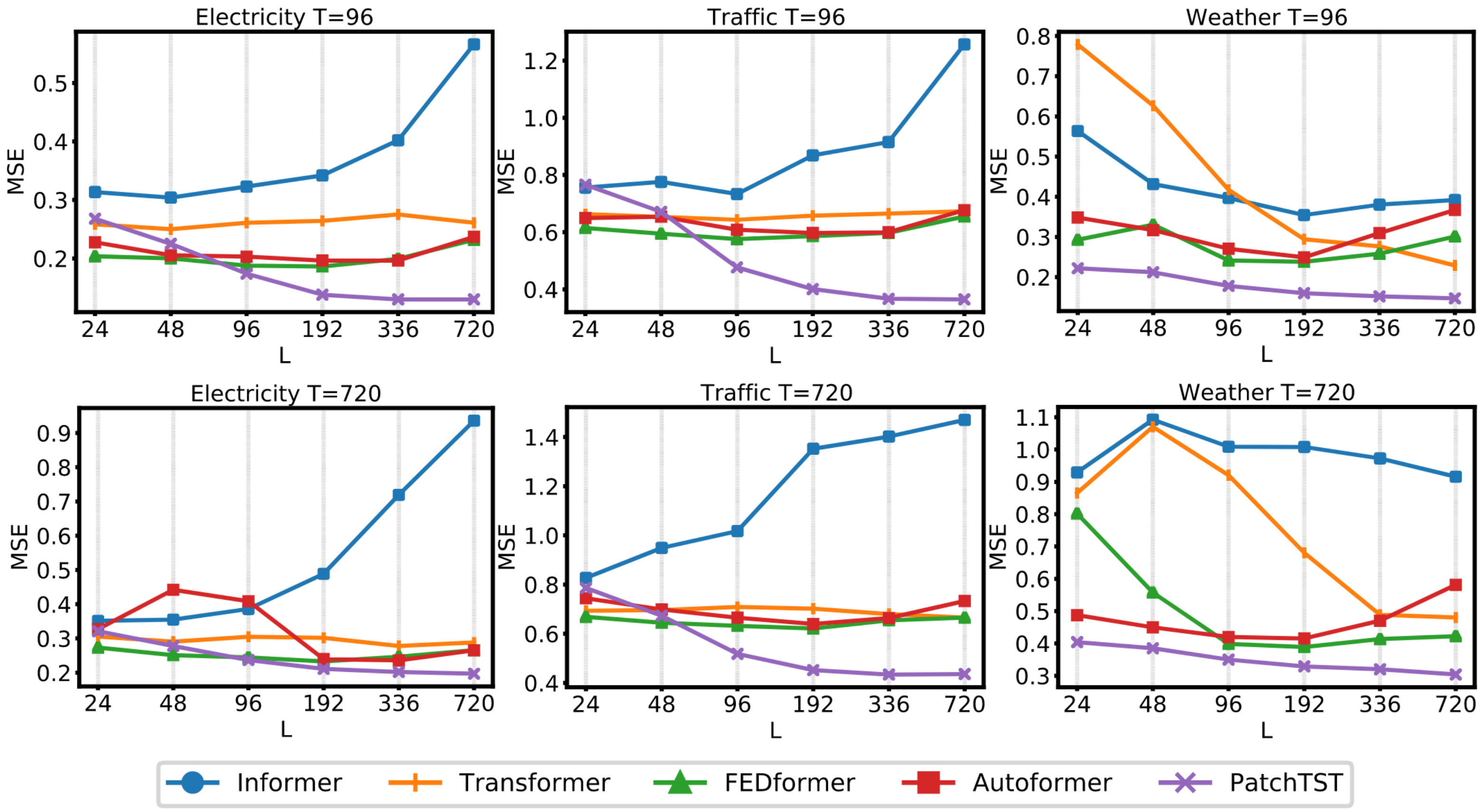
Overlength Context





Rethinking Long-Context Transformers

How Long Should be Inputted? Is Longer Context Better?



Performance (MSE) - Context Length (L)

Tokenization

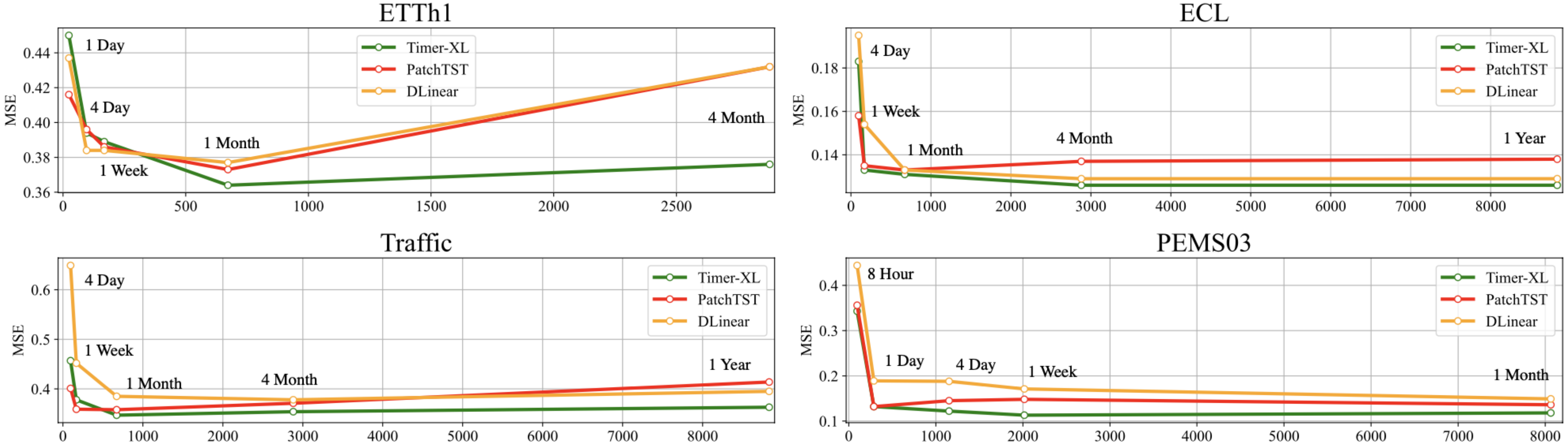
- Point-Level
- ✓ **Patch-Level**

Prediction Length

- Long-Term
- ✓ **Short-Term**

A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *ICLR 2023.*

Rethinking Long-Context Transformers

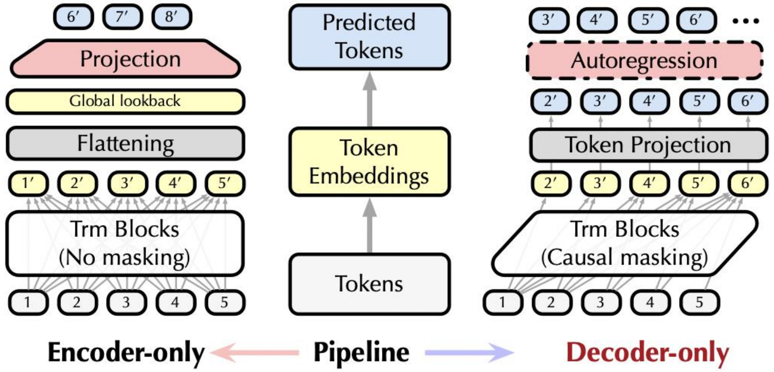


Performance - Context Length

Models	Timer-XL	PatchTST	DLinear
Metric	MSE | MAE	MSE | MAE	MSE | MAE
Lookback-8 (1 Day)	0.0847 | 0.2100	0.0897 | 0.2196	0.0970 | 0.2276
Lookback-32 (4 Day)	0.0713 | 0.1928	0.0778 | 0.2080	0.0841 | 0.2113
Lookback-56 (1 Week)	0.0688 | 0.1891	0.0785 | 0.2082	0.0814 | 0.2081
Lookback-224 (1 Month)	0.0675 | 0.1868	0.0745 | 0.2042	0.0788 | 0.2048
Lookback-960 (4 Month)	0.0667 | 0.1863	0.1194 | 0.2696	0.0773 | 0.2031
Lookback-2944 (1 Year)	0.0663 | 0.1857	0.1109 | 0.2638	0.0763 | 0.2024

Architecture

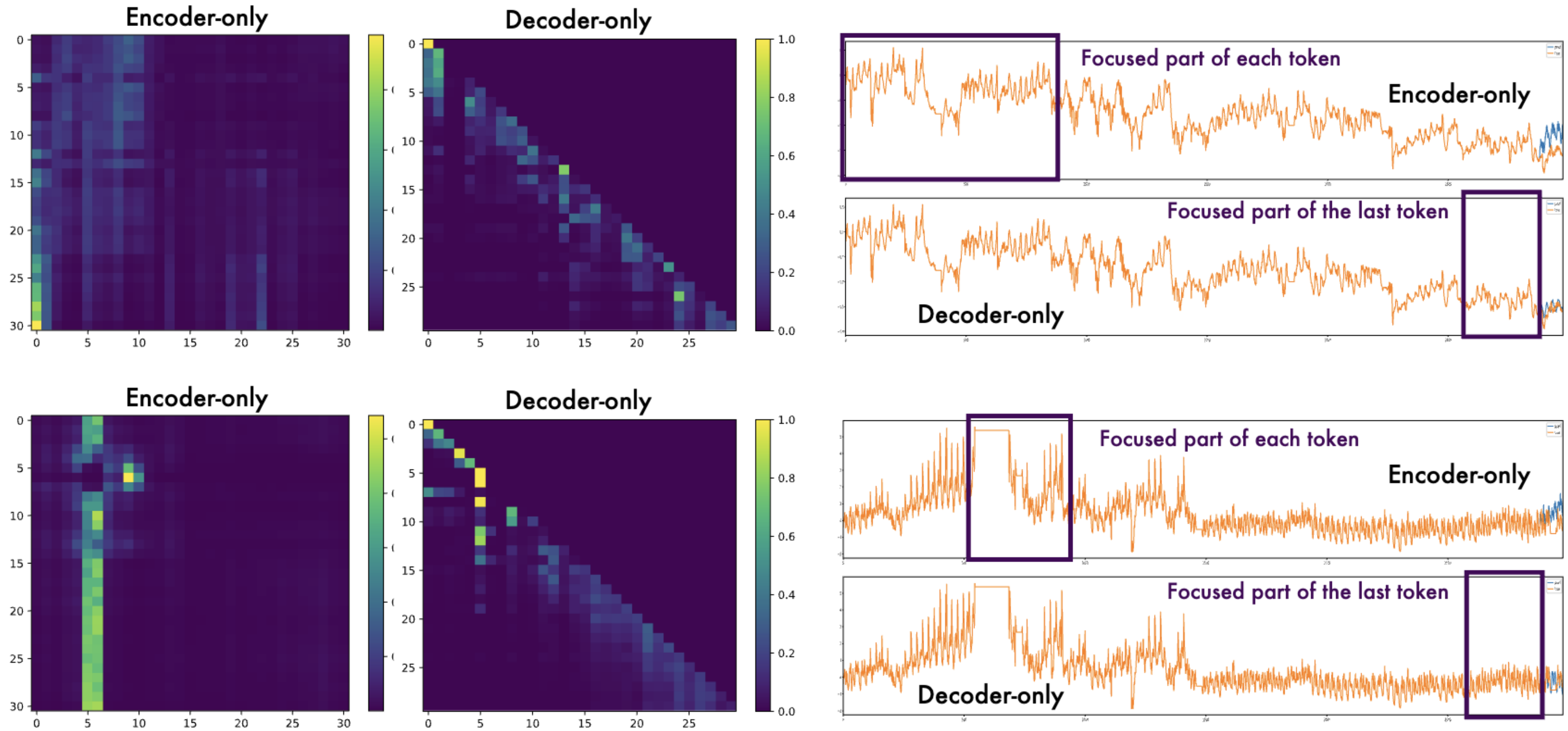
- Encoder-Only
- ✓ **Decoder-Only**



Decoder-Only Transformers Outperform Encoder-Only Models on Long-Context Sequences

Rethinking Long-Context Transformers

Attention - Raw Time Series



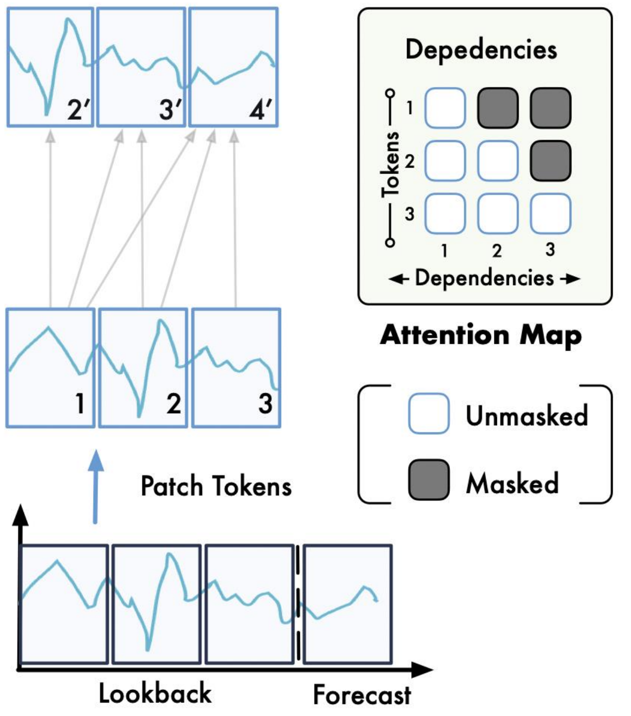
Decoder-Only Transformers Can Selectively Focus on Long-Context Sequences

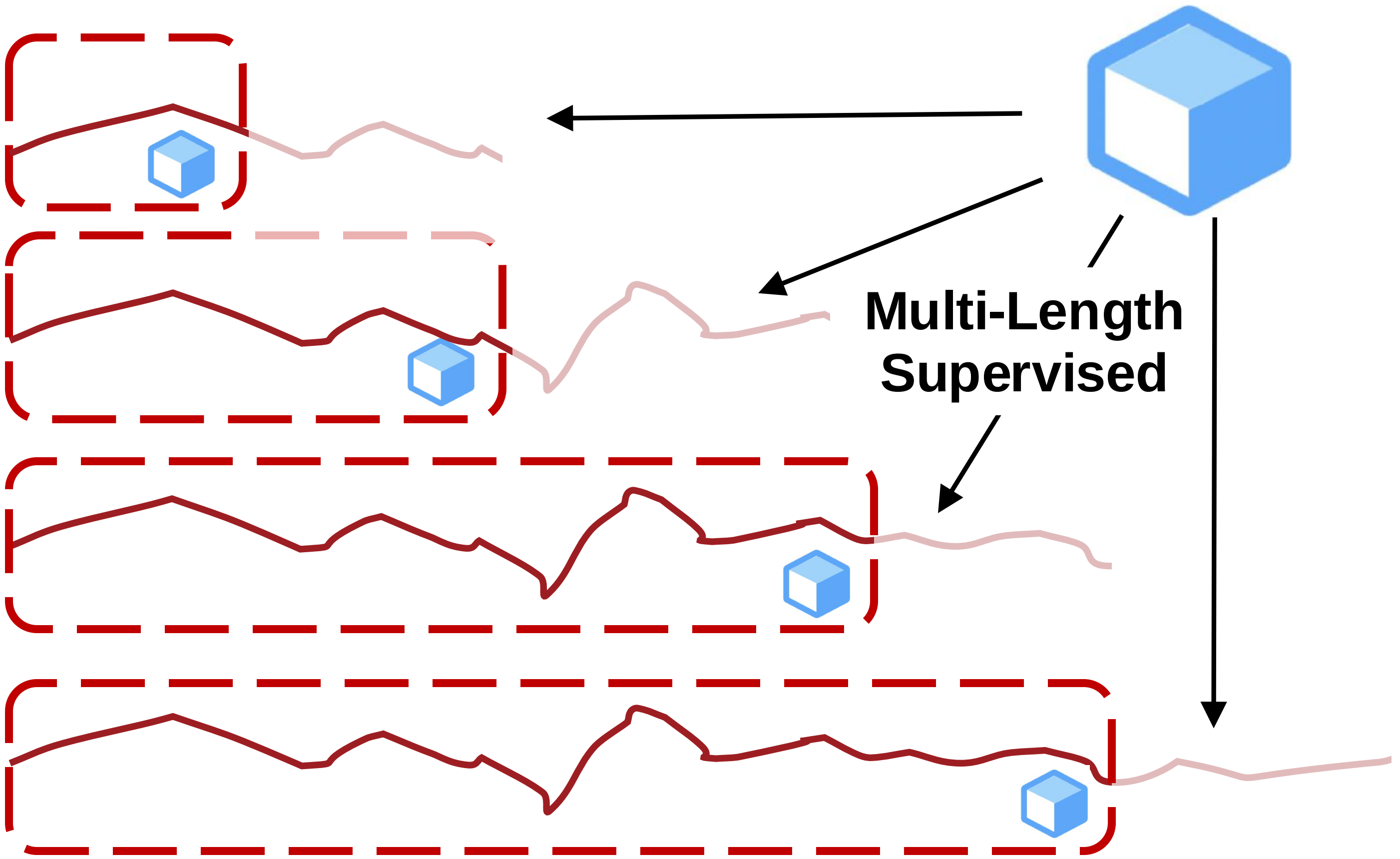
Extending 1D Sequences to 2D Time Series

Next Token Prediction (Patch Tokenization) $\mathbf{x}_i = \{x_{(i-1)P+1}, \dots, x_{iP}\}$

$$P(\mathbf{X}) = \prod_{i=1}^{T} p(\mathbf{x}_{i+1} | \mathbf{x}_{\leq i})$$

(a) Univariate





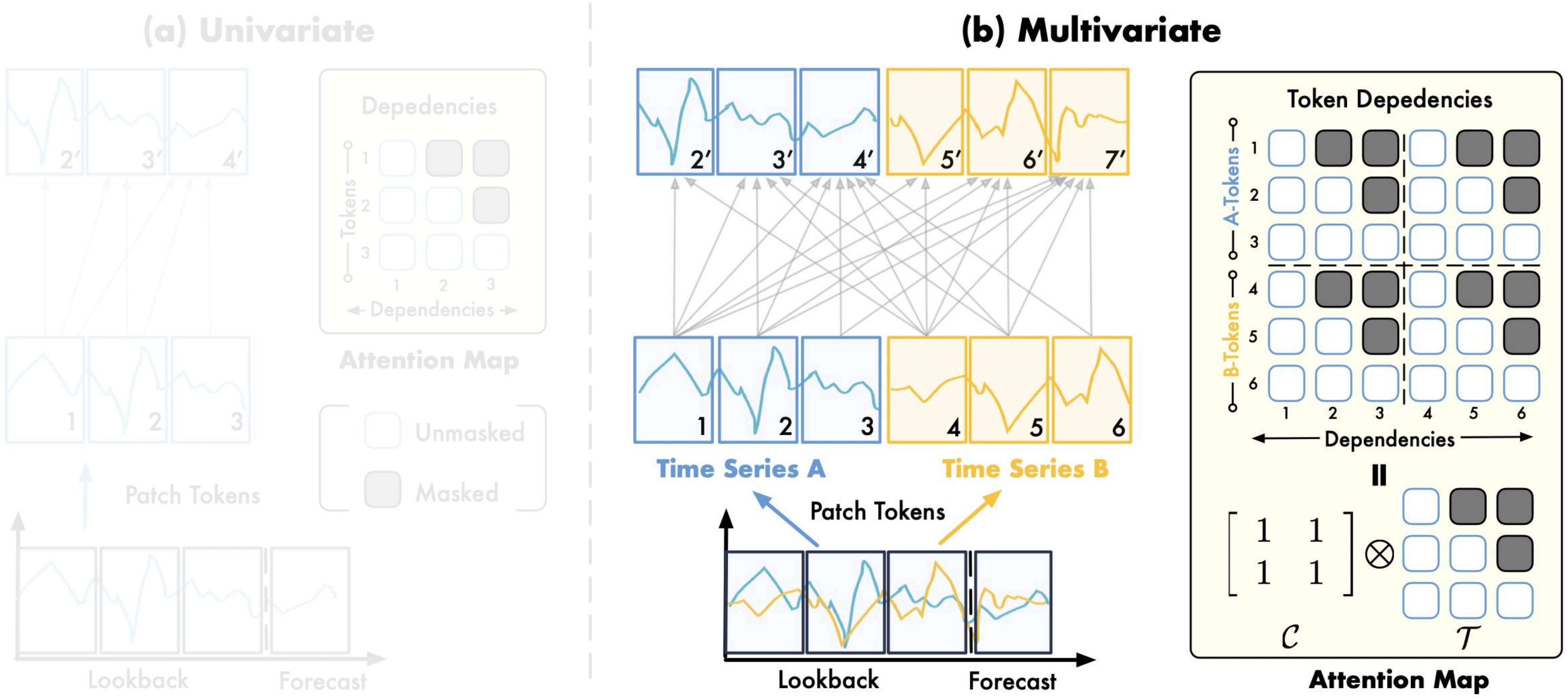
Decoder-Only Transformers Are One-For-All-Length Models

Extending 1D Sequences to 2D Time Series

Next Token Prediction -> Multivariate Next Token Prediction

$$P(\mathbf{X}) = \prod_{i=1}^{T} p(\mathbf{x}_{i+1} | \mathbf{x}_{\le i})$$

$$P(\mathbf{X}) = \prod_{m=1}^{N} \prod_{i=1}^{T} p(\mathbf{x}_{m,i+1} | \mathbf{x}_{:,\le i}) \qquad \mathbf{x}_{m,i} = \{\mathbf{X}_{m,(i-1)P+1}, \dots, \mathbf{X}_{m,iP}\}$$



Kronecker Product

- Temporal Causality

$$\mathcal{T}_{i,j} = \begin{cases} 1 & \text{if } j \le i, \\ 0 & \text{otherwise.} \end{cases}$$

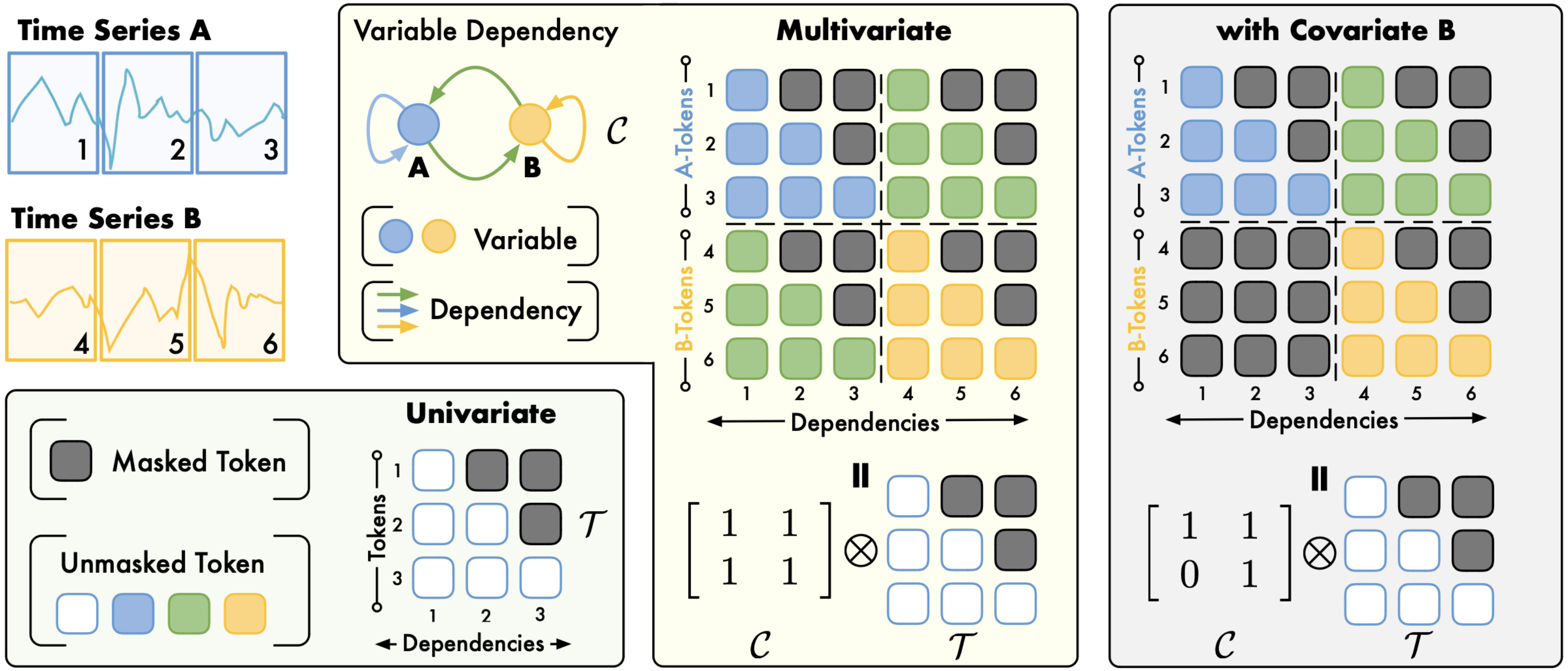
- Variable Dependence

$$\mathcal{C}_{m,n} = \begin{cases} 1 & \text{if variable } m \text{ is dependent on } n, \\ 0 & \text{otherwise.} \end{cases}$$

TimeAttention

A Versatile Masking Mechanism for Multidimensional Time Series

$$\mathrm{TimeAttention}(\mathbf{H}) = \mathrm{Softmax}\left(\frac{\mathrm{Mask}(\mathcal{C} \otimes \mathcal{T}) + \mathcal{A}}{\sqrt{d_k}}\right)\mathbf{H}\mathbf{W}_v,\ \mathrm{Mask}(\mathcal{M}) = \begin{cases} 0 & \text{if } \mathcal{M}_{i,j} = 1, \\ -\infty & \text{if } \mathcal{M}_{i,j} = 0. \end{cases}$$



Kronecker Product

- Temporal Causality

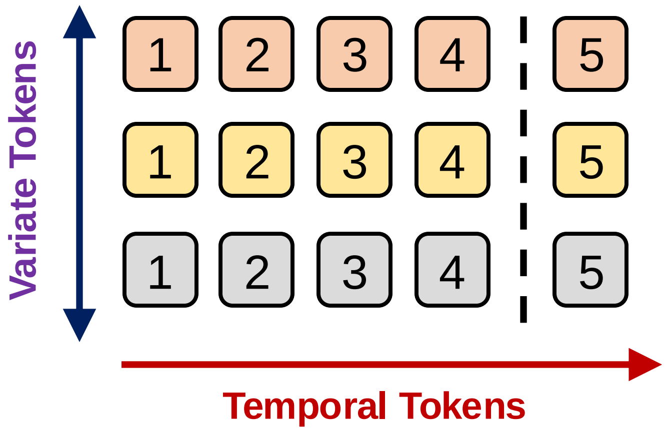
$$\mathcal{T}_{i,j} = \begin{cases} 1 & \text{if } j \leq i, \\ 0 & \text{otherwise.} \end{cases}$$

- Variable Dependence

$$\mathcal{C}_{m,n} = \begin{cases} 1 & \text{if variable } m \text{ is dependent on } n, \\ 0 & \text{otherwise.} \end{cases}$$

Position Embedding in Self-Attention

$$\mathcal{A}_{mn,ij} = \underbrace{\mathbf{h}_{m,i}^{\top}\mathbf{W}_q\mathbf{R}_{\theta,i-j}\mathbf{W}_k^{\top}\mathbf{h}_{n,j}}_{\text{RoPE}} + \underbrace{u \cdot \mathbb{1}(m = n) + v \cdot \mathbb{1}(m \neq n)}_{\text{Alibi}}$$



Tokens of multivariate time series are both **temporal tokens** and **variate tokens**

Permutation-Invariant

$\mathcal{H} : \mathbb{R}^T \to \mathbb{R}$

$\mathcal{H}(x_1, \dots, x_T) = \mathcal{H}(\pi\{x_1, \dots, x_T\})$

π : permutation of temporal tokens

RoPE: Avoid PI (inherent in self-attention) on the Temporal dimension

Permutation-Equivalent

$\mathcal{H} : \mathbb{R}^N \to \mathbb{R}^N$

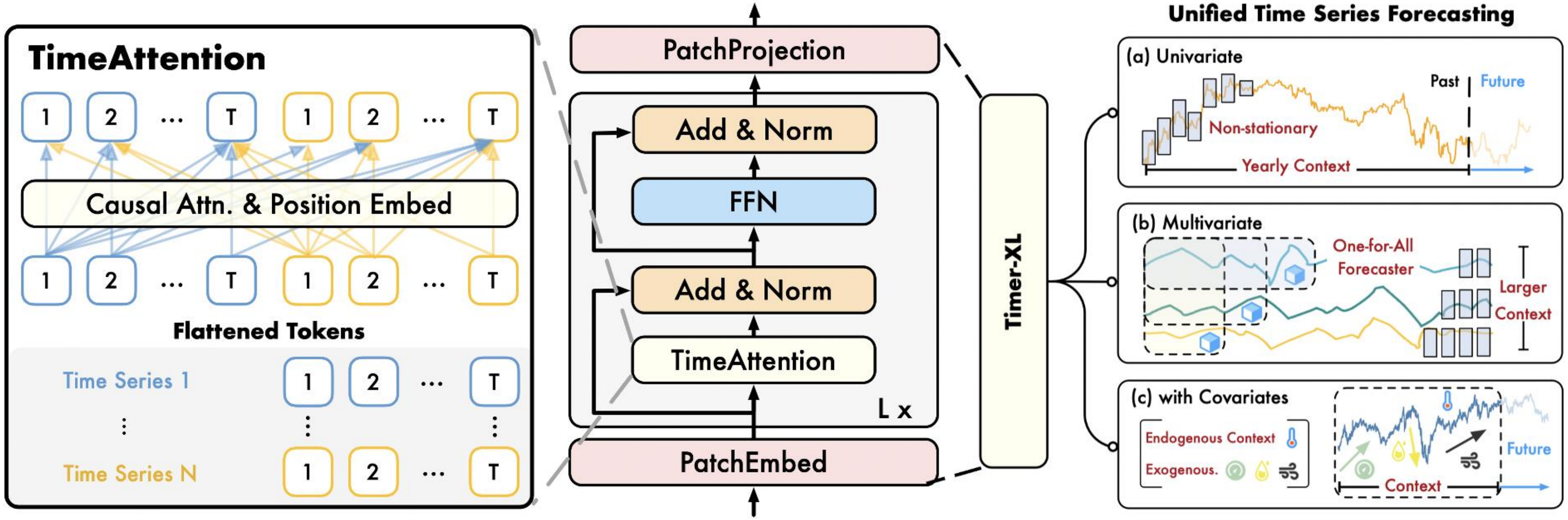
$\pi\{\mathcal{H}(x_1, \dots, x_N)\} = \mathcal{H}(\pi\{x_1, \dots, x_N\})$

π : permutation of variate tokens

Learnable Alibi: Maintain PE on the Variate dimension (only distinguish endo-/exo-variates)

Timer-XL

A Decoder-Only Long-Context Transformer for Unified Forecasting



Unified

Context

Timer-XL can be used for (1) task-specific training and (2) scalable pre-training, handling arbitrary-length and any-variable time series

Timer-XL

A Decoder-Only Long-Context Transformer for Unified Forecasting

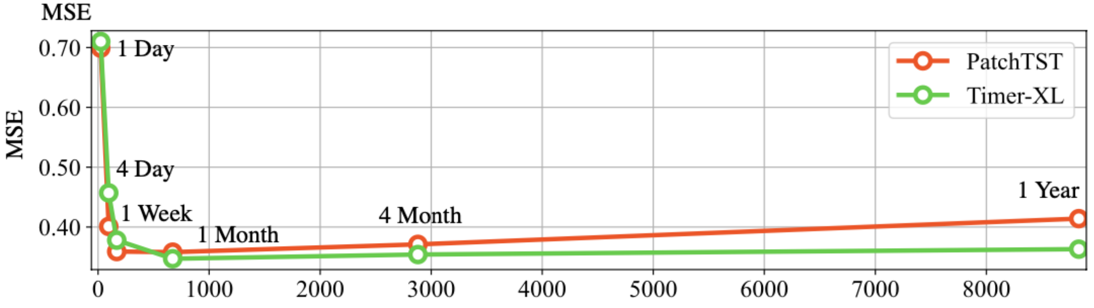
Table 1: Comparison among representative time-series Transformers.

Model	PatchTST (2022)	iTrans. (2023)	TimeXer (2024b)	UniTST (2024a)	Moirai (2024)	Timer (2024c)	**Timer-XL (Ours)**
Intra-Series	✓	✗	✓	✓	✓	✓	✓
Inter-Series	✗	✓	✓	✓	✓	✗	✓
Causal Trm.	✗	✗	✗	✗	✗	✓	✓
Pre-Trained	✗	✗	✗	✗	✓	✓	✓

Timer-XL can be used for (1) task-specific training and (2) scalable pre-training, handling arbitrary-length and any-variable time series

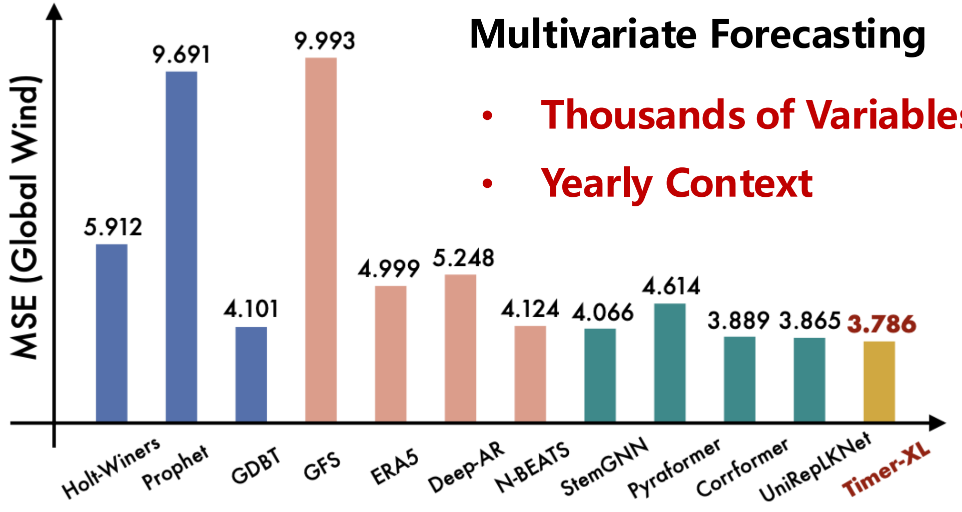
Supervised Training Performance

Univariate Forecasting (Non-Stationary)



Multivariate Forecasting

- Thousands of Variables
- Yearly Context



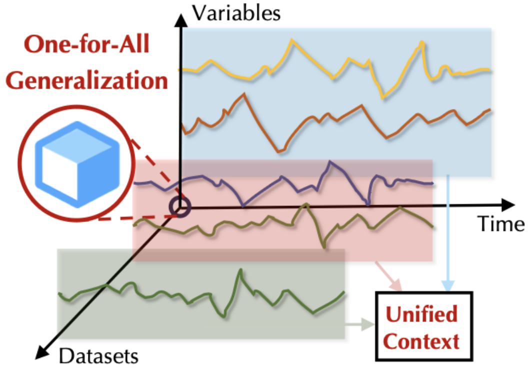
Forecasting with Covariates

Models	Timer-XL (Ours)		Timer-XL (Noncausal)		TimeXer (2024b)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE
NP	0.234	0.262	0.237	0.265	0.238	0.268
PJM	0.089	0.187	0.092	0.188	0.088	0.188
BE	0.371	0.243	0.410	0.279	0.379	0.243
FR	0.381	0.204	0.406	0.220	0.384	0.208
DE	0.434	0.415	0.435	0.415	0.440	0.418
Average	0.302	0.262	0.316	0.273	0.306	0.265

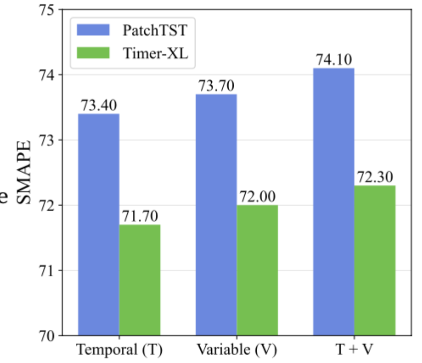
Outperform Task-Specific Models

Timer-XL

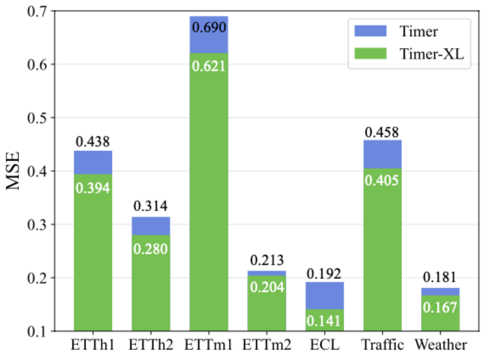
Large-Scale Pre-Training & Zero-Shot Forecasting



(a) ERA5 (Pred-96, 4920 Stations)



(b) TSLib (Pred-96, UTSD Pre-Trained)



Pre-Training Large Time-Series Model

Zero-Shot Forecasting (Pre-trained on 260B Time Points)

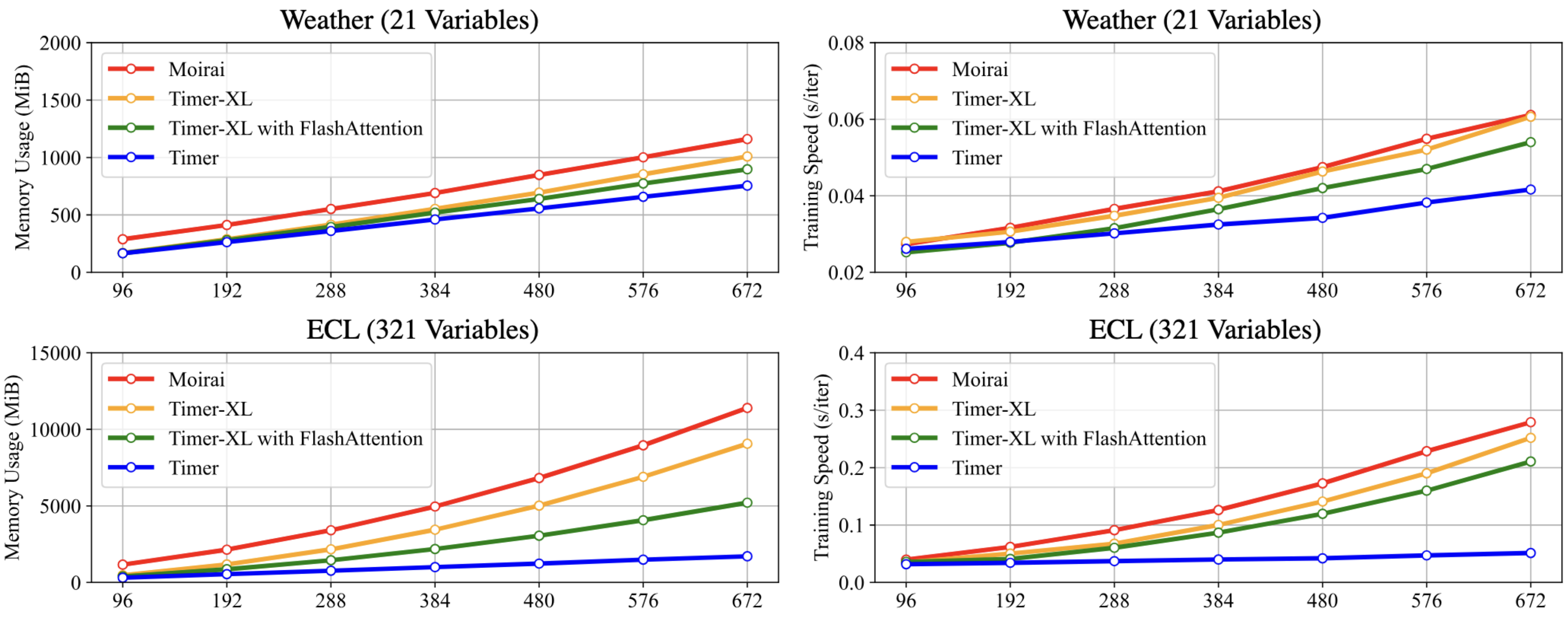
Table 7: Averaged results of zero-shot forecasting. Full results of all prediction lengths are provided in Table 13. 1st Count represents the number of wins achieved by a model under all prediction lengths and datasets. The configuration of **Timer-XL**$_{Base}$ shown in Table 11 is comparable with **Moirai**$_{Base}$, which is pre-trained on UTSD (Liu et al., 2024c) and LOTSA (Woo et al., 2024).

Models	**Timer-XL**$_{Base}$ **(Ours)**		**Time-MoE**$_{Base}$ (2024)		**Time-MoE**$_{Large}$ (2024)		**Time-MoE**$_{Ultra}$ (2024)		**Moirai**$_{Small}$ (2024)		**Moirai**$_{Base}$ (2024)		**Moirai**$_{Large}$ (2024)		**TimesFM** (2023)		**MOMENT** (2024)		**Chronos**$_{Base}$ (2024)		**Chronos**$_{Large}$ (2024)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	0.373	0.392	0.394	0.415	0.376	0.405	**0.356**	0.391	0.436	0.410	0.406	**0.385**	0.422	0.391	0.433	0.418	0.670	0.536	0.645	0.500	0.555	0.465
ETTm2	**0.273**	**0.336**	0.317	0.365	0.316	0.361	0.288	0.344	0.307	0.347	0.311	0.337	0.329	0.343	0.328	0.346	0.316	0.365	0.310	0.350	0.295	0.338
ETTh1	0.404	**0.417**	0.400	0.424	**0.394**	0.419	0.412	0.426	0.428	0.427	0.417	0.419	0.480	0.439	0.473	0.443	0.683	0.566	0.591	0.468	0.588	0.466
ETTh2	**0.347**	0.388	0.366	0.404	0.405	0.415	0.371	0.399	0.361	0.384	0.362	0.382	0.367	**0.377**	0.392	0.406	0.361	0.409	0.405	0.410	0.455	0.427
ECL	**0.174**	0.278	-	-	-	-	-	-	0.218	0.303	0.187	0.274	0.186	**0.270**	-	-	0.765	0.686	0.214	0.278	0.204	0.273
Weather	**0.256**	0.294	0.265	0.297	0.270	0.300	**0.256**	0.288	0.275	0.286	0.287	0.281	0.264	**0.273**	-	-	0.294	0.326	0.292	0.315	0.279	0.306
1st Count	**15**	**10**	2	1	3	0	10	7	0	0	0	5	1	**10**	0	1	2	0	0	0	0	2

The model checkpoint is available at: https://huggingface.co/thuml/timer-base-84m.

Model Efficiency

Evaluating Memory/FLOPS of Time-Series Transformers



Efficiency - Context Length

Model Efficiency

Computational Complexity of Time-Series Transformer

- **FFN**: Linear growth with the context length - $O(NT)$ ***Dominate Term in TS!***
- **Attention**: Quadratic growth with the context length - $O(N^2T^2)$

Table 8: Parameters count and computational complexity of Transformers for multivariate time series.

Metric	Type	Count	Complexity
FLOPs (Training Speed)	Channel Independence	$12\big(PDNT + L(D+H)NT^2 + (2+\alpha)LD^2NT\big)$	$\mathcal{O}\big(LDNT(D+T)\big)$
	Channel Dependence	$12\big(PDNT + L(D+H)N^2T^2 + (2+\alpha)LD^2NT\big)$	$\mathcal{O}\big(LDNT(D+NT)\big)$
Parameters	Encoder-Only	$(4+2\alpha)LD^2 + 4LD + (1+T)PD$	$\mathcal{O}(LD^2)$
	Decoder-Only	$(4+2\alpha)LD^2 + 4LD + 2PD$	$\mathcal{O}(LD^2)$
Memory Footprint	Self-Attention	$4(D+P)NT + (32+8\alpha)LDNT + 4LHN^2T^2$	$\mathcal{O}\big(LHN^2T^2\big)$
	FlashAttention	$4(D+P)NT + (32+8\alpha)LDNT$	$\mathcal{O}\big(LDNT\big)$

$*$ L is the block number of Transformers. D is the dimension of embeddings (the hidden dimension of FFN D_{ff} is set as αD). H is the head number and the dimension of query, key, and value $d_k = D/H$. The overhead is to train on a multivariate time series (N-variables and TP time points) with patch token length P and context length T. Set $N = 1$ for training on univariate time series.

Model Analysis

Non-stationary Forecasting

Table 16: Evaluations (672-pred-96) on the effect of ReVIN (Kim et al., 2021) on Transformers.

Models	Timer-XL with ReVIN	Timer-XL w/o ReVIN	PatchTST with ReVIN	PatchTST w/o ReVIN
Metric	MSE | MAE	MSE | MAE	MSE | MAE	MSE | MAE
ETTh1	0.364 | 0.397	0.370 | 0.401	0.370 | 0.399	0.421 | 0.448
Weather	0.157 | 0.205	0.151 | 0.205	0.149 | 0.198	0.173 | 0.242
ECL	0.127 | 0.219	0.130 | 0.225	0.129 | 0.222	0.138 | 0.244

Small Gap

Big Gap

- Long-context Transformers do not rely on Stationarization

Ablation Study

Table 14: Embedding ablation in TimeAttention. For the temporal dimension, we compare prevalent relative and absolute position embeddings. As for the variable dimension, we explore the effectiveness of the variable embedding that distinguishes endogenous and exogenous variables.

Design	Temporal	Variable	Traffic MSE	Traffic MAE	Weather MSE	Weather MAE	Solar-Energy MSE	Solar-Energy MAE	ERA5-MS MSE	ERA5-MS MAE
Timer-XL	**RoPE** (2024)	**with**	**0.340**	**0.238**	**0.157**	**0.205**	**0.162**	0.221	**0.164**	0.307
Replace	ALiBi (2021)	with	0.351	0.246	0.162	0.212	0.188	0.210	0.167	0.308
	Relative (2020)	with	0.361	0.250	0.163	0.214	0.197	0.215	0.168	0.309
	Absolute (2017)	with	0.381	0.270	0.159	0.207	0.171	**0.204**	0.165	**0.306**
w/o	RoPE (2024)	w/o	0.361	0.254	0.171	0.217	0.181	0.221	0.235	0.373
	w/o	w/o	0.363	0.253	0.164	0.215	0.194	0.215	0.167	0.309

$$\mathcal{A}_{mn,ij} = \underbrace{\boxed{\mathbf{h}_{m,i}^\top \mathbf{W}_q \mathbf{R}_{\theta,i-j} \mathbf{W}_k^\top \mathbf{h}_{n,j}}}_{\text{Temporal}} + \underbrace{\boxed{u \cdot \mathbb{1}(m=n) + v \cdot \mathbb{1}(m \neq n)}}_{\text{Variable}}$$

- RoPE outperformers other counterparts

- It is helpful to distinguish endogenous and exogenous variables

Interpretability

Attention Map

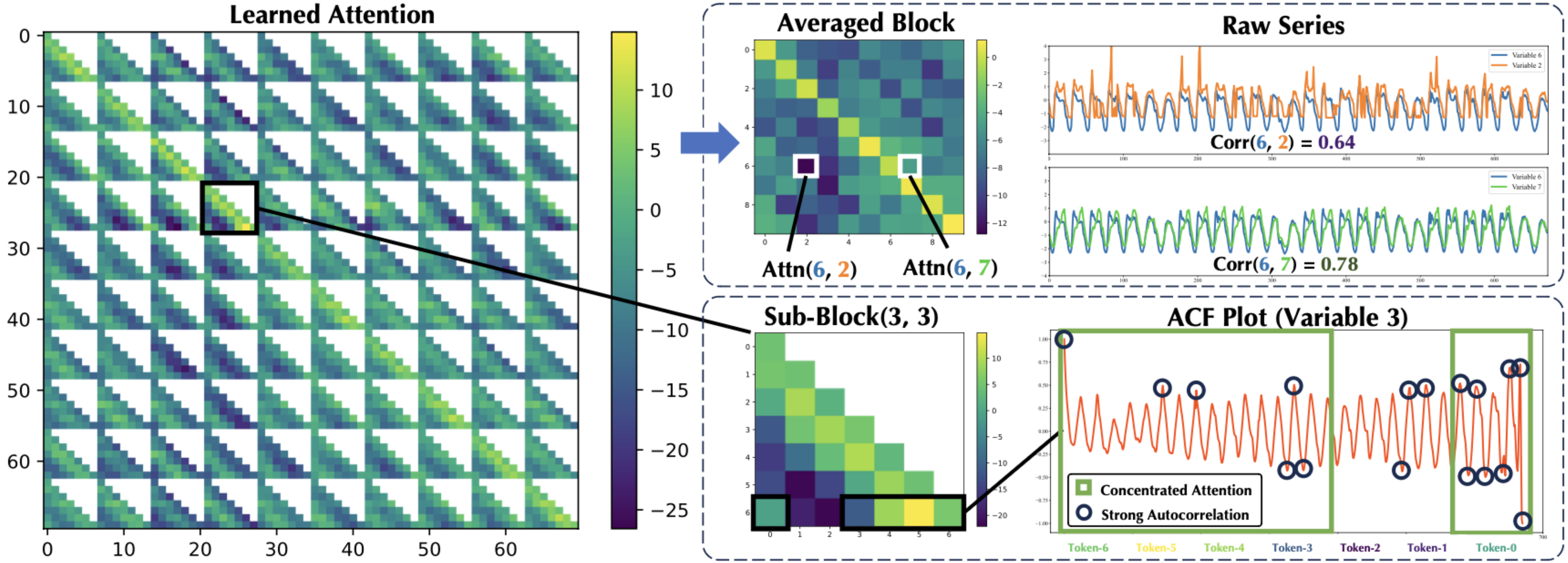


Figure 7: Visualization of TimeAttention. It is from the first sample of a length 672 in the test split of Traffic. We visualize the last 10 variables with each contains 7 tokens. We present auto-correlation function plot. Auto-correlation can be reflected by the distribution of attention scores (bottom right). We average TimeAttention across sub-blocks, which indicates Pearson correlations (upper right).

Thank You!

GitHub: https://github.com/thuml/Timer-XL