

GTM: A General Time-Series Model for Enhanced Representation Learning of Time-Series Data

Cheng HE^{1,2}, Xu Huang^{1,2}, Gangwei Jiang^{1,2}, Zhaoyi Li^{1,2}, Defu Lian^{1,2}, Hong Xie^{1,2}, Enhong Chen^{1,2}, Xijie Liang³, Zengrong Zheng⁴, Patrick P. C. Lee⁵

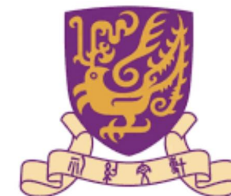
¹University of Science and Technology of China

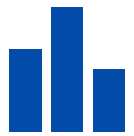
²State Key Laboratory of Cognitive Intelligence

³Shanghai Black Wing Asset Management

⁴Di-Matrix (Shanghai) Information Technology

⁵The Chinese University of Hong Kong





Background & Motivation

Key Challenges in building TSFMs

Limited expressiveness of TS data

- Scalar, temporally indexed sequences provide limited informational.

Heterogeneity of Downstream Tasks

- Current TSFMs require task-specific modifications at various levels.

General Time-Series Model

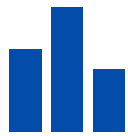
GTM Objectives

Enhanced Representation Learning

- Improve TS representation by exploring new dimensions.

Unified Pre-training Framework

- Establish the first generative-task-agnostic model for TS analysis.



Key Observations



Joint distributions of amplitude-frequency and phase-frequency differ significantly across time granularities.

Empirical Analysis

- A non-parametric estimation method: 2-D Kernel Density Estimation.

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^n K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right)$$

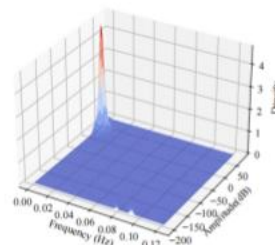
- 2-D Gaussian kernel function

$$K(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x - \mu_x)^2}{2\sigma_x^2} - \frac{(y - \mu_y)^2}{2\sigma_y^2}\right)$$

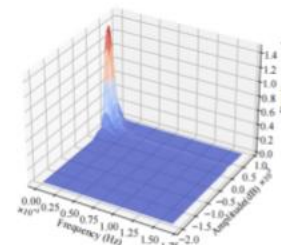
- 2-D Scott's rule as bandwidth function.

$$h_x = h_y = n^{-\frac{1}{6}} (\sigma_x \sigma_y)^{\frac{1}{2}}$$

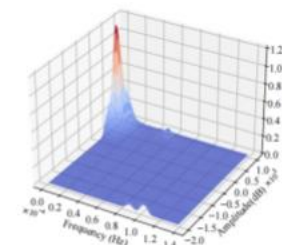
- Implication: Time granularity is a critical but unexplored dimension in TS representation learning.



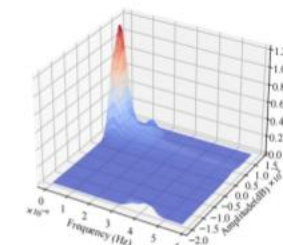
(a) 4-s amp



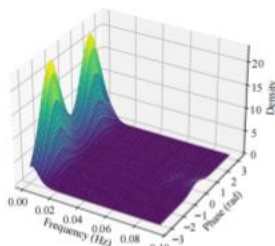
(b) 5-min amp



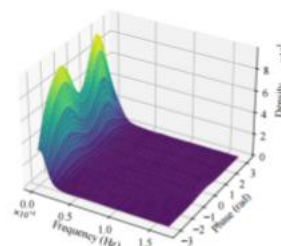
(c) 1-hour amp



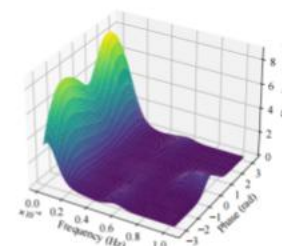
(d) 1-day amp



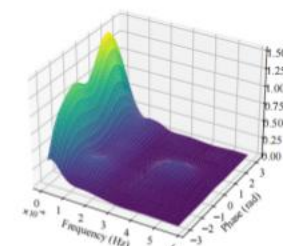
(e) 4-s phase



(f) 5-min phase



(g) 1-hour phase



(h) 1-day phase



GTM Architecture Overview

General Time Series Model

Input Embedding

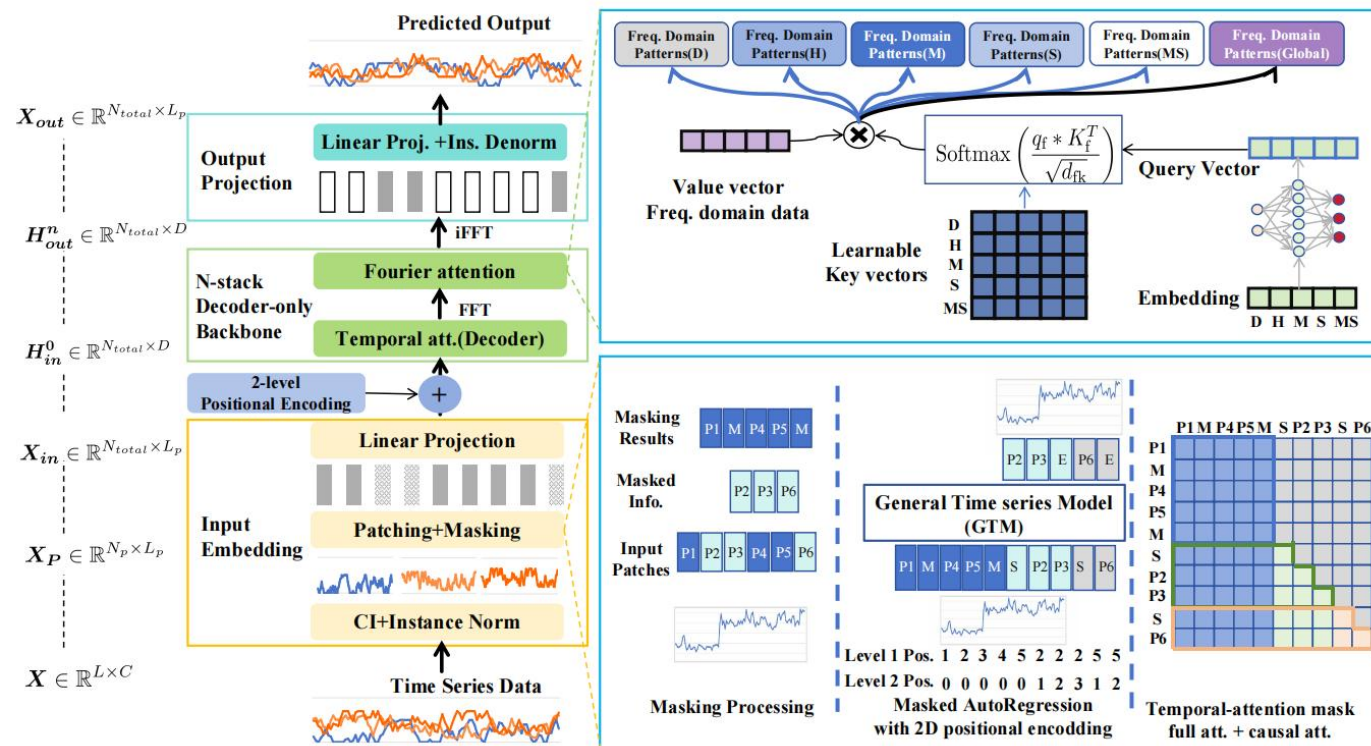
- Instance norm + Channel Independence (CI)
- Patching + Masking
- Linear embedding

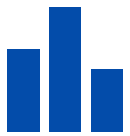
Backbone

- N-stack Decoder-only Transformer backbone.
- Attention Modules: Jointly models temporal and frequency patterns.

Output Projection

- A unified linear projection layer
- Followed by instance denormalization



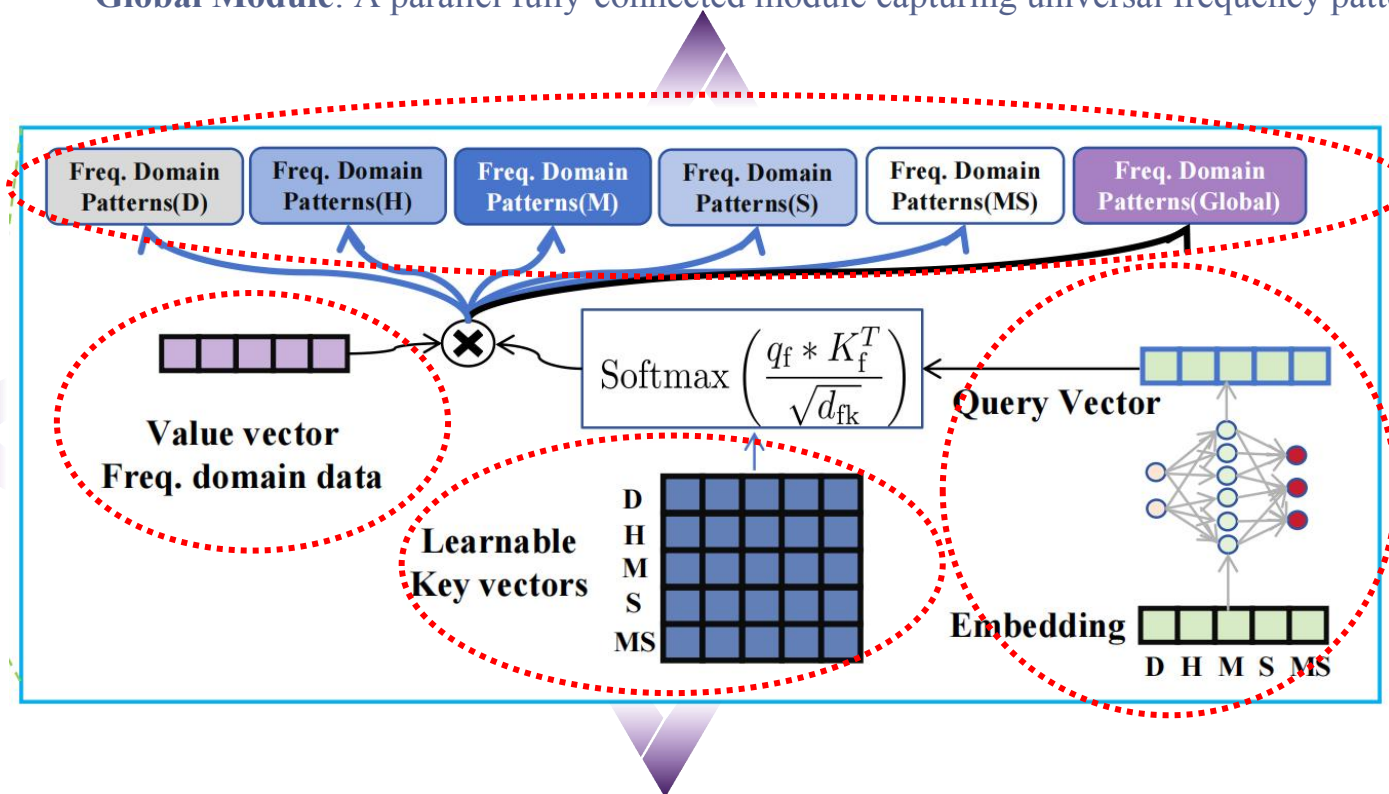


Fourier Attention Module

Granularity-aware Fourier Attention Mechanism

- **Granularity Experts:** Five low-rank modules to capture granularity-specific frequency patterns.
- **Global Module:** A parallel fully-connected module capturing universal frequency patterns.

- **Value Vector:** Frequency-domain features directly transformed from temporal representation via FFT.



- **Granularity Encoding:** Encode time granularity as a quintuple (e.g., [0, 0, 15, 0, 0] for 15-min).
- **Query Generation:** Map the quintuple into a query vector via linear embedding.

- **Key Embeddings:** Five learnable vectors, each strictly corresponding to a typical temporal granularity (Day, Hour, Min, Sec, MS).



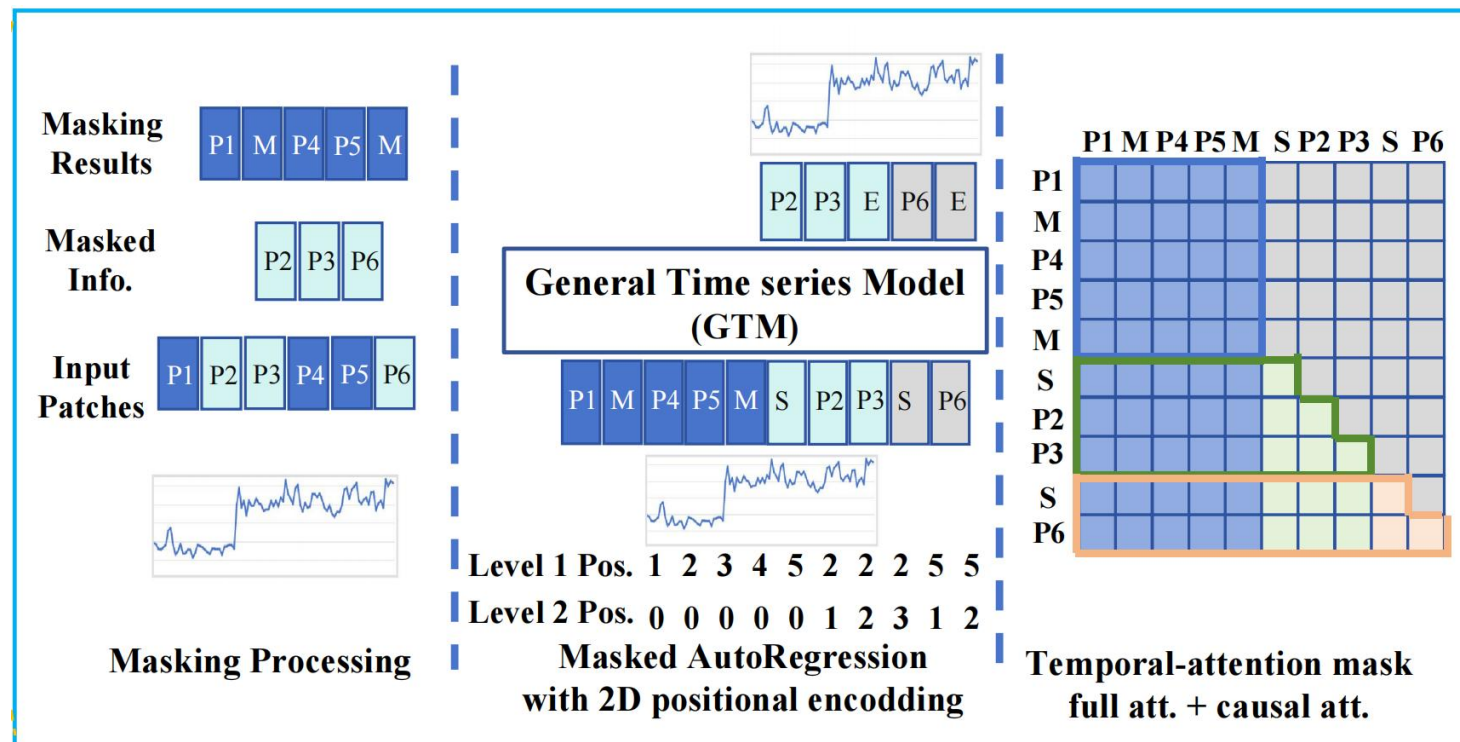
Unified Pre-training Framework

Hybrid Masking Strategy

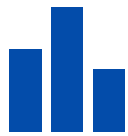
- Span shuffling
- Random and controlled consecutive tail masking
- 2D positional encoding

Hybrid Temporal Attention Masks

- Full attention for reconstruction
- Causal attention mask for autoregressive forecasting



GOAL: Unify mask reconstruction and autoregressive forecasting under a single objective.



Main Results: Generative Tasks

Long-term

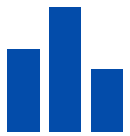
- Consistent SOTA model
- Directly utilize model for time series

Models		GTM	GPT4TS	TimesNet	PatchTST	DLinear	Fedformer	Informer	—
Dataset	Mask Ratio	MSE MAE	MSE MAE	MSE MAE	MSE MAE	MSE MAE	MSE MAE	MSE MAE	er IAE
ETTh1	12.5%	.034 .125	<u>.043</u> <u>.140</u>	.057 .159	.093 .201	.151 .267	.070 .190	.114 .234	713
	25%	.046 .143	<u>.054</u> <u>.156</u>	.069 .178	.107 .217	.180 .292	.106 .236	.140 .262	792
	37.5%	.059 .163	<u>.072</u> <u>.180</u>	.084 .196	.120 .230	.215 .318	.124 .258	.174 .293	809
	50%	.073 .179	<u>.107</u> <u>.216</u>	.102 .215	.141 .248	.257 .347	.165 .299	.215 .325	865
	Avg.	.053 .152	<u>.069</u> <u>.173</u>	.078 .187	.115 .224	.201 .306	.117 .246	.161 .279	795
	12.5%	.015 .082	<u>.017</u> <u>.085</u>	.023 .101	.041 .130	.080 .193	.052 .166	.063 .180	571

Models	GTM	UP2ME	GPT4TS	TimesNet	PatchTST	FEDformer	DLinear	Autoformer	Informer
Dataset	F1(%)	F1(%)	F1(%)	F1(%)	F1(%)	F1(%)	F1(%)	F1(%)	F1(%)
MSL	82.53	-	82.45	81.84	78.70	78.57	84.88	79.05	84.06
SMAP	77.57	-	72.88	69.39	68.82	70.76	69.26	71.12	69.92
SWaT	94.78	93.85	<u>94.23</u>	93.02	85.72	93.19	87.52	92.74	81.43
SMD	85.47	83.31	86.89	84.61	84.62	85.08	77.10	85.11	81.65
PSM	<u>95.43</u>	97.16	97.13	97.34	96.08	<u>97.23</u>	93.55	93.29	77.10
Average	87.01	-	<u>86.72</u>	85.24	82.79	84.97	82.46	84.26	78.83



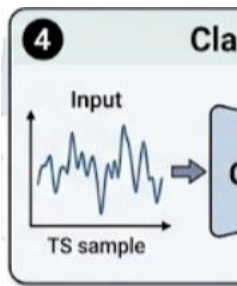
Electricity	25%	<u>.084</u> <u>.199</u>	<u>.087</u> <u>.203</u>	.089 .206	.065 .175	.118 .247	.120 .251	.042 .100	416		
	37.5%	<u>.090</u> <u>.206</u>	.094 .211	.094 .213	.076 .189	.144 .276	.136 .266	.049 .111	368		
	50%	<u>.096</u> <u>.215</u>	.101 .220	.100 .221	.091 .208	.175 .305	.158 .284	.053 .114	386		
	Avg.	<u>.086</u> <u>.202</u>	.090 .207	.092 .210	.072 .183	.132 .260	.130 .259	.045 .104	394		
	[Avg]	.161 <u>.254</u>	.167 .263	.184 .282	.158 -	<u>.159</u> .252	.192 .295	.166 .263	.207 .321	.214 .326	.311 .397



Main Results: More experiments

Classification

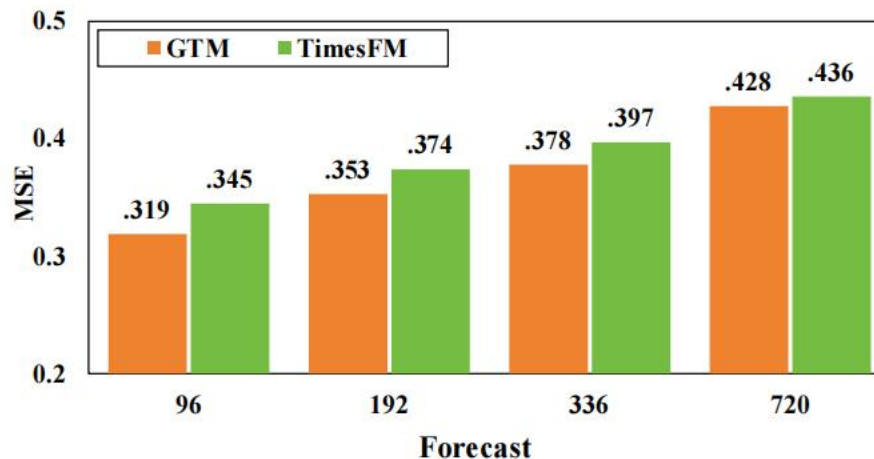
- GTM achieves the most best- and second-best results compared to TSFMs.



Dat	Dataset
Ethano	ETTh1
Fac	ETTm1
Ha	Weather
I	Electricity
Japa	Traffic
P	Average
SelfRe	
SelfRegulation	
SpokenArabicDigits	
UWaveGestureLibrary	
Best Count	

Zero-Shot

GTM ranks first on average



Few-Shot

GTM outperforms

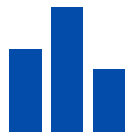
across 4 forecasting

M | CHRONOS-S1

0.571
0.632

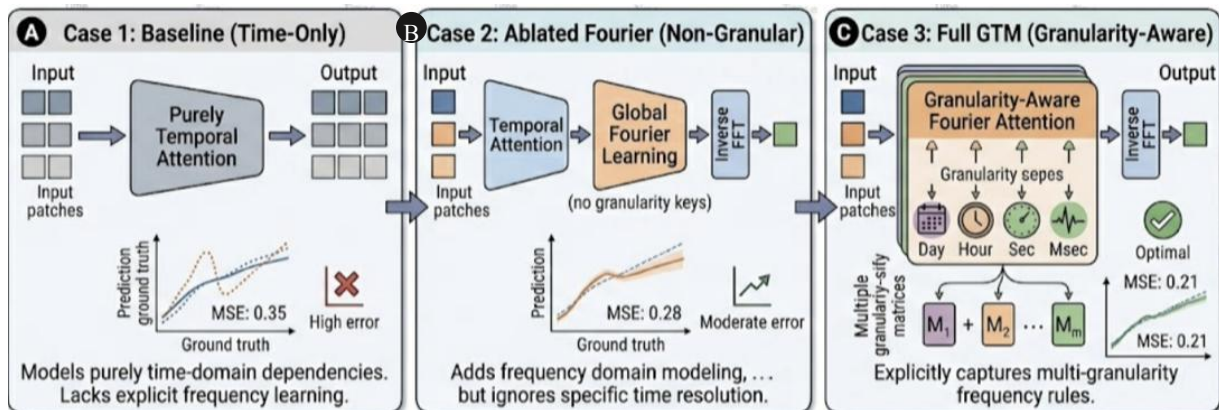
-
-
-

	96	192	336	720	96	96
Best Count	5	0	0	2	3	0



Main Results: Ablation & Efficiency

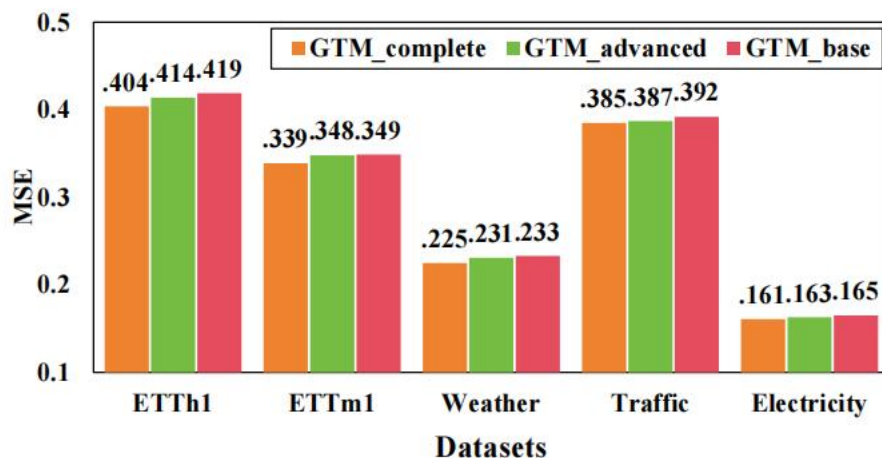
Ablation Study



Baseline GTM

Advanced GTM

Complete GTM



Model	Parameter	Training Speed	Inference Speed	Training Mem	Inference Mem
GTM	35.73M	0.290s/iter	0.165s/iter	8324.00MB	1250.00MB
Time-Moe(base)	50.00M	0.840s/iter	0.095s/iter	1812.48MB	226.70MB
GPT-2(6)-768	82.28M	0.104s/iter	0.054s/iter	5230.00MB	2566.00MB
FEDformer-768	30.75M	0.467s/iter	0.172s/iter	9535.00MB	1880.19MB
TimesNet-768	42.21M	1.849s/iter	0.547s/iter	35871.00MB	1904.18MB

Efficiency Analysis



Conclusion & Key Contributions

1. Model design

- Pioneered the first **generative-task-agnostic** foundation model for time-series analysis.
- Introduced a novel **Granularity-Aware Fourier Attention mechanism** to capture frequency-domain distributional differences.

2. Pre-training

- Developed a **unified pre-training framework** that jointly optimizes reconstruction and autoregressive objectives.
- Leverages a **hybrid masking strategy** alongside 2D positional encoding to enhance model robustness and generalizability.

3. Experimental Results

- Consistently outperforms SOTA baselines across diverse benchmarks with minimal or zero task-specific adaptation.
- Exhibits clear **scaling behavior** and highly **cost-effective efficiency**, making it highly suitable for real-time industrial applications.

Thank you!

Q & A