Machine Intelligence
Of Damo

# FEDformer

## Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting

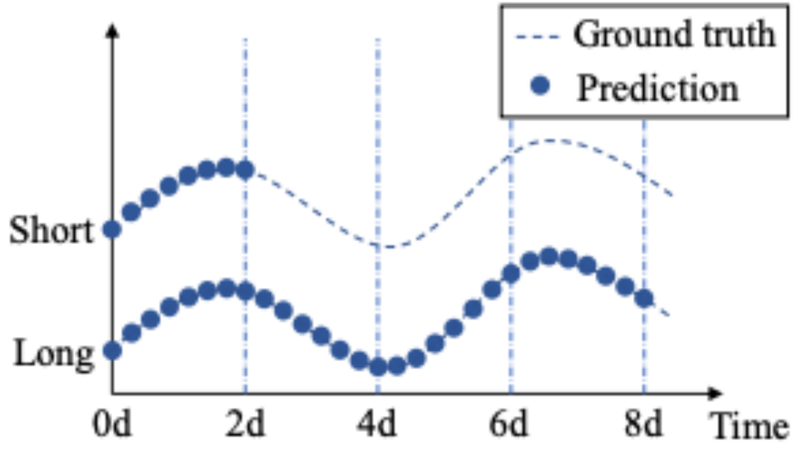
Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, Rong Jin

ICML22

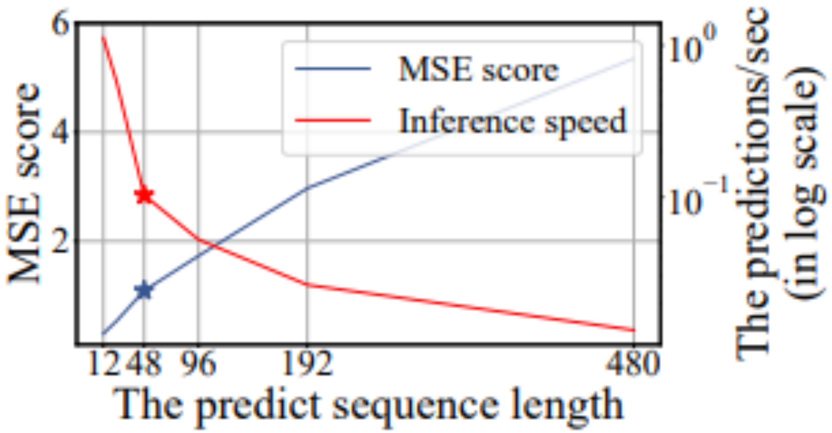
# Contents

目录

# Problem



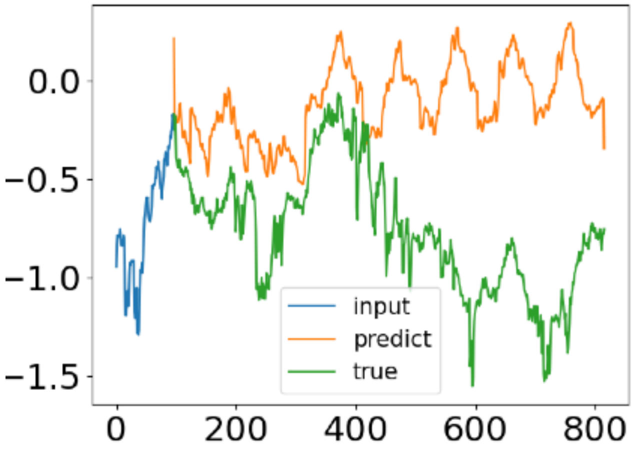

(a) Sequence Forecasting.

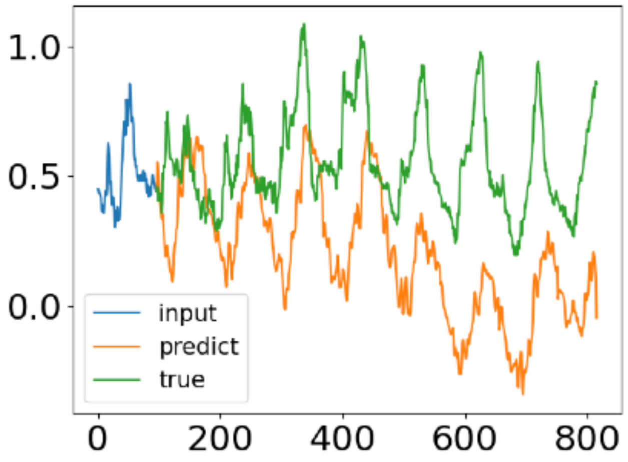



(b) Run LSTM on sequences.

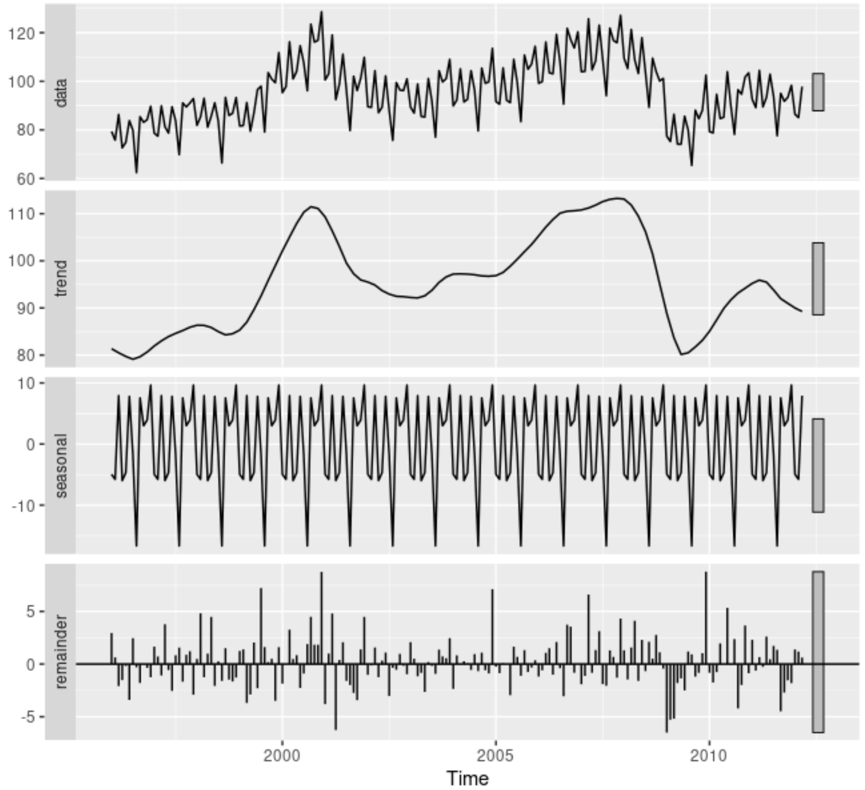
Zhou, Haoyi, et al. "Informer: Beyond efficient transformer for long sequence time-series forecasting." *Proceedings of AAAI*. 2021.

# Motivations




Trend and Seasonality Discrepancy




Seasonal and Trend decomposition

# Motivations

We can get a compact Representation of Time Series in Frequency Domain

**Theorem 1.** *Assume that $\mu(A)$, the coherence measure of matrix $A$, is $\Omega(k/n)$. Then, with a high probability, we have*

$$|A - P_{A'}(A)| \leq (1+\epsilon)|A - A_k|$$

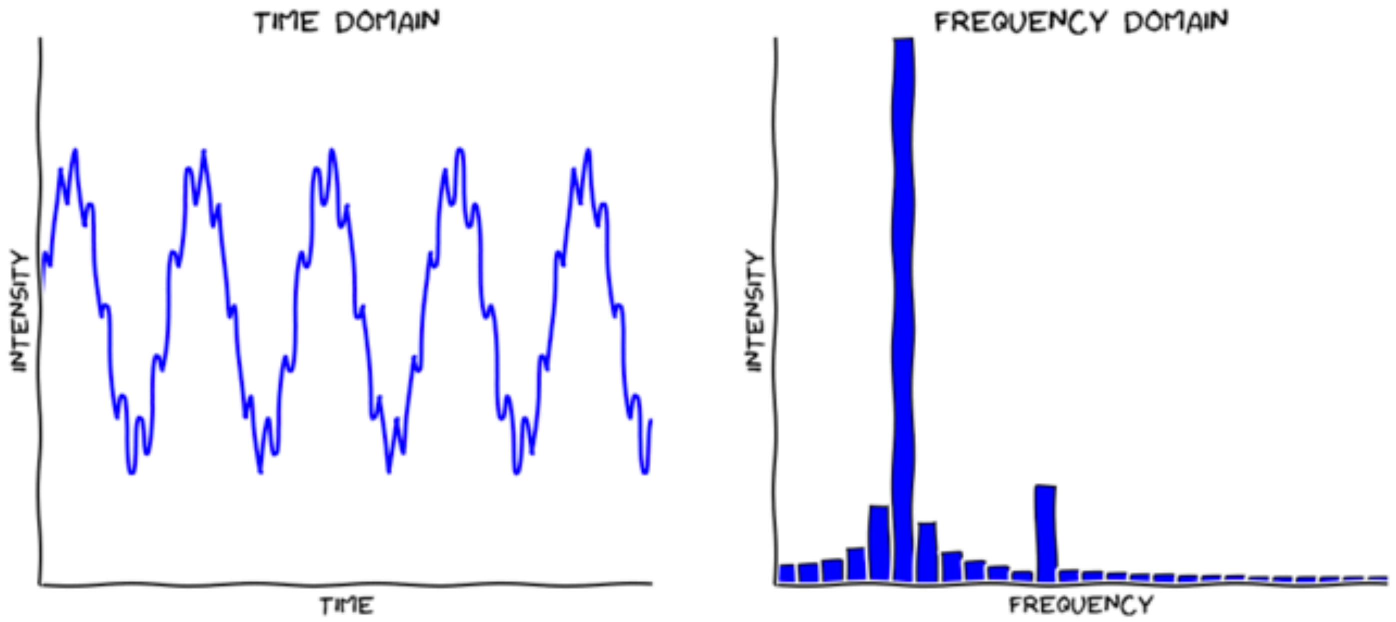
*if $s = O(k^2/\epsilon^2)$.*

Fourier transform

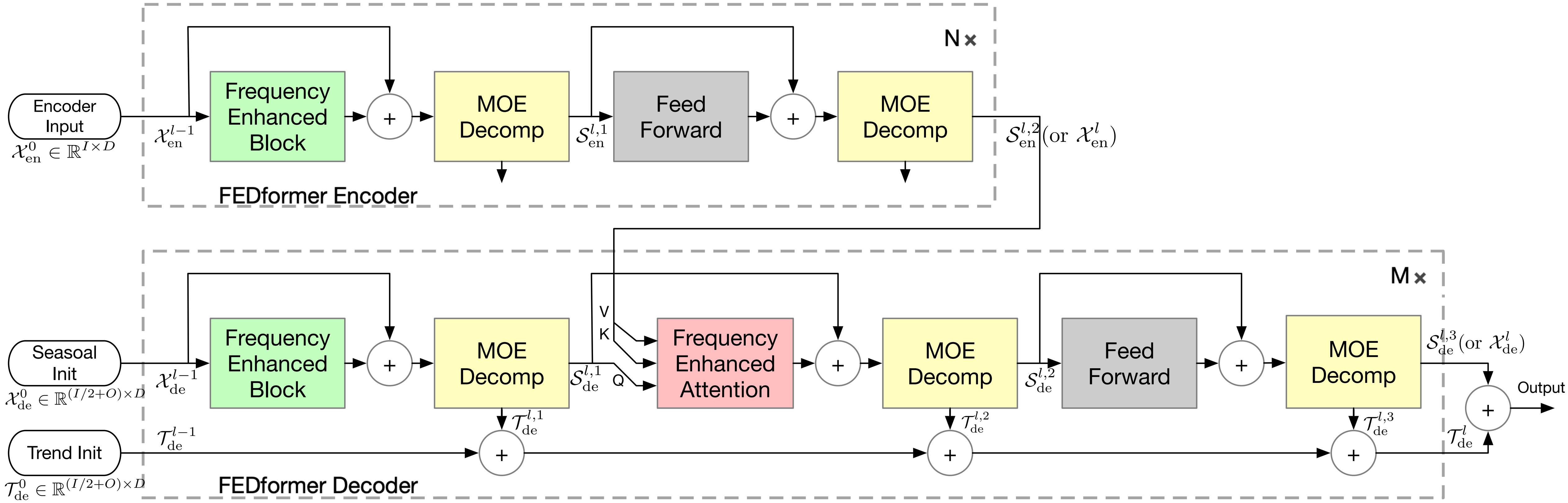
$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall\, \xi \in \mathbb{R}.$$

Fourier inverse transform

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{i2\pi\xi x}\, d\xi, \quad \forall\, x \in \mathbb{R},$$



# Model Structures

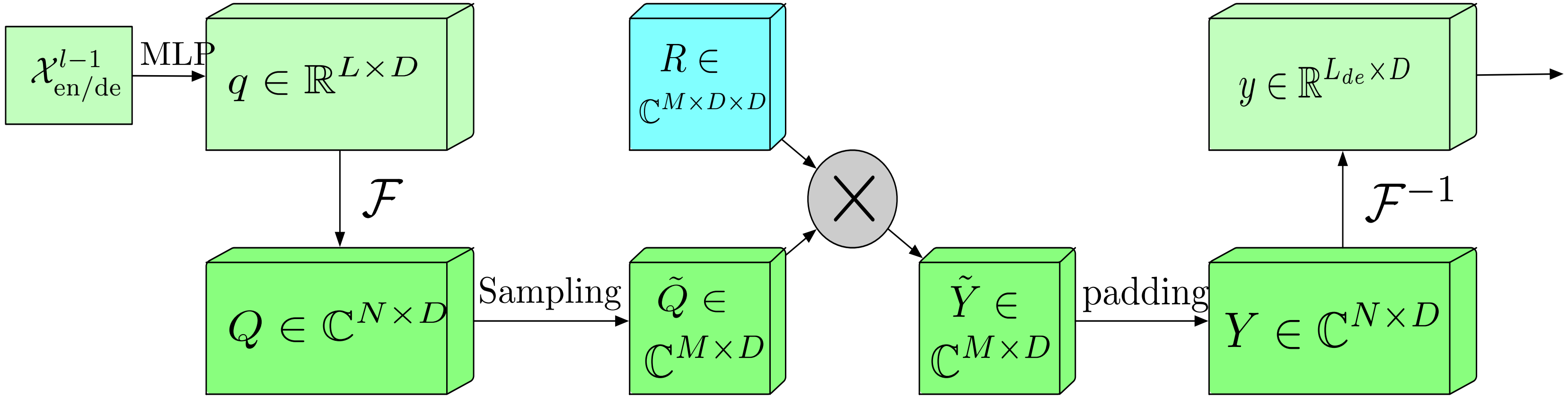



Frequency Enhanced Block: **Feature representation** for encoder and decoder signal separately

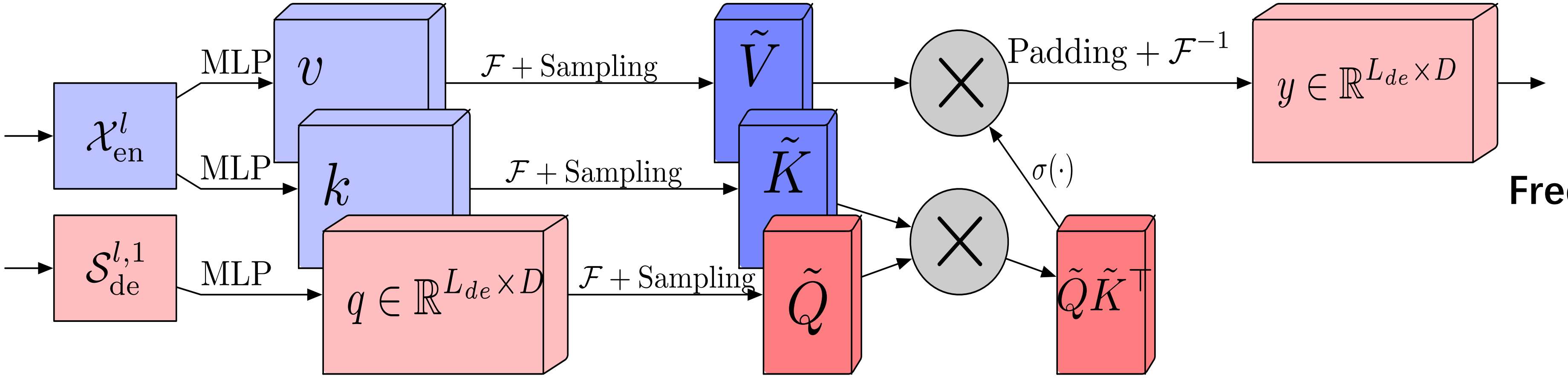
Frequency Enhanced Attention: **Cross feature** interaction between encoder and decoder signal

MOE Decomposition: **STL** decomposition

# Model Structures



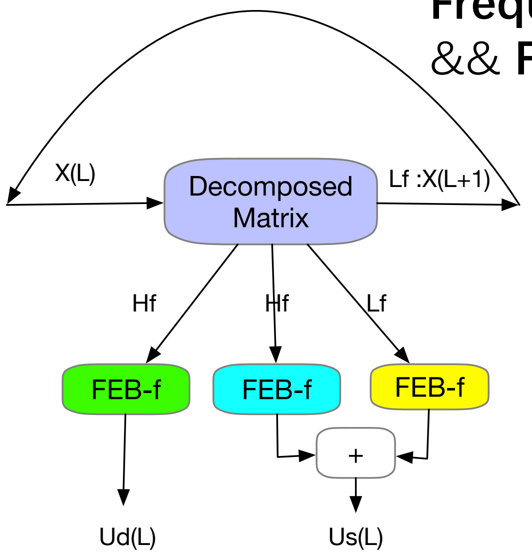

Frequecny Enhanced Blocks_f

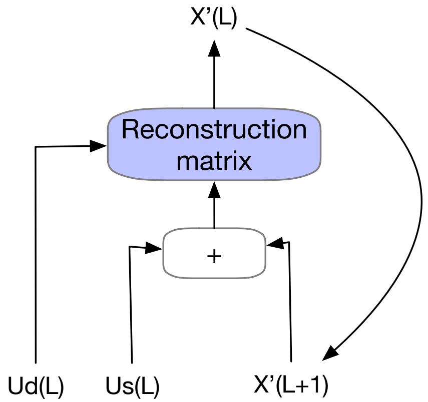



Frequency Enhanced attention_f

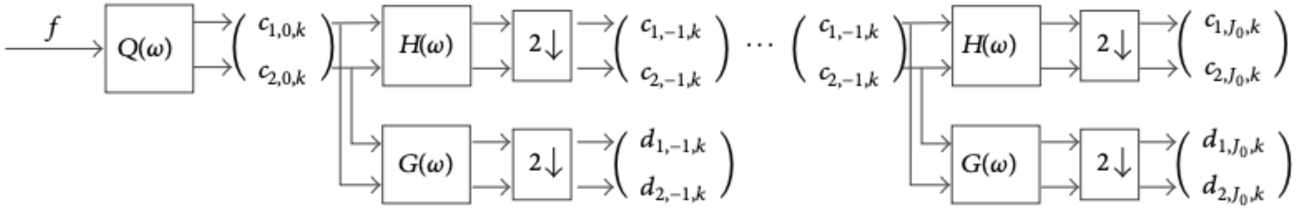
# Model Structures

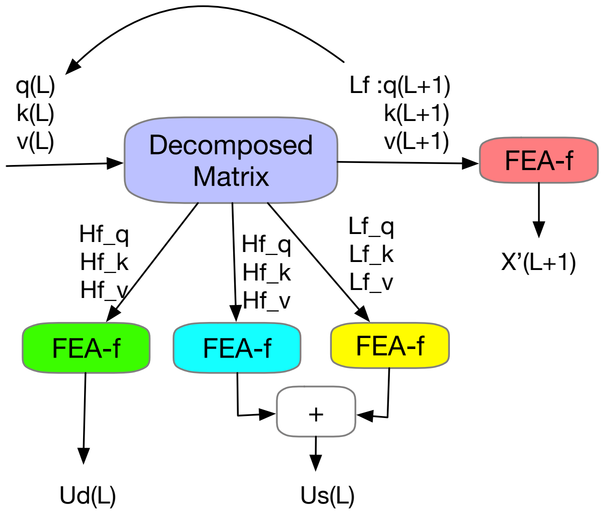
**Frequecny Enhanced Blocks_w**
&& **Frequency Enhanced attention_w**




Discrete Multiwavelet Decomposition




**MOE** Seasonal-Trend Decomposition

$$\mathbf{X_{trend}} = \mathbf{Softmax}(L(x)) * (F(x)), \qquad (10)$$

where $F(\cdot)$ is a set of average pooling filters and $\mathrm{Softmax}(L(x))$ is the weights for mixing these extracted trends.

# Experiments

*Table 2.* Multivariate long-term series forecasting results on six datasets with input length $I = 96$ and prediction length $O \in \{96, 192, 336, 720\}$ (For ILI dataset, we use input length $I = 36$ and prediction length $O \in \{24, 36, 48, 60\}$). A lower MSE indicates better performance, and the best results are highlighted in bold.

| Methods | Metric | ETTm2 | | | | Electricity | | | | Exchange | | | | Traffic | | | | Weather | | | | ILI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 96 | 192 | 336 | 720 | 96 | 192 | 336 | 720 | 96 | 192 | 336 | 720 | 96 | 192 | 336 | 720 | 96 | 192 | 336 | 720 | 24 | 36 | 48 | 60 |
| FEDformer-f | MSE | **0.203** | **0.269** | **0.325** | **0.421** | 0.193 | 0.201 | 0.214 | 0.246 | 0.148 | 0.271 | 0.460 | 1.195 | 0.587 | 0.604 | 0.621 | 0.626 | **0.217** | **0.276** | **0.339** | **0.403** | 3.228 | 2.679 | 2.622 | 2.857 |
| | MAE | **0.287** | **0.328** | **0.366** | **0.415** | 0.308 | 0.315 | 0.329 | 0.355 | 0.278 | 0.380 | 0.500 | 0.841 | 0.366 | 0.373 | 0.383 | 0.382 | **0.296** | **0.336** | **0.380** | **0.428** | 1.260 | 1.080 | 1.078 | 1.157 |
| FEDformer-w | MSE | 0.204 | 0.316 | 0.359 | 0.433 | **0.183** | **0.195** | **0.212** | **0.231** | **0.139** | **0.256** | **0.426** | **1.090** | **0.562** | **0.562** | **0.570** | **0.596** | 0.227 | 0.295 | 0.381 | 0.424 | **2.203** | **2.272** | **2.209** | **2.545** |
| | MAE | 0.288 | 0.363 | 0.387 | 0.432 | **0.297** | **0.308** | **0.313** | **0.343** | **0.276** | **0.369** | **0.464** | **0.800** | **0.349** | **0.346** | **0.323** | **0.368** | 0.304 | 0.363 | 0.416 | 0.434 | **0.963** | **0.976** | **0.981** | **1.061** |
| Autoformer | MSE | 0.255 | 0.281 | 0.339 | 0.422 | 0.201 | 0.222 | 0.231 | 0.254 | 0.197 | 0.300 | 0.509 | 1.447 | 0.613 | 0.616 | 0.622 | 0.660 | 0.266 | 0.307 | 0.359 | 0.419 | 3.483 | 3.103 | 2.669 | 2.770 |
| | MAE | 0.339 | 0.340 | 0.372 | 0.419 | 0.317 | 0.334 | 0.338 | 0.361 | 0.323 | 0.369 | 0.524 | 0.941 | 0.388 | 0.382 | 0.337 | 0.408 | 0.336 | 0.367 | 0.395 | 0.428 | 1.287 | 1.148 | 1.085 | 1.125 |
| Informer | MSE | 0.365 | 0.533 | 1.363 | 3.379 | 0.274 | 0.296 | 0.300 | 0.373 | 0.847 | 1.204 | 1.672 | 2.478 | 0.719 | 0.696 | 0.777 | 0.864 | 0.300 | 0.598 | 0.578 | 1.059 | 5.764 | 4.755 | 4.763 | 5.264 |
| | MAE | 0.453 | 0.563 | 0.887 | 1.338 | 0.368 | 0.386 | 0.394 | 0.439 | 0.752 | 0.895 | 1.036 | 1.310 | 0.391 | 0.379 | 0.420 | 0.472 | 0.384 | 0.544 | 0.523 | 0.741 | 1.677 | 1.467 | 1.469 | 1.564 |
| LogTrans | MSE | 0.768 | 0.989 | 1.334 | 3.048 | 0.258 | 0.266 | 0.280 | 0.283 | 0.968 | 1.040 | 1.659 | 1.941 | 0.684 | 0.685 | 0.7337 | 0.717 | 0.458 | 0.658 | 0.797 | 0.869 | 4.480 | 4.799 | 4.800 | 5.278 |
| | MAE | 0.642 | 0.757 | 0.872 | 1.328 | 0.357 | 0.368 | 0.380 | 0.376 | 0.812 | 0.851 | 1.081 | 1.127 | 0.384 | 0.390 | 0.408 | 0.396 | 0.490 | 0.589 | 0.652 | 0.675 | 1.444 | 1.467 | 1.468 | 1.560 |
| Reformer | MSE | 0.658 | 1.078 | 1.549 | 2.631 | 0.312 | 0.348 | 0.350 | 0.340 | 1.065 | 1.188 | 1.357 | 1.510 | 0.732 | 0.733 | 0.742 | 0.755 | 0.689 | 0.752 | 0.639 | 1.130 | 4.400 | 4.783 | 4.832 | 4.882 |
| | MAE | 0.619 | 0.827 | 0.972 | 1.242 | 0.402 | 0.433 | 0.433 | 0.420 | 0.829 | 0.906 | 0.976 | 1.016 | 0.423 | 0.420 | 0.420 | 423 | 0.596 | 0.638 | 0.596 | 0.792 | 1.382 | 1.448 | 1.465 | 1.483 |

*Table 1.* A subset of the benchmark showing both Mean and STD.

| MSE | | ETTm2 | Electricity | Exchange | Traffic |
|---|---|---|---|---|---|
| FED-f | 96 | 0.203 ± 0.0042 | 0.194 ± 0.0008 | 0.148 ± 0.002 | 0.217 ± 0.008 |
| | 192 | 0.269 ± 0.0023 | 0.201± 0.0015 | 0.270± 0.008 | 0.604 ± 0.004 |
| | 336 | 0.325 ± 0.0015 | 0.215± 0.0018 | 0.460± 0.016 | 0.621 ± 0.006 |
| | 720 | 0.421 ± 0.0038 | 0.246± 0.0020 | 1.195± 0.026 | 0.626 ± 0.003 |
| Autoformer | 96 | 0.255 ± 0.020 | 0.201± 0.003 | 0.197± 0.019 | 0.613± 0.028 |
| | 192 | 0.281 ± 0.027 | 0.222± 0.003 | 0.300± 0.020 | 0.616± 0.042 |
| | 336 | 0.339 ± 0.018 | 0.231± 0.006 | 0.509± 0.041 | 0.622± 0.016 |
| | 720 | 0.422 ± 0.015 | 0.254± 0.007 | 1.447± 0.084 | 0.419± 0.017 |

# Experiments

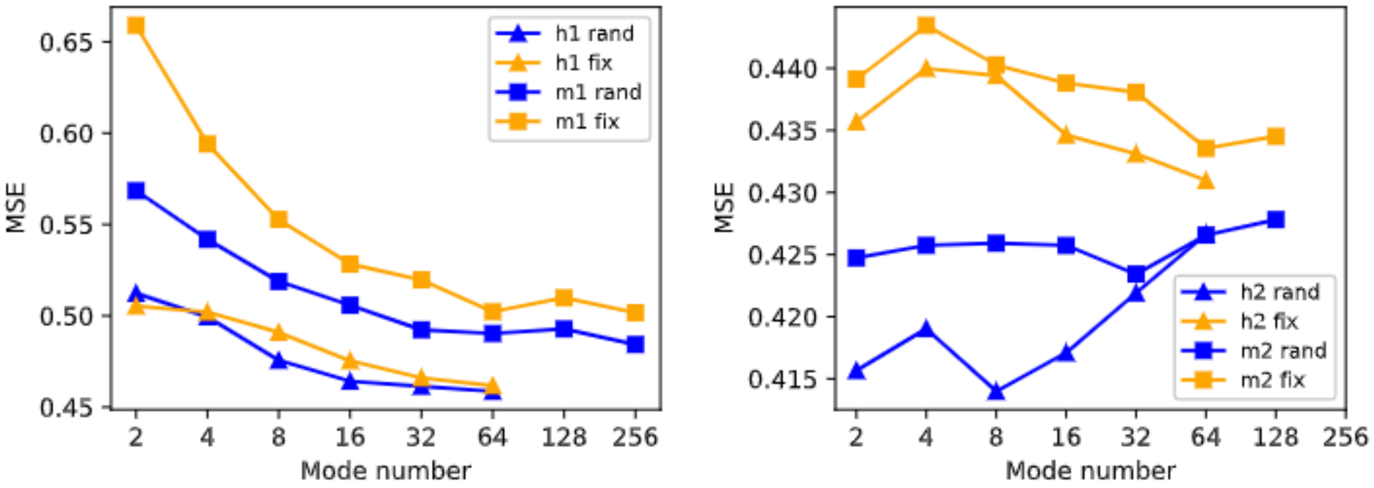



*Figure 6.* Comparison of two base-modes selection method (Fix&Rand). Rand policy means randomly selecting a subset of modes, Fix policy means selecting the lowest frequency modes. Two policies are compared on a variety of base-modes number $M \in \{2, 4, 8 \ldots 256\}$ on ETT full-benchmark (h1, m1, h2, m2).

**Random** policy is better

Mode **saturation** at 32

# Experiments

**Rule of thumb** for mode selection && model selection

*Table 3.* Perm Entropy Complexity comparison for multi vs uni

| Permutation Entropy | Electricity | Traffic | Exchange | Illness |
|---|---|---|---|---|
| Multivariate | 0.910 | 0.792 | 0.961 | 0.960 |
| Univariate | 0.902 | 0.790 | 0.949 | 0.867 |

**Wavelet** for **complex** dataset
**Fourier** for **less complex** dataset

*Table 2.* Complexity experiments for datasets

| Methods | ETTh1 | ETTh2 | ETTm1 | ETTm2 |
|---|---|---|---|---|
| Permutation Entropy | 0.954 | 0.866 | 0.959 | 0.788 |
| SVD Entropy | 0.807 | 0.495 | 0.589 | 0.361 |

**More mode** for more complex dataset

# Experiments

The Kolmogorov-Smirnov statistic is

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

$$D_{n,m} > \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{n+m}{n \cdot m}},$$

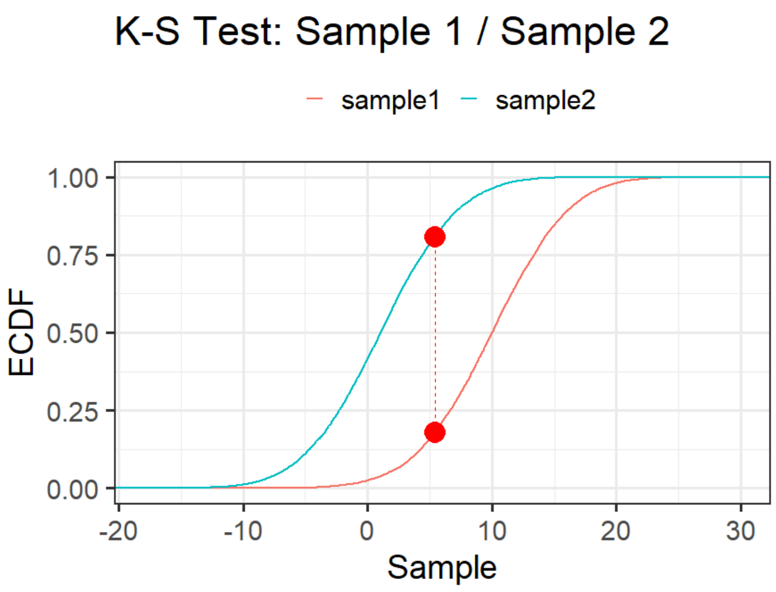
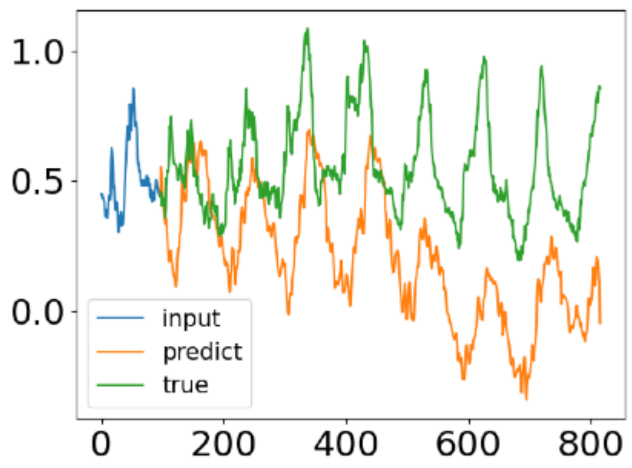



*Table 5.* P-values of Kolmogrov-Smirnov test of different transformer models for long-term forecasting output on ETTm1 and ETTm2 dataset. Larger value indicates the hypothesis (the input sequence and forecasting output come from the same distribution) is less likely to be rejected. The best results are highlighted.

| Methods | | Transformer | Informer | Autoformer | FEDformer | True |
|---|---|---|---|---|---|---|
| ETTm1 | 96 | 0.0090 | 0.0055 | 0.020 | **0.048** | 0.023 |
| | 192 | 0.0052 | 0.0029 | 0.015 | **0.028** | 0.013 |
| | 336 | 0.0022 | 0.0019 | 0.012 | **0.015** | 0.010 |
| | 720 | 0.0023 | 0.0016 | 0.008 | **0.014** | 0.004 |
| ETTm2 | 96 | 0.0012 | 0.0008 | **0.079** | 0.071 | 0.087 |
| | 192 | 0.0011 | 0.0006 | **0.047** | 0.045 | 0.060 |
| | 336 | 0.0005 | 0.00009 | 0.027 | **0.028** | 0.042 |
| | 720 | 0.0008 | 0.0002 | **0.023** | 0.021 | 0.023 |
| Count | | 0 | 0 | 3 | 5 | NA |



# Experiments

Running Time: O(L)

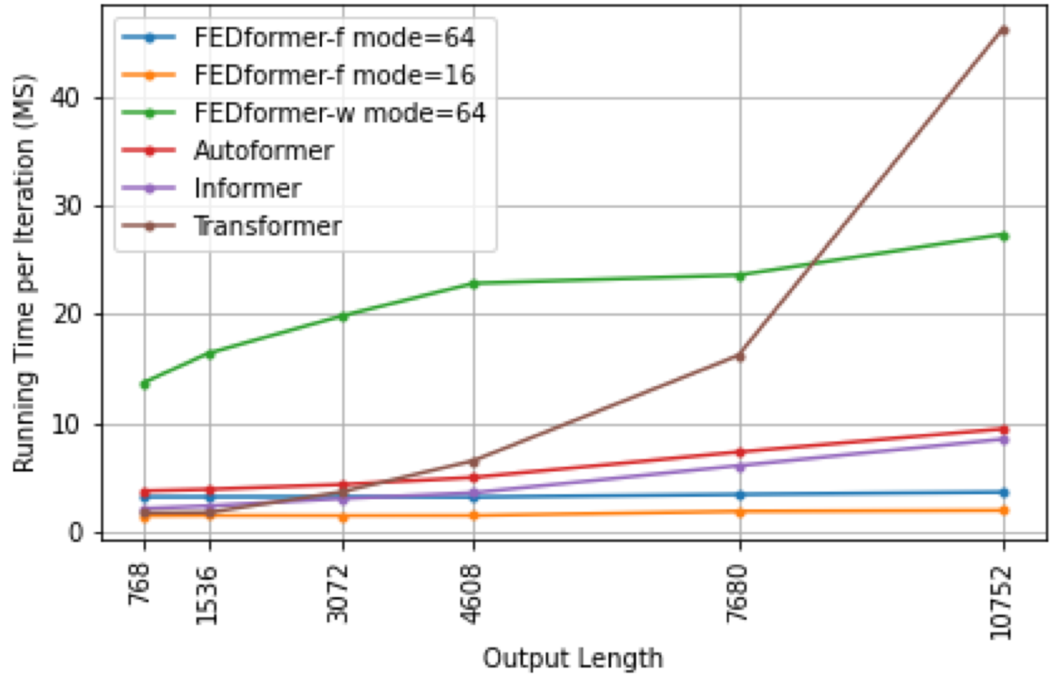



*Table 1.* Complexity analysis of different forecasting models.

| Methods | Training | | Testing |
|---|---|---|---|
| | Time | Memory | Steps |
| FEDformer | $\mathcal{O}(L)$ | $\mathcal{O}(L)$ | 1 |
| Autoformer | $\mathcal{O}(L \log L)$ | $\mathcal{O}(L \log L)$ | 1 |
| Informer | $\mathcal{O}(L \log L)$ | $\mathcal{O}(L \log L)$ | 1 |
| Transformer | $\mathcal{O}\left(L^2\right)$ | $\mathcal{O}\left(L^2\right)$ | $L$ |
| LogTrans | $\mathcal{O}(L \log L)$ | $\mathcal{O}\left(L^2\right)$ | 1 |
| Reformer | $\mathcal{O}(L \log L)$ | $\mathcal{O}(L \log L)$ | $L$ |
| LSTM | $\mathcal{O}(L)$ | $\mathcal{O}(L)$ | $L$ |

# Conclusion

- We propose a frequency enhanced decomposed Transformer architecture with mixture of experts for seasonal-trend decomposition in order to better capture global properties of time series.

- We propose **Fourier enhanced blocks and Wavelet enhanced blocks** in the Transformer structure that allows us to capture important structures in time series through frequency domain mapping. They serve as substitutions for both self-attention and cross-attention blocks.

- By **randomly selecting** a **fixed number** of Fourier components, the proposed model achieves linear computational complexity and memory cost. The effectiveness of this selection method is verified both theoretically and empirically.

- We conduct extensive experiments over **6 benchmark datasets** across multiple domains (energy, traffic, economics, weather and disease). Our empirical studies show that the proposed model improves the performance of state-of-the-art methods by **14.8% and 22.6%** for multivariate and univariate forecasting, respectively.

# Thank You

https://arxiv.org/abs/2201.12740

https://github.com/MAZiqing/FEDformer