SPM: Modeling Dependencies across Arbitrary Positions for Effective High-Dimensional Long-Term Series Forecasting

Anonymous Authors[1]

		Time 1D FFT	Vari 1D FFT	w/o FFT	Multiply	SPM
Dataset	Prediction	MSE / MAE	MSE / MAE	MSE / MAE	MSE / MAE	MSE / MAE
Traffic	96	0.423 / 0.275	0.443 / 0.314	0.429 / 0.279	0.387 / 0.263	**0.379 / 0.259**
Traffic	192	0.436 / 0.279	0.458 / 0.318	0.440 / 0.281	0.405 / **0.270**	**0.401** / 0.271
Traffic	336	0.450 / 0.288	0.487 / 0.333	0.453 / 0.289	0.423 / 0.278	**0.418 / 0.277**
Traffic	720	0.457 / 0.288	0.52 / 0.347	0.490 / 0.309	0.457 / 0.297	**0.455 / 0.298**
Traffic	AVG	0.442 / 0.285	0.477 / 0.328	0.453 / 0.290	0.418 / 0.277	**0.413 / 0.276**
Electricity	96	0.146 / 0.244	0.141 / 0.244	0.145 / 0.244	**0.135** / 0.239	0.136 / **0.238**
Electricity	192	0.164 / 0.254	0.156 / 0.250	0.163 / 0.254	**0.157 / 0.250**	0.158 / 0.252
Electricity	336	0.181 / 0.273	0.177 / 0.273	0.180 / 0.271	0.177 / 0.273	**0.174 / 0.271**
Electricity	720	0.220 / 0.305	0.218 / 0.309	0.223 / 0.307	0.214 / 0.306	**0.199 / 0.292**
Electricity	AVG	0.178 / 0.269	0.173 / 0.269	0.178 / 0.269	0.171 / 0.267	**0.167 / 0.263**

Figure 1. Ablation study with a thorough isolation.

Traffic-96	Time (s/epoch)	Memory (GB)	Param (M)
Autoformer	17.564	1.406	1.369
FEDformer	41.203	1.286	4.515
Crossformer	63.143	15.546	2.332
PatchTST	24.632	6.522	0.249
iTransformer	11.455	1.776	**0.125**
Leddam	118.715	2.130	0.305
xPatch	19.135	4.828	0.146
S-Mamba	11.661	**1.278**	0.189
SPM	**10.154**	1.582	0.140

Figure 2. Thoroughly analysis of computing complexity.

1. Reviewer FtbZ

We thank Reviewer FtbZ for such a detailed review. We hope we have addressed all your concerns as follows:

Q1: ablation studies to isolate the impact of different components of the proposed method.

We regard FFT+Hadamard product as a whole method derived from our theoretical proofs, thus there is no thing to isolate. If without FFT+Hadamard, the model cannot work. To address your concern, we conduct an ablation study with a thorough isolation (Figure 1).

’Time 1D FFT’, ’Vari 1D FFT’, ’w/o FFT’, ’Maltiply’, and ’Add’ represents 1D FFT at time dimension, 1D FFT at variable dimension, without FFT, matrix multiplication, and matrix addition, respectively. It can be seen, replacing 2D FFT by 1D FFT at time or variable dimension causes larger error (Traffic average MSE/MAE of 0.442/0.285 and 0.477/0.328). This indicates the effectiveness of SPM’s global-wise modeling paradigm. Without FFT makes the same phenomenon (Traffic average MSE/MAE of 0.453/0.290) and indication. Further, we hold 2D FFT and replace Hadamard product to matrix multiplication. From the ’Multiply’ results, 12 metrics of MSE/MAE are larger than SPM, while only 4 metrics are smaller. Moreover, its average MSE/MAE of Traffic (0.418/0.277) and Electricity (0.171/0.267) is higher than SPM with 0.413/0.276 and 0.167/0.263. This indicates the effectiveness of SPM’s Hadamard product.

Q2: the actual runtime and memory usage comparisons with other methods could be more thoroughly analyzed.

We conduct a thoroughly analysis on large-scale dataset Traffic and Electricity with full 4 prediction length, shown as Figure 2-9. On the metric of ’Time (s/epoch)’, SPM ranks first at full 4 prediction length of the all two datasets. This validates the low complexity of SPM in time. In terms of ’Memory (GB)’, SPM ranks first at Traffic-720, second at Electricity-336, third at Traffic-192/336 and Electricity-96/192/720, and fouth at Traffic-96. This validates the competition of SPM in space complexity. In terms of ’Param (M)’, ranks first at Traffic/Electricity-720, second at Traffic-96 and Electricity-96/192/336, third at Traffic-192/336. This validates the light parameter number of SPM.

Q3: how these theoretical results translate into practical improvements in the context of time series forecasting.

SPM projects the input signal into Q, K, and V for autocorrelation calculation, extending it to deep learning for parameterized learning and enhanced non-linear expression ability. In the frequency domain, Q and K are first mul-

Traffic-192	Time (s/epoch)	Memory (GB)	Param (M)
Autoformer	25.063	2.282	1.468
FEDformer	45.659	1.572	4.515
Crossformer	61.121	26.348	2.999
PatchTST	26.813	6.254	0.397
iTransformer	14.808	2.088	**0.137**
Leddam	166.338	1.686	3.293
xPatch	23.154	5.624	0.401
S-Mamba	15.988	**1.432**	0.202
SPM	**14.016**	1.594	0.149

Figure 3. Thoroughly analysis of computing complexity.

Traffic-336	Time (s/epoch)	Memory (GB)	Param (M)
Autoformer	29.063	3.134	**1.468**
FEDformer	49.687	2.160	4.515
Crossformer	99.538	42.114	3.999
PatchTST	31.090	6.962	0.618
iTransformer	19.731	2.342	0.156
Leddam	170.349	1.776	3.367
xPatch	28.273	6.850	1.008
S-Mamba	21.254	**1.706**	0.220
SPM	**18.639**	1.998	0.163

Figure 4. Thoroughly analysis of computing complexity.

tiplied, which represents the auto-correlation calculation in the time domain. Then, the result is multiplied by V, which corresponds to the weighted integration of the auto-correlation scores with the input signal through convolution in the time domain, enabling adaptive attention to different parts of the input signal.

Q4: more detailed hyperparameter tuning and sensitivity analysis to ensure that the results are robust to different settings.

The hyperparameter search setting of SPM is as follows. ('d model':[32, 256, 2], 'd ff':[32, 256, 2], "hidden size1":[32, 256, 2], "hidden size2":[32, 256, 2], 'batch size':[32, 256, 2], 'learning rate':[0.00001, 0.01, 0.00001]), where they represent embedding dimension, mapping dimension of feed-forward network in Transformer, first layer, second layer of MLP, batch size, learning rate, respectively. [a, b, c] represents the section is [a, b] and the sample interval is c.

We perform a hyperparameter sensitivity analysis on Electricity dataset for important hyperparameter (encoder head number, MLP layer number, MLP layer hidden size, encoder layer number, d model/d ff) and show each hyperparameter in order.

In terms of encoder head number (Figure 10), when the prediction length is the shortest (96), the mean value ± standard deviation of MSE and MAE are 0.135±0.0004 and 0.237±0.0004, respectively. When the prediction length is the longest (720), they are 0.231±0.019 and 0.318±0.015, respectively. The best encoder head number is 1.

In terms of MLP layer number (Figure 11), the layer hidden size is set to 192. When the prediction length is the shortest (96), the mean value ± standard deviation of MSE and MAE are 0.149±0.011 and 0.250±0.008, respectively. When the prediction length is the longest (720), they are 0.231±0.008 and 0.320±0.005, respectively. The best MLP layer number is around 3.

In terms of MLP layer hidden size (Figure 12), the layer number is set to 2. When the prediction length is the shortest (96), the mean value ± standard deviation of MSE and MAE are 0.143±0.006 and 0.244 ± 0.005, respectively. When the prediction length is the longest (720), they are 0.226±0.007 and 0.315±0.004, respectively. The best MLP layer hidden size is 672 or 768.

In terms of encoder layer number (Figure 13), when the prediction length is the shortest (96), the mean value ± standard deviation of MSE and MAE are 0.528±0.748 and 0.426±0.329, respectively. When the prediction length is the longest (720), they are 0.238±0.010 and 0.323±0.008, respectively. The best encoder layer number is 1.

In terms of d model/d ff (Figure 14), when the prediction length is the shortest (96), the mean value ± standard deviation of MSE and MAE are 0.135±0.001 and 0.237±0.002, respectively. When the prediction length is the longest (720), they are 0.238±0.010 and 0.323±0.008, respectively. The best encoder layer number is around 64.

Q5: generalizability of the method to different types of time series (e.g., irregularly sampled, non-stationary).

SPM is designed for regularly sampled time series. According to the paper (Frequency Adaptive Normalization For Non-stationary Time Series Forecasting, NeurIPS 2024), all benchmarks employed by SPM are non-stationary.

Q6: discussion on the limitations of the method (e.g., po-

Traffic-720	Time (s/epoch)	Memory (GB)	Param (M)
Autoformer	42.284	5.440	1.468
FEDformer	65.099	3.510	4.515
Crossformer	109.653	85.598	6.665
PatchTST	39.568	7.542	1.209
iTransformer	31.556	2.880	0.205
Leddam	179.943	2.528	3.565
xPatch	44.463	9.900	3.944
S-Mamba	32.552	2.806	0.270
SPM	**30.195**	**2.390**	**0.200**

Figure 5. Thoroughly analysis of computing complexity.

Electricity-96	Time (s/epoch)	Memory (GB)	Param (M)
Autoformer	18.778	1.122	0.676
FEDformer	55.051	1.020	3.822
Crossformer	37.504	6.096	1.501
PatchTST	14.642	2.672	0.249
iTransformer	8.620	0.742	**0.125**
Leddam	101.916	1.816	3.238
xPatch	11.094	2.086	0.145
S-Mamba	8.855	**0.684**	0.189
SPM	**7.673**	0.860	0.140

Figure 6. Thoroughly analysis of computing complexity.

tential issues with very long sequences or extremely high-dimensional data).

We gradually increase input variable and time dimensions at a 1:1 ratio on Traffic dataset to explore the potential issues brought by extreme input-size (Figure 15).

As the input-size of variable and time dimensions increases, the error gradually decreases. The error reaches its minimum at an input-size of 500 or 600. Then, when input-size is greater than 600, the error starts to rise. Therefore, extremely large input-size (800) results in larger prediction error.

Q7: In Table 5, SPM shows some limitations.

Please review Q2, where SPM shows complete and strong advantages.

Through adaptively adjusting QKV weights for non-linear expression, SPM achieves dynamic attention to handle the dependencies of abrupt accident.

2. Reviewer BWeG

We thank Reviewer BWeG for such a detailed review. We hope we have addressed all your concerns as follows:

Q4: The experiment corresponding to figure 5 is meaningless and cannot explain anything.

We conduct more ratio experiment with L:C at a 1:2 (Figure 16) and L:C at a 2:1 (Figure 17). It can be seen that SPM achieves lowest MSE in the whole process. Moreover, SPM' prediction error decreases progressively during the increase with L and C, it indicates that SPM can effectively model the dependencies among the input samples before and after the increase.

Electricity-192	Time (s/epoch)	Memory (GB)	Param (M)
Autoformer	20.777	1.740	0.676
FEDformer	57.125	1.166	3.822
Crossformer	57.791	9.962	1.752
PatchTST	15.741	2.550	0.397
iTransformer	10.261	0.790	**0.137**
Leddam	104.741	2.626	3.287
xPatch	12.687	2.342	0.399
S-Mamba	11.055	**0.784**	0.202
SPM	**9.925**	0.938	0.149

Figure 7. Thoroughly analysis of computing complexity.

Electricity-336	Time (s/epoch)	Memory (GB)	Param (M)
Autoformer	24.038	2.250	0.676
FEDformer	57.438	1.476	3.822
Crossformer	98.105	15.780	2.129
PatchTST	17.467	2.894	0.618
iTransformer	12.704	0.928	**0.156**
Leddam	75.784	1.022	0.365
xPatch	16.052	2.808	1.006
S-Mamba	14.286	**0.924**	0.220
SPM	**11.973**	1.020	0.163

Figure 8. Thoroughly analysis of computing complexity.

Electricity-720	Time (s/epoch)	Memory (GB)	Param (M)
Autoformer	33.295	3.946	0.676
FEDformer	66.014	2.154	3.822
Crossformer	104.632	32.138	3.134
PatchTST	23.194	3.056	1.209
iTransformer	20.719	1.174	0.205
Leddam	113.332	3.056	3.558
xPatch	26.115	3.950	3.943
S-Mamba	21.516	**1.110**	0.270
SPM	**19.156**	1.208	**0.200**

Figure 9. Thoroughly analysis of computing complexity.

n_head		1	2	4	8	16
Dataset	prediction length	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE
Electricity	96	0.136/0.238	**0.135/0.237**	**0.135/0.237**	**0.135/0.237**	**0.135/0.237**
Electricity	720	**0.199/0.292**	0.247/0.328	0.248/0.329	0.239/0.326	0.223/0.313

Figure 10. Hyperparameter sensitivity analysis of encoder head number.

MLP layer number		1	2	3	4	5
Dataset	prediction length	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE
Electricity	96	0.168/0.263	0.145/0.246	**0.143/0.245**	**0.143**/0.246	0.145/0.249
Electricity	720	0.248/0.324	**0.224/0.312**	0.234/0.325	**0.224**/0.318	0.226/0.319

Figure 11. Hyperparameter sensitivity analysis of MLP layer number.

MLP layer hidden size		96	192	288	384	480	576	672	768
Dataset	prediction length	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE
Electricity	96	0.156/0.255	0.145/0.246	0.142/0.243	0.141/0.243	0.143/0.244	0.140/0.242	0.140/0.242	**0.138/0.239**
Electricity	720	0.236/0.319	0.224/0.312	0.222/0.313	0.230/0.319	0.225/0.316	0.226/0.319	**0.217/0.309**	0.225/0.316

Figure 12. Hyperparameter sensitivity analysis of MLP layer hidden size.

encoder layer number		1	2	3	4	5
Dataset	prediction length	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE
Electricity	96	**0.136/0.238**	0.141/0.242	0.143/0.243	0.348/0.416	1.874/0.992
Electricity	720	**0.199/0.292**	0.211/0.303	0.211/0.305	0.418/0.463	1.575/0.976

Figure 13. Hyperparameter sensitivity analysis of encoder layer number.

d_model/d_ff		32	64	128	256	512
Dataset	prediction length	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE
Electricity	96	**0.134**/0.236	**0.134/0.235**	**0.134**/0.236	0.136/0.238	0.136/0.239
Electricity	720	**0.225**/0.316	0.228/**0.313**	0.247/0.329	0.242/0.328	0.246/0.331

Figure 14. Hyperparameter sensitivity analysis of d model/d ff.

Input-size	100	200	300	400	500	600	700	800
Traffic	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE	MSE/MAE
SPM	0.311/0.273	0.304/0.261	0.294/0.259	0.292/0.259	**0.290**/0.260	0.294/**0.256**	0.300/0.258	0.317/0.265

Figure 15. Potential issues brought by extreme input-size.

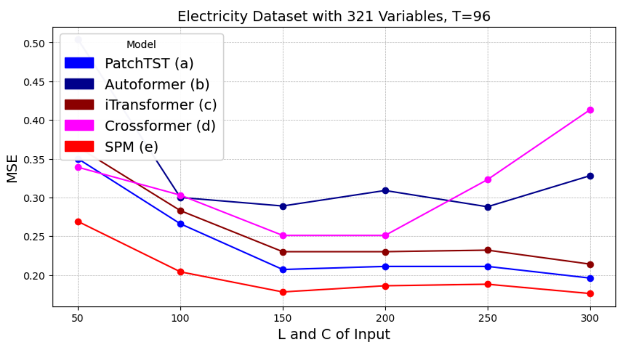

Figure 16. Experiment of simulating a uniform 2D plane to modeling dependencies across arbitrary positions. L:C at a 1:2

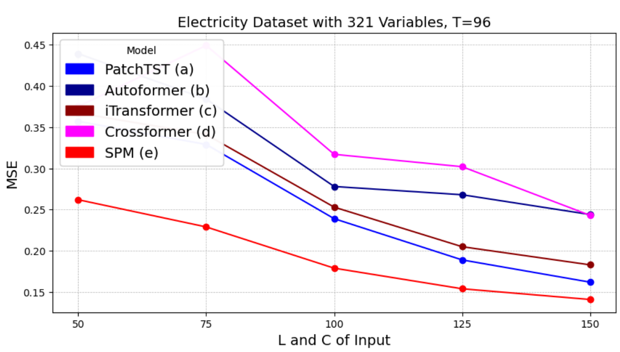

Figure 17. Experiment of simulating a uniform 2D plane to modeling dependencies across arbitrary positions. L:C at a 2:1