北京航空航天大學
BEIHANG UNIVERSITY

2024 ICLR

CARD

Channel Aligned Robust Blend Transformer for Time Series Forcasting

Xue Wang* Tian Zhou* Qingsong Wen Jinyang Gao Bolin Ding Rong Jin
{xue.w, tian.zt, qingsong.wen, jinyang.gjy, bolin.ding, jinrong.jr}@alibaba-inc.com

2024.10.27

问题描述

- **Channel Dependent & Channel Independent 如何抉择?**
  - CD捕捉不同预测变量之间的依赖关系：多变量的时序预测
  - CI不同变量独立预测： PatchTST, DLinear
  - CD建模能力强，CI鲁棒性好
  - CD（特别是在Transformer类模型中）在长回望窗口时带来过拟合问题

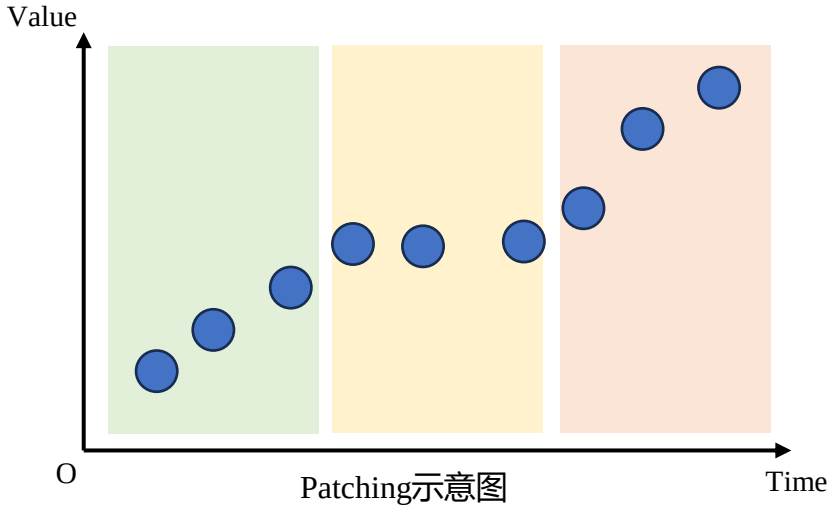
- **如何有效利用不同信道（变量）之间的依赖性?**
- **如何缓解时间序列预测中过拟合噪声的问题?**

创新点Innovation

- **捕获跨通道信息 & 变量相关性**
  - Patch Tokens：在每个token内部对齐局部信息；
  - Attention关注不同的通道和隐藏层间的联系；
  - Token Blend模块：将同一头内临近的token合并为新token；
- **解决过拟合噪声问题，提高鲁棒性**
  - 注意力机制内的指数移动平均层（EMA）、动态映射模块（DP）；
  - 新的基于信号衰减的鲁棒性损失函数；

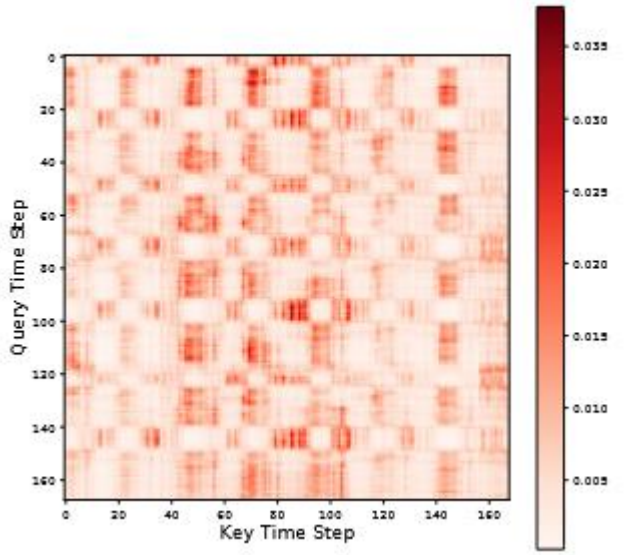
Related Work

- **Patch思想的应用**
  - NLP：基于字符编码到基于子词划分（语义信息）；
  - CV：将图像切割为若干小块；
  - 语音识别：将原语音序列切分为若干子序列；



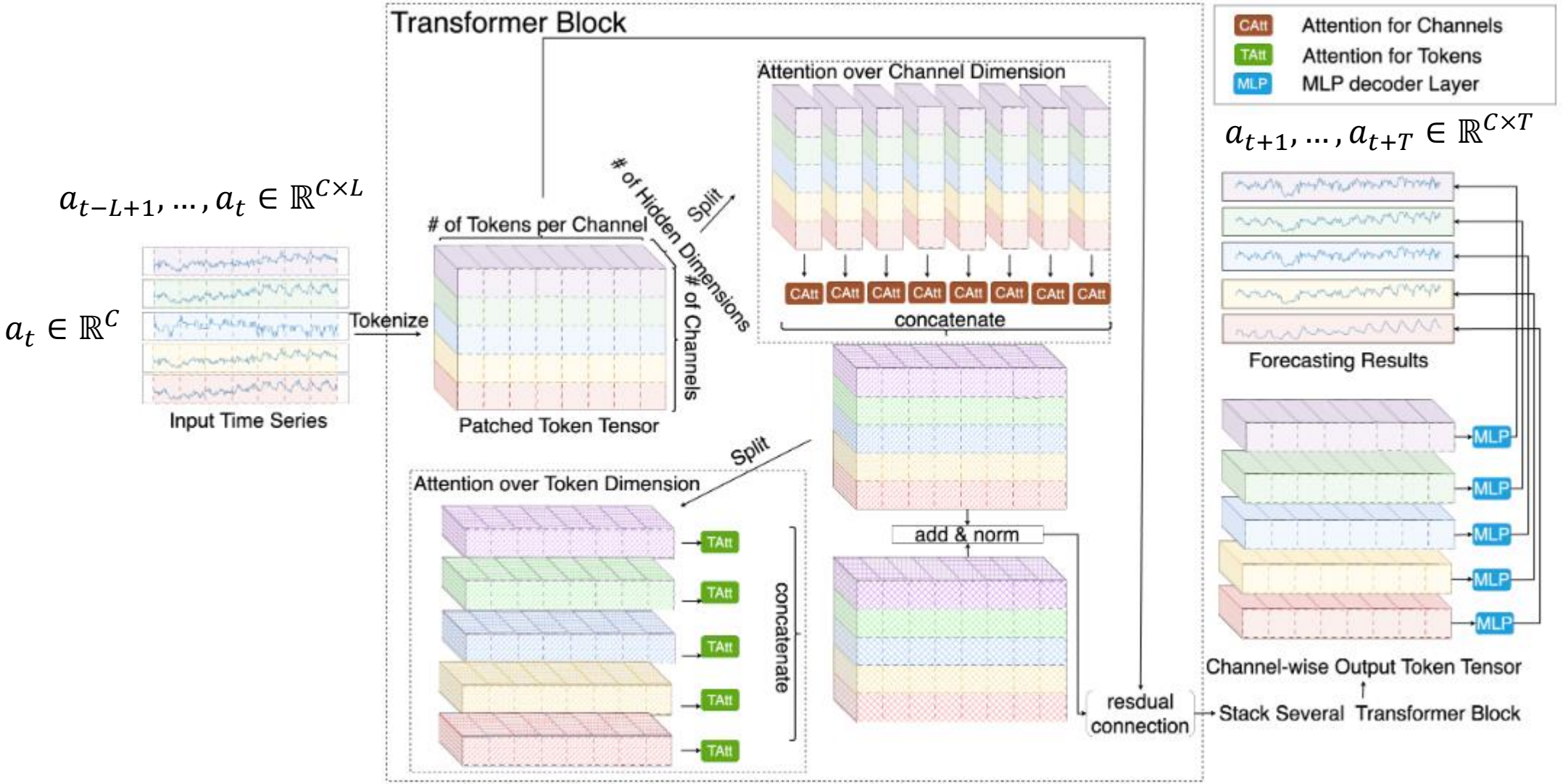
Patching示意图

- **Patched Transformers（2023 ICLR）**
  - PatchTST：率先提出了CI的思想，并使用patch取得良好效果
  - CrossFormer：patch+使用多层交叉注意力机制
  - Patch-wise Attention BETTER THAN Point-wise Attention



CrossFormer：在ETTh1数据集上训练两层Transformer的自注意力分数

系统整体Overview



模型结构：Tokenization

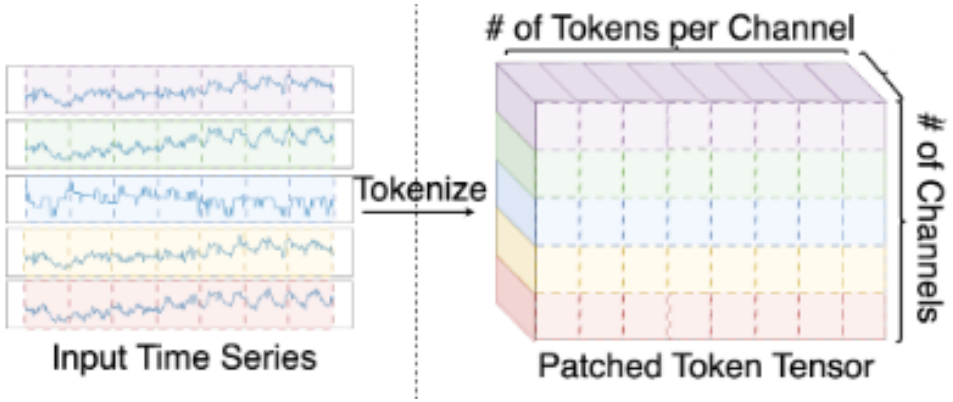
- **分段：** $\boldsymbol{A} = [\boldsymbol{a}_{t-L+1}, \ldots, \boldsymbol{a}_t] \in \mathbb{R}^{C \times L} \xrightarrow{\text{Patching}} \tilde{\boldsymbol{X}} \in \mathbb{R}^{C \times N \times P}$ 长度为L序列 ⟶ N个长度为P序列
- **Tokenization：** $X = [\mathbf{T_0}, F_1(\tilde{\boldsymbol{X}}) + \boldsymbol{E}], \quad X \in \mathbb{R}^{C \times (N+1) \times d}$

注入静态特征　　投影层　　Positional Encoding

MLP layer： $F_1 : P \to d$

Positional Encoding： $\boldsymbol{E} \in \mathbb{R}^{C \times N \times d}$

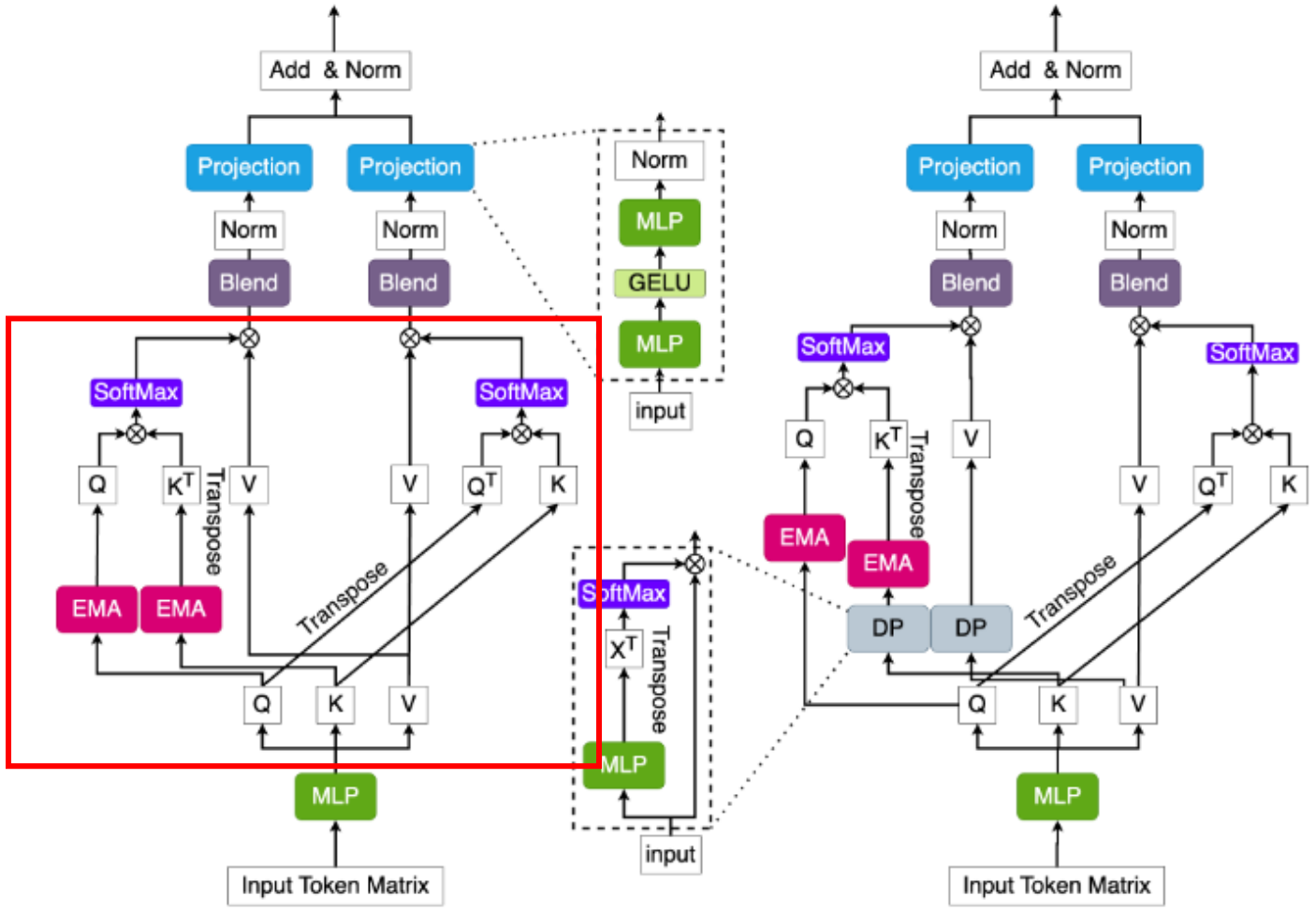
Extra token： $\mathbf{T_0} \in \mathbb{R}^{C \times d}$



$$Q = F_q(X), K = F_k(X), V = F_v(X), \quad Q, K, V \in \mathbb{R}^{C \times (N+1) \times d}$$

$$Q_i, K_i, V_i \in \mathbb{R}^{C \times (N+1) \times d_{\text{head}}}, \quad i = 1, 2, \ldots, H$$

模型结构：CARD Attention



模型结构：CARD Attention over Tokens

- **变量分离：** $Q_i, K_i, V_i \in \mathbb{R}^{C\times(N+1)\times d_{\text{head}}} \xrightarrow{\text{Slicing}} Q_i^{c:}, K_i^{c:}, V_i^{c:} \in \mathbb{R}^{(N+1)\times d_{\text{hcad}}},\ c=1,2,\ldots,C$
  - **Patch之间：**

$$A_{i1}^{c:} = \text{softmax}\left(\frac{1}{\sqrt{d}} \cdot \text{EMA}(Q_i^c)\left(\text{EMA}(K_i^c)\right)^{\top}\right),\quad A_{i1}^{c:} \in \mathbb{R}^{(N+1)\times(N+1)}$$

指数滑动平均：使得近期的数据点具有更高的权重

$$EMA(x_{t+1}) = \alpha x_{t+1} + (1-\alpha)EMA(x_t)$$

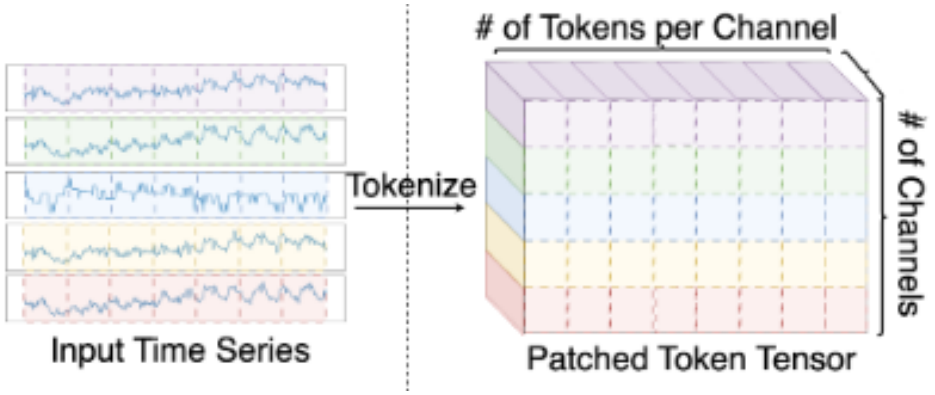
  - **隐藏层之间：**

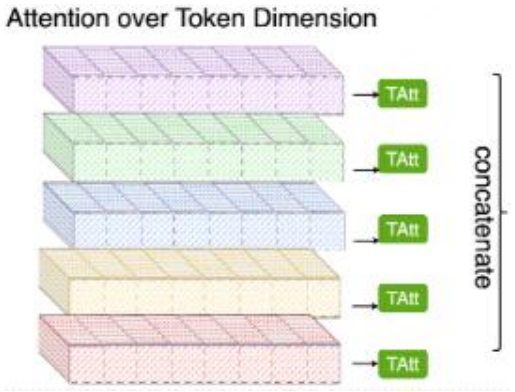
$$A_{i2}^{c:} = \text{softmax}\left(\frac{1}{\sqrt{N}} \cdot (Q_i^{c:})^{\top} K_i^c\right),\quad A_{i2}^{c:} \in \mathbb{R}^{d_{\text{head}}\times d_{\text{head}}}$$

$$\left.\begin{aligned} O_{i1}^{c:} &= A_{i1}^{c:} V_i^c \\ O_{i2}^{c:} &= V_i^{c:} A_{i2}^{c:} \end{aligned}\right\}$$

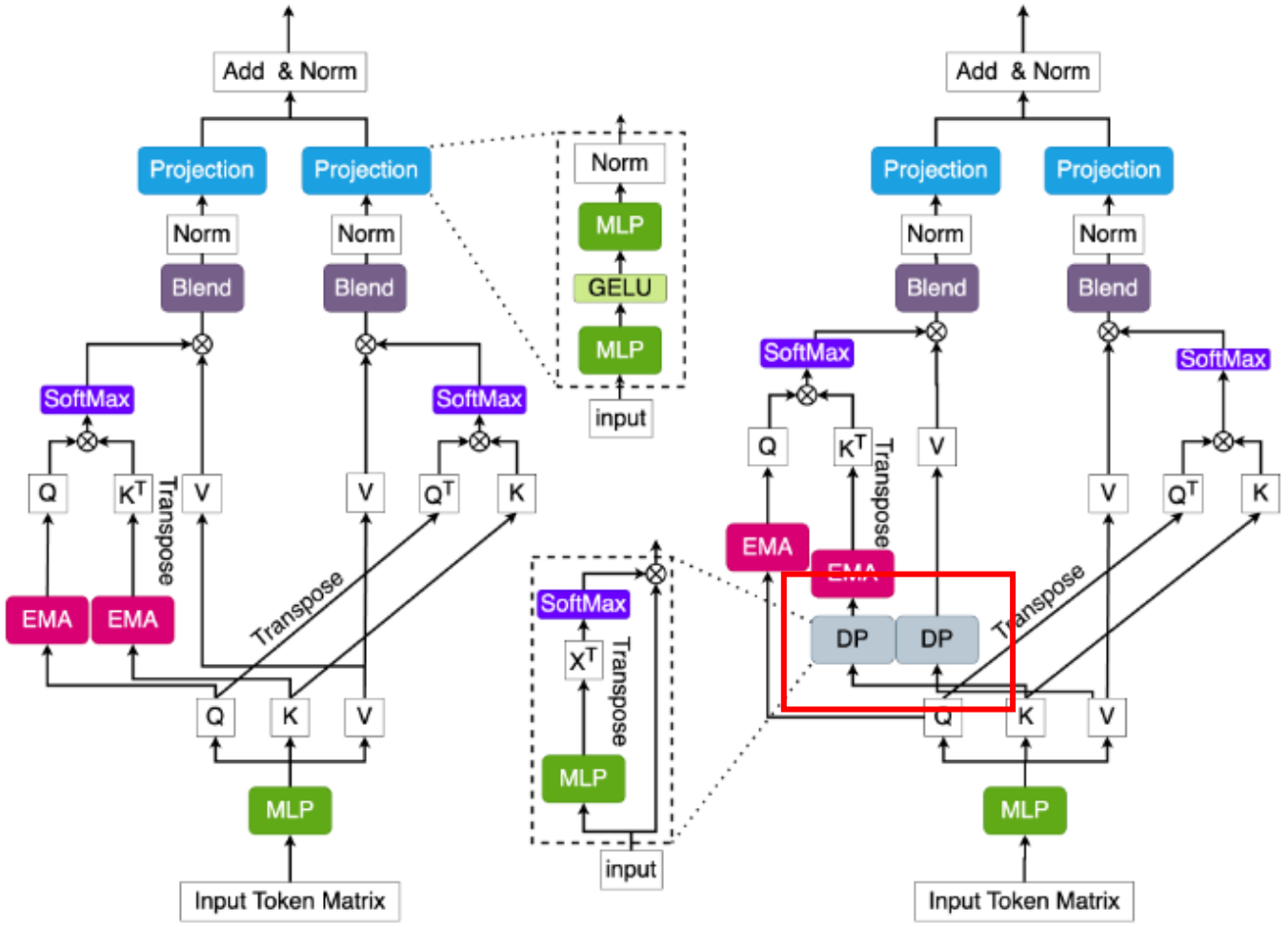
- **时间复杂度：**

$$\mathcal{O}(C \cdot d^2 \cdot L^2) \Rightarrow \mathcal{O}(C \cdot d^2 \cdot L^2/S^2)$$





模型结构：CARD Attention



模型结构：CARD Attention over Channels

- **切片分离：** $Q_i, K_i, V_i \in \mathbb{R}^{C\times(N+1)\times d_{\text{head}}} \xrightarrow{\text{Slicing}} Q_i^{:n}, K_i^{:n}, V_i^{:n} \in \mathbb{R}^{C\times d_{\text{head}}},\ n = 1, 2, \ldots, N+1$
- **动态投影：**

$$P_{ki}^{:n} = \text{softmax}(F_{pk}(K_i^{:n}))$$
$$P_{vi}^{:n} = \text{softmax}(F_{pv}(V_i^{:n}))$$

动态投影：计算K和V，利用低秩矩阵降维

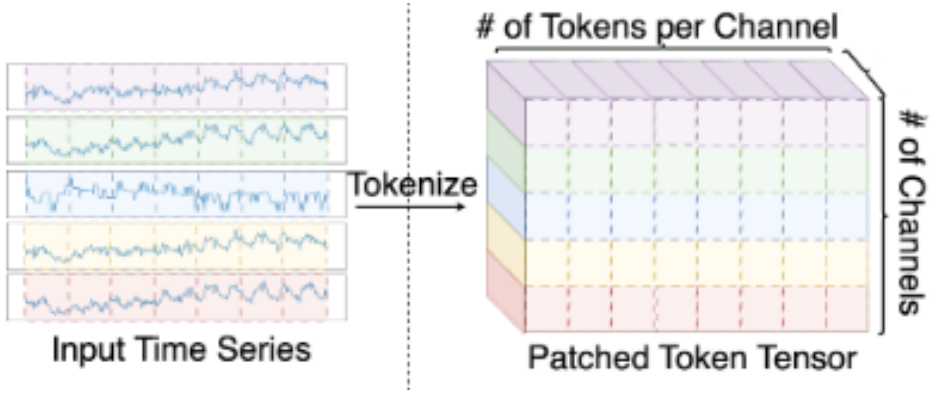
$F_{pk}, F_{pv} : d_{head} \to r, r \ll C$

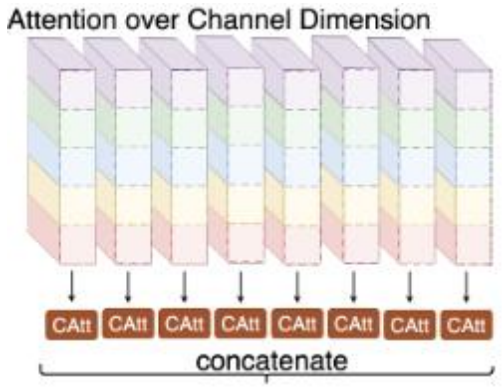
$$P_{ki}^{:n}, P_{vi}^{:n} \in \mathbb{R}^{C\times r}$$

$$\tilde{K}_i^{:n} = (P_{ki}^{:n})^\top K_i^{:n}, \quad \tilde{V}_i^{:n} = (P_{vi}^{:n})^\top V_i^{:n}$$

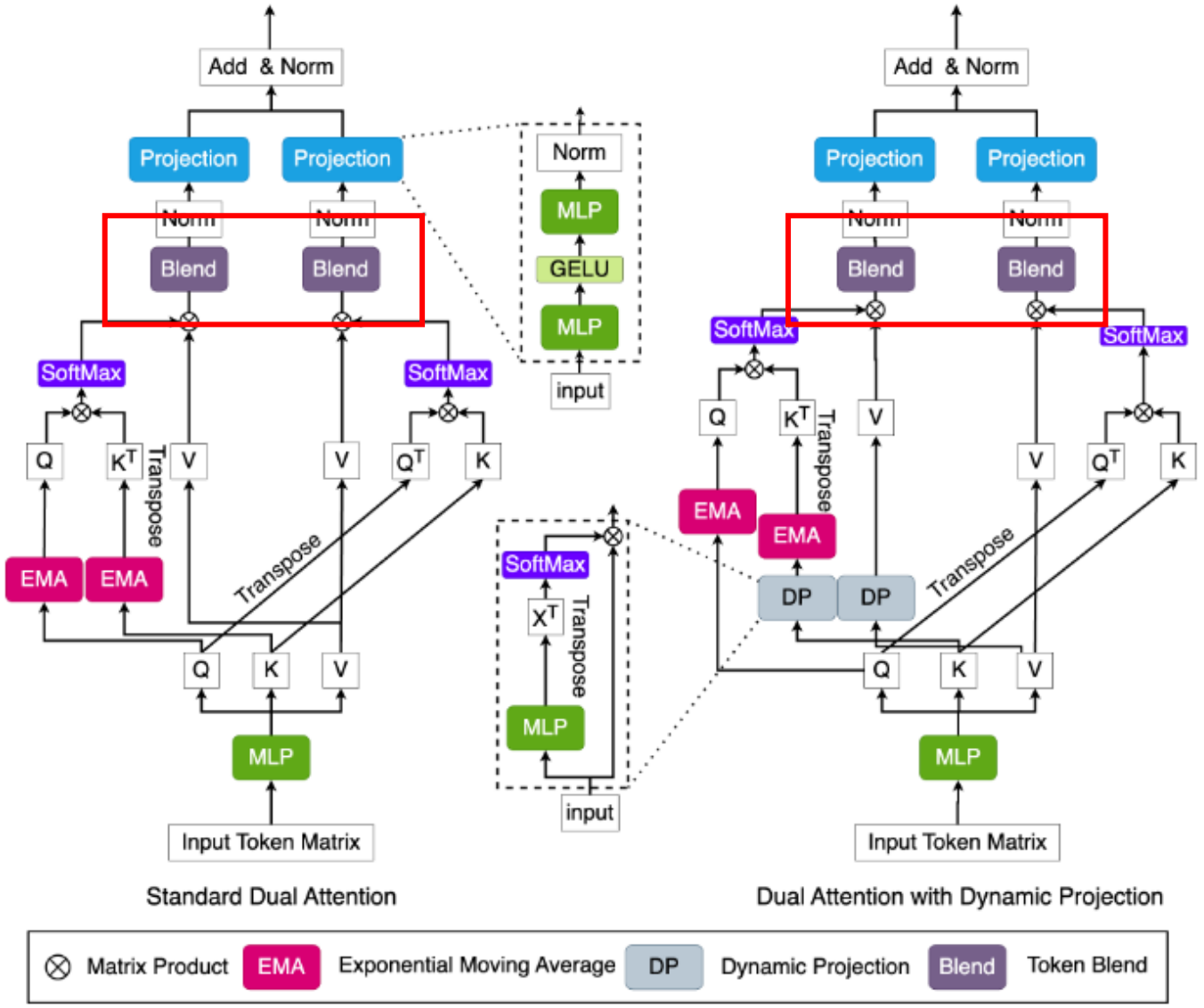
$$\tilde{K}_i^{:n}, \tilde{V}_i^{:n} \in \mathbb{R}^{r\times d_{\text{head}}}$$

- **时间复杂度：** $\mathcal{O}(L/S \cdot C \cdot r \cdot d^2) \to \mathcal{O}(L/S \cdot C^2 \cdot d^2)$



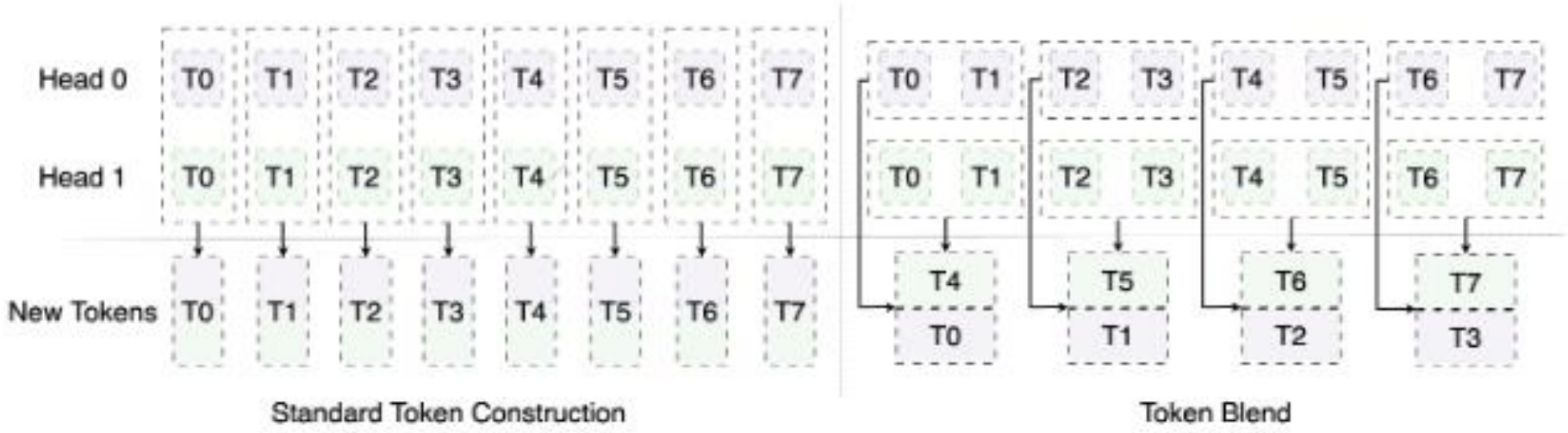


模型结构：CARD Attention



模型结构：Token Blend Module

- **多尺度信息——Token融合：**
  - **思想：**产生的新token包含了更长时间段内的信息



$$C \times H \times (N+1) \times d_{\text{head}}$$

$$\downarrow$$

$$C \times H(N+1) \times d_{\text{head}} \xrightarrow{\text{解耦}} H(N+1) \rightarrow h_1 \times h_2 \times h_3 \rightarrow \begin{cases} h_1 = H/h_3 \\ h_2 = N+1 \\ h_3 \geq 1 \quad \text{(blend size)} \end{cases}$$

模型结构：Signal Decay-Based Loss Function

- **基于信号衰减的损失函数**
  - 思路：历史信息对近未来观测的相关性 < 对远未来观测的相关性
  - 远未来观测值具有更高的方差 $Var(a_{t+l}) \preccurlyeq l\sigma^2 I$
  - 近期损失比远期损失对泛化性能的改进更大

- **MSE：**

$$\min \quad \mathbb{E}_{\boldsymbol{A}}\left[\frac{1}{L}\sum_{l=1}^{L}\|\hat{\boldsymbol{a}}_{t+l}(A)-\boldsymbol{a}_{t+l}(A)\|_2^2\right]$$

$$\mathbb{E}_{\boldsymbol{A}}\left[\frac{1}{L}\sum_{l=1}^{L} l^{-1}\|\hat{\boldsymbol{a}}_{t+l}(A)-\boldsymbol{a}_{t+l}(A)\|_2^2\right]$$

$l \in [t, t+L]$ 用于对远未来信息放缩

- **新损失函数：**

$$\min \mathbb{E}_{\boldsymbol{A}}\left[\frac{1}{L}\sum_{l=1}^{L} l^{-1/2}\|\hat{\boldsymbol{a}}_{t+l}(\boldsymbol{A})-\boldsymbol{a}_{t+l}(\boldsymbol{A})\|_1\right]$$

高斯分布换成拉普拉斯分布（MSE到MAE）
对异常值抗干扰能力更强

Experiments-1：长时预测任务

Table 1: Long-term forecasting tasks. The lookback length is set as 96. All models are evaluated on 4 different prediction horizons {96, 192, 336, 720} and average MSE/MAE results of ten repeats are reported. The best model is in boldface and the second best is underlined.

Models	CARD		PatchTST		MICN		TimesNet		Crossformer		Dlinear		LightTS		FiLM		ETSformer		FEDformer	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	**0.383**	**0.383**	0.395	0.408	0.387	0.411	0.400	0.406	0.435	0.417	0.403	0.407	0.435	0.437	0.408	0.399	0.429	0.425	0.448	0.452
ETTm2	**0.271**	**0.316**	0.283	0.327	0.284	0.340	0.291	0.333	0.609	0.521	0.350	0.401	0.409	0.436	0.287	0.328	0.292	0.342	0.305	0.349
ETTh1	0.443	**0.429**	0.455	0.444	**0.440**	0.462	0.458	0.450	0.486	0.481	0.456	0.452	0.491	0.479	0.461	0.456	0.452	0.510	**0.440**	0.460
ETTh2	**0.367**	**0.390**	0.384	0.406	0.402	0.437	0.414	0.427	0.966	0.690	0.559	0.515	0.602	0.543	0.384	0.406	0.439	0.452	0.437	0.449
Weather	**0.240**	**0.262**	0.257	0.280	0.243	0.299	0.259	0.287	0.250	0.310	0.265	0.317	0.261	0.312	0.269	0.339	0.271	0.334	0.309	0.360
Electricity	**0.169**	**0.258**	0.216	0.318	0.187	0.295	0.192	0.295	0.273	0.363	0.212	0.300	0.229	0.329	0.223	0.303	0.208	0.323	0.214	0.327
Traffic	**0.450**	**0.278**	0.488	0.327	0.542	0.316	0.620	0.336	0.593	0.332	0.625	0.383	0.622	0.392	0.639	0.389	0.621	0.396	0.610	0.376

Experiments-1：长时预测任务（lookback length 720

Models	CARD		PatchTST		MICN		TimesNet		Crossformer		Dlinear		LightTS		FilM		ETSformer		FEDformer	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	**0.350**	**0.368**	0.351	0.381	0.387	0.411	0.400	0.406	0.424	0.439	0.362	0.379	0.435	0.437	0.408	0.399	0.429	0.425	0.448	0.452
ETTm2	**0.254**	**0.310**	0.255	0.315	0.284	0.340	0.291	0.333	0.509	0.522	0.256	0.331	0.409	0.436	0.259	0.321	0.292	0.342	0.305	0.349
ETTh1	**0.401**	**0.421**	0.413	0.431	0.440	0.462	0.458	0.450	0.437	0.461	0.423	0.437	0.491	0.479	0.461	0.456	0.452	0.510	0.440	0.460
ETTh2	**0.321**	**0.373**	0.330	0.379	0.402	0.437	0.414	0.427	0.454	0.446	0.259	0.321	0.602	0.543	0.384	0.406	0.439	0.452	0.437	0.449
Weather	**0.219**	**0.248**	0.226	264	0.243	0.299	0.259	0.287	0.232	0.295	0.240	0.300	0.261	0.312	0.261	0.299	0.271	0.334	0.309	0.360
Electricity	**0.157**	**0.251**	0.159	0.253	0.187	0.295	0.192	0.295	0.280	0.343	0.177	0.224	0.229	0.329	0.194	0.290	0.208	0.323	0.214	0.327
Traffic	**0.381**	**0.251**	0.391	0.264	0.542	0.316	0.620	0.336	0.534	0.304	0.434	0.295	0.622	0.392	0.442	0.308	0.621	0.396	0.610	0.376

Experiments-2：基于重构的异常检测

Table 2: Anomaly detection. F1 scores are reported. The best model is in boldface and the second best is underlined.

Models	CARD	PatchTST	MICN	TimesNet	Crossformer	ETSformer	LightTS	Dlinear	FEDformer	Stationary	Autoformer	Informer
SMD	**0.872**	<u>0.866</u>	0.800	0.858	0.778	0.831	0.825	0.771	0.851	0.847	0.851	0.855
MSL	0.817	0.823	0.816	**0.852**	0.820	<u>0.850</u>	0.790	0.849	0.786	0.775	0.791	0.841
SMAP	**0.857**	0.695	0.656	<u>0.715</u>	0.674	0.695	0.692	0.693	0.708	0.711	0.711	0.699
SWaT	**0.945**	0.909	0.875	0.921	0.886	0.849	<u>0.933</u>	0.875	0.932	0.799	0.927	0.814
PSM	0.957	0.951	0.933	**0.975**	0.921	0.918	0.972	0.936	0.972	<u>0.973</u>	0.933	0.771
Avg	**0.890**	0.849	0.816	<u>0.864</u>	0.816	0.829	0.842	0.825	0.849	0.821	0.843	0.789

Experiments-3：基于信号衰减的损失函数

Table 3: Influence for signal decay-based loss function. The lookback length is set as 96. All models are evaluated on 4 different predication lengths {96, 192, 336, 720}. The average results are reported, and the full table is deferred to Table 19 in the Appendix. The model name with * uses the robust loss proposed in this work. The better results are in boldface.

Models	CARD		CARD*		MICN-regre		MICN-regre*		TimesNet		TimesNet*		FEDformer		FEDformer*		Autoformer		Autoformer*	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	0.390	0.399	**0.383**	**0.383**	0.392	0.414	**0.383**	**0.393**	0.400	0.406	**0.392**	**0.395**	0.448	0.452	**0.413**	**0.415**	0.588	0.528	**0.523**	**0.475**
ETTh1	0.449	0.440	**0.443**	**0.425**	0.559	0.535	**0.527**	**0.499**	0.458	0.450	**0.449**	**0.438**	0.440	0.460	**0.436**	**0.442**	**0.496**	0.487	0.514	**0.481**

Experiments-4：对Blend Size

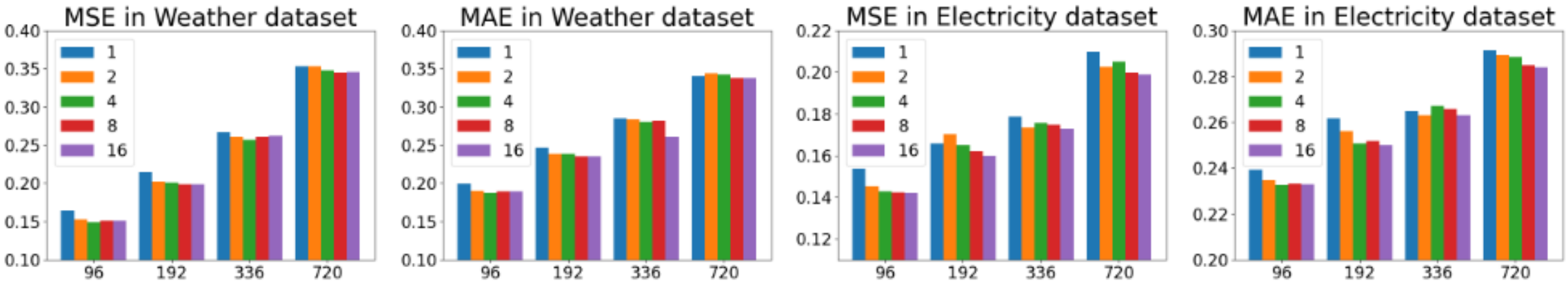


Figure 4: Experiments on token blend size. The blend size is varying in 1, 2, 4, 8, and 16.

Table 6: Model configurations of CARD.

Dataset	patch	stride	model dim	FFN dim	dropout	blend size	learning rate	warm-up	batch size
ETTm1	16	8	16	32	0.3	2	1e-4	0	128
ETTm2	16	8	16	32	0.3	2	1e-4	0	128
ETTh1	16	8	16	32	0.3	2	1e-4	0	128
ETTh2	16	8	16	32	0.3	2	1e-4	0	128
Weather	16	8	128	256	0.2	16	1e-4	0	128
Electricity	16	8	128	256	0.2	16	1e-4	20	32
Traffic	16	8	128	256	0.2	16	1e-4	20	24

Experiments-5：注意力图

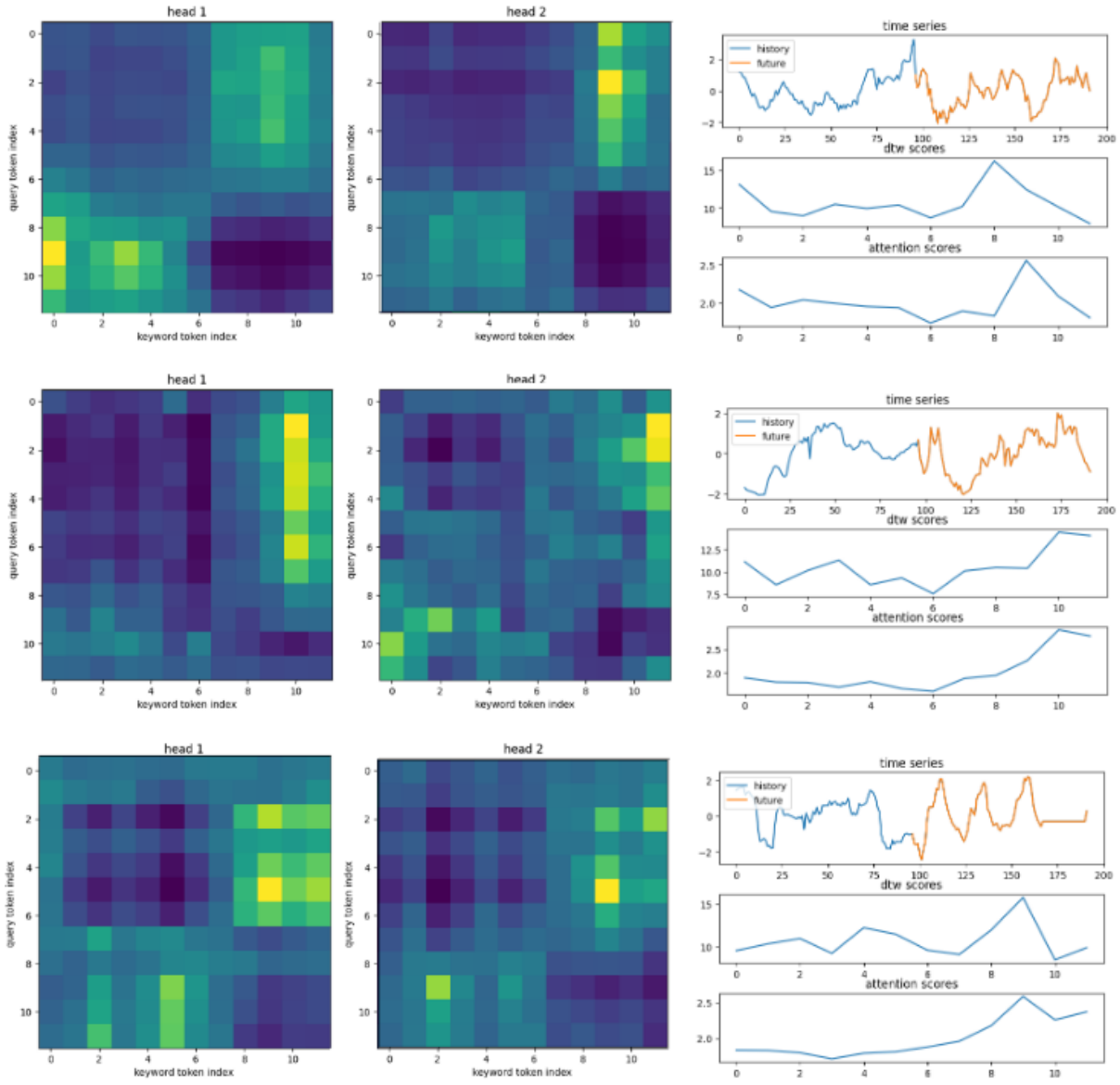


Figure 36: Attention Map Samples of ETTh1 task.

- 注意力图平滑；
- 注意力得分与DTW得分正相关；

Experiments-6：消融实验（注意力层设计）

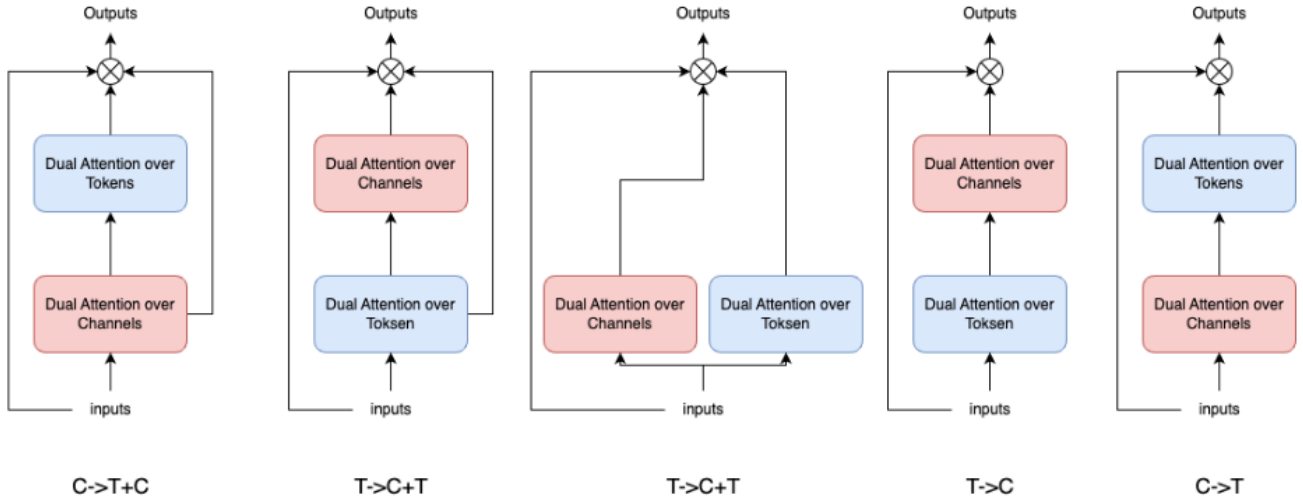
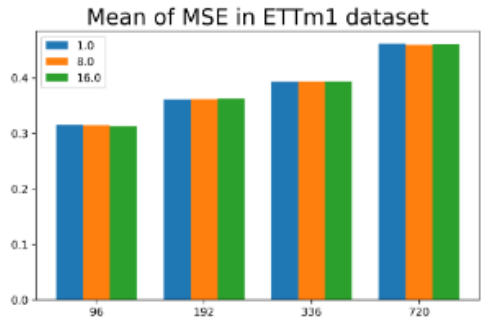


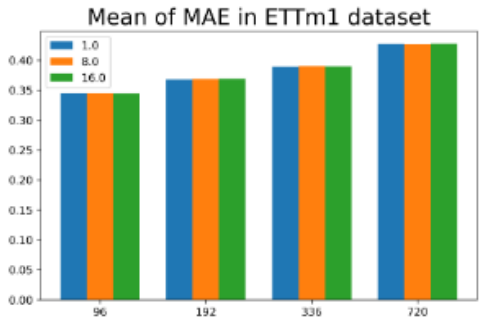
Figure 38: Architecture Variants

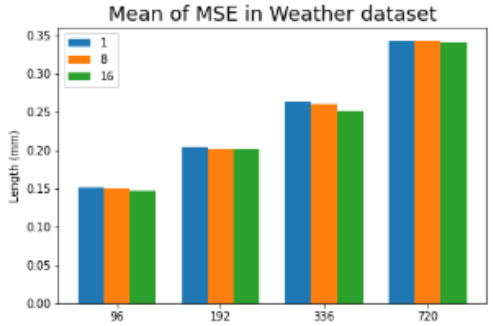
Table 25: Model variants. All models are evaluated on 4 different predication lengths $\{96, 192, 336, 720\}$. The best results are in boldface.

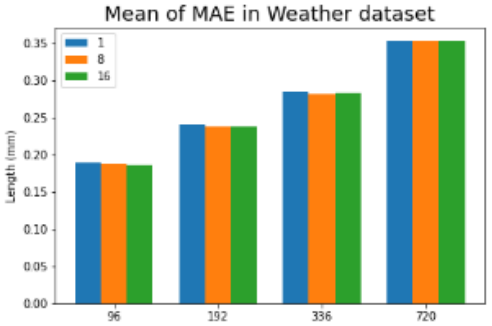
Models		c->t+c (CARD)		t->c+t		t+c		t->c		c->t	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	**0.316**	**0.347**	0.318	0.346	0.318	0.346	0.326	0.363	0.334	0.368
	192	**0.363**	**0.370**	0.367	0.370	0.366	0.369	0.366	0.385	0.372	0.387
	336	**0.393**	**0.390**	0.399	0.391	0.396	0.391	0.400	0.404	0.401	0.407
	720	**0.458**	**0.426**	0.466	0.429	0.463	0.428	0.459	0.440	0.458	0.438
	avg	**0.383**	**0.384**	0.388	0.384	0.386	0.384	0.388	0.398	0.391	0.400
Weather	96	**0.150**	**0.188**	0.153	0.193	0.152	0.189	0.152	0.191	0.152	0.192
	192	0.202	0.238	0.203	0.239	**0.201**	**0.236**	**0.201**	0.239	0.203	0.240
	336	**0.260**	0.282	0.269	0.288	0.261	**0.281**	0.263	0.284	0.262	0.284
	720	**0.343**	**0.335**	0.345	0.339	0.344	0.337	0.347	0.339	0.344	0.337
	avg	**0.239**	**0.261**	0.243	0.265	0.240	**0.261**	0.241	0.263	0.240	0.263

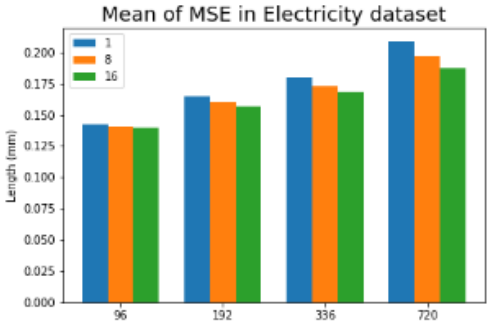
Experiments-7：消融实验（动态投影维度）

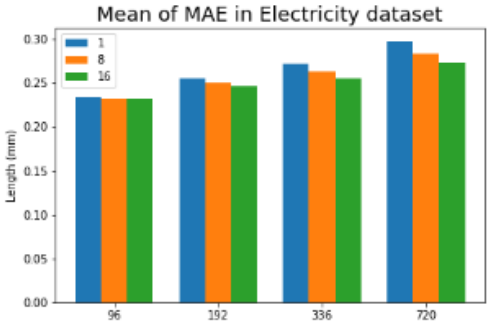


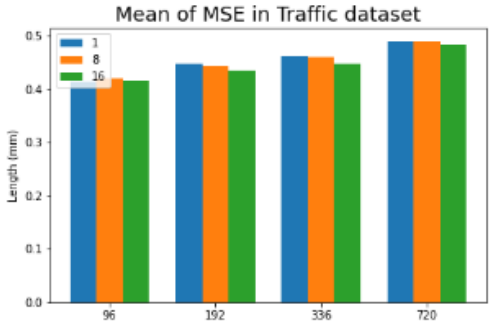












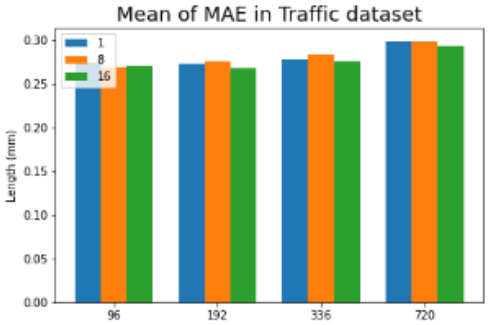
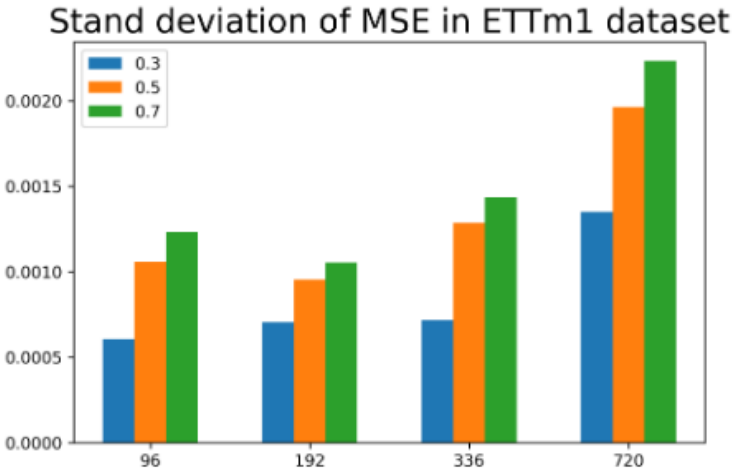
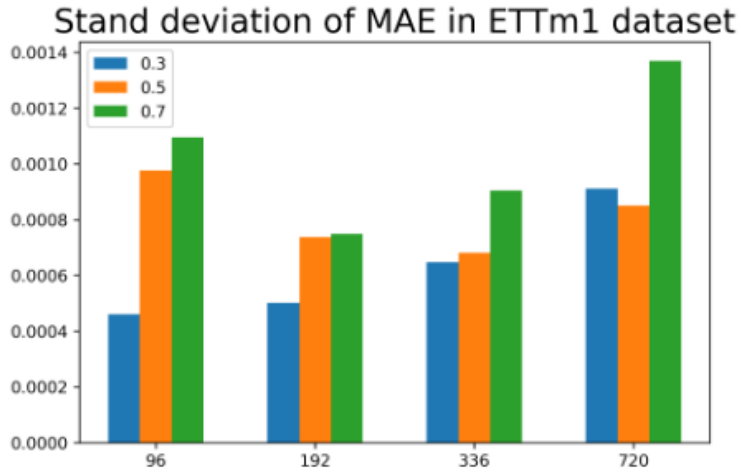
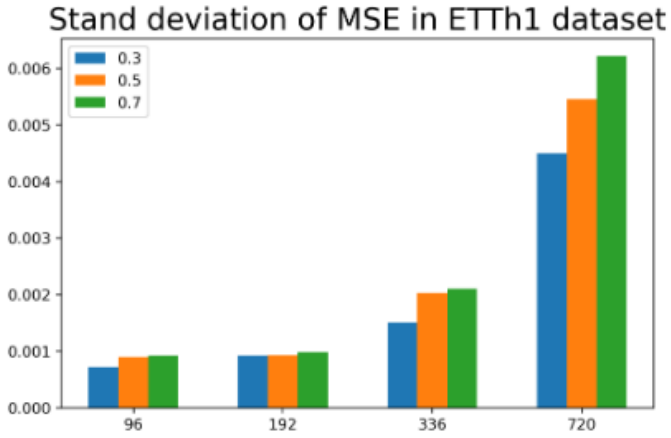


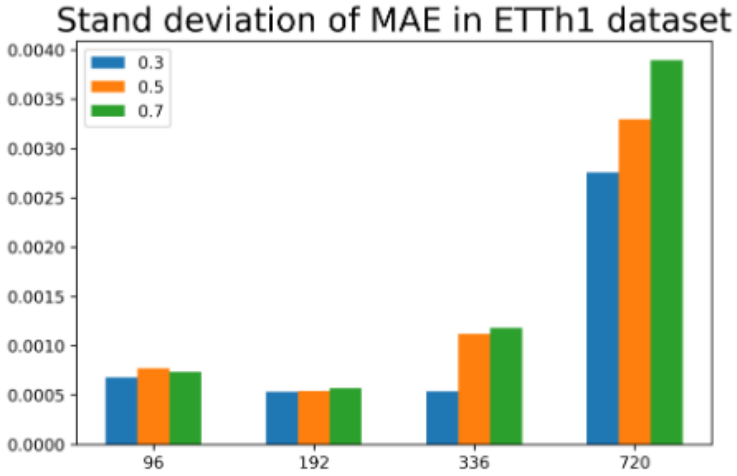
Figure 39: Experiments on dynamic projection dimensions. The projection dimension is varying in 1, 8, and 16.

Experiments-8：消融实验（EMA参数）









- EMA参数越小，平滑程度越重

指数滑动平均：使得近期的数据点具有更高的权重

$$EMA(x_{t+1}) = \alpha x_{t+1} + (1-\alpha) EMA(x_t)$$

感谢各位聆听！

敬请批评指正！