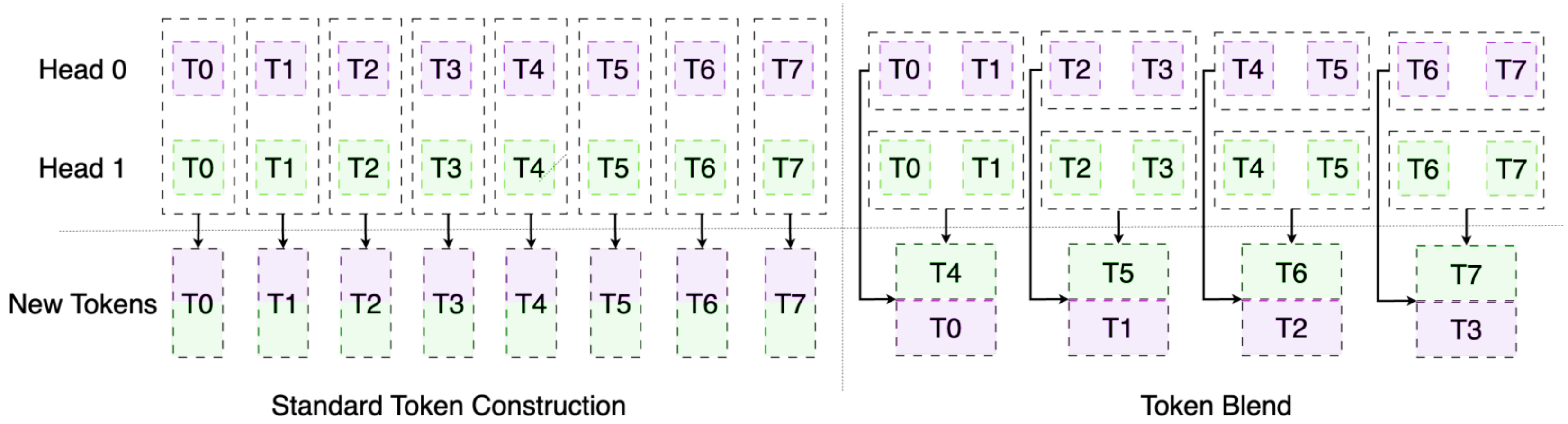
2024 ICLR CCF-A

CARD

Channel **A**ligned **R**obust Blen**d** Transformer For Time Series Forecasting

Make Transformer Great Again

for Time Series Forecasting: Channel Aligned Robust Dual Transformer



2024 ICLR CCF-A

CARD

Channel Aligned Robust Blend Transformer For Time Series Forecasting

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

（Make Transformer Great Again for Time Series Forecasting: Channel Aligned Robust Dual Transformer）

24.4.30

Presented by Yyyq

01 问题描述

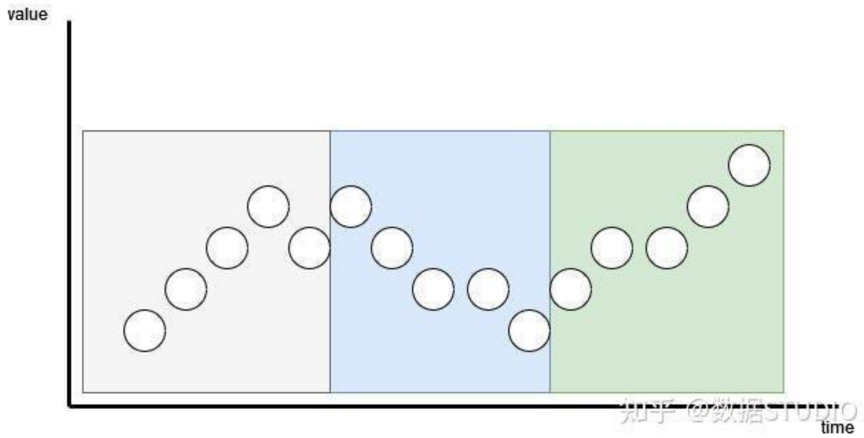
- **channel-dependent (CD) & channel-independent (CI)**
  - CD利用不同预测变量之间的依赖关系：多变量时序预测
  - CI提高训练鲁棒性：PatchTST 和 Dlinear
  - CI 更具有鲁棒性，CD 更高的建模能力
- **Transformer For Time Series Forecasting**
  - 有效利用信道（即预测变量）之间的依赖性
  - 缓解时间序列预测中的过拟合噪声问题

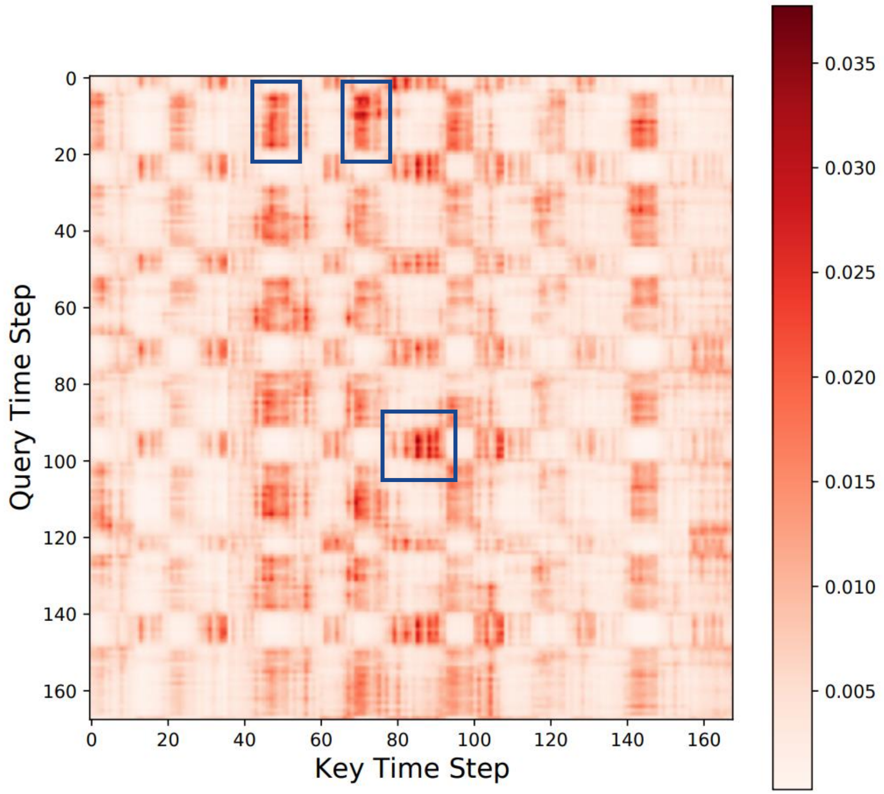
02 创新点

- **跨通道信息，变量间相关性**
  - Patch-token：在每个token内对齐局部信息
  - Attention关注不同的通道和隐藏维度
  - Token混合模块：将同一注意力头内的相邻token合并为新token
- **提高鲁棒性，减轻过拟合噪声的问题**
  - 注意力机制中加入：指数平滑层、动态投影模块
  - 基于信号衰减的鲁棒损失函数

03 相关工作

- **Patched Transformers**
  - NLP：基于子词的标记化（优于字符）
  - CV：将图像分割成小块
  - 语音领域：原始音频的子序列级别信息
- **PatchTST 和 CrossFormer** （ICLR2023）
  - Patch-wise Attention 优于 Point-wise Attention





04

模型结构：Overview

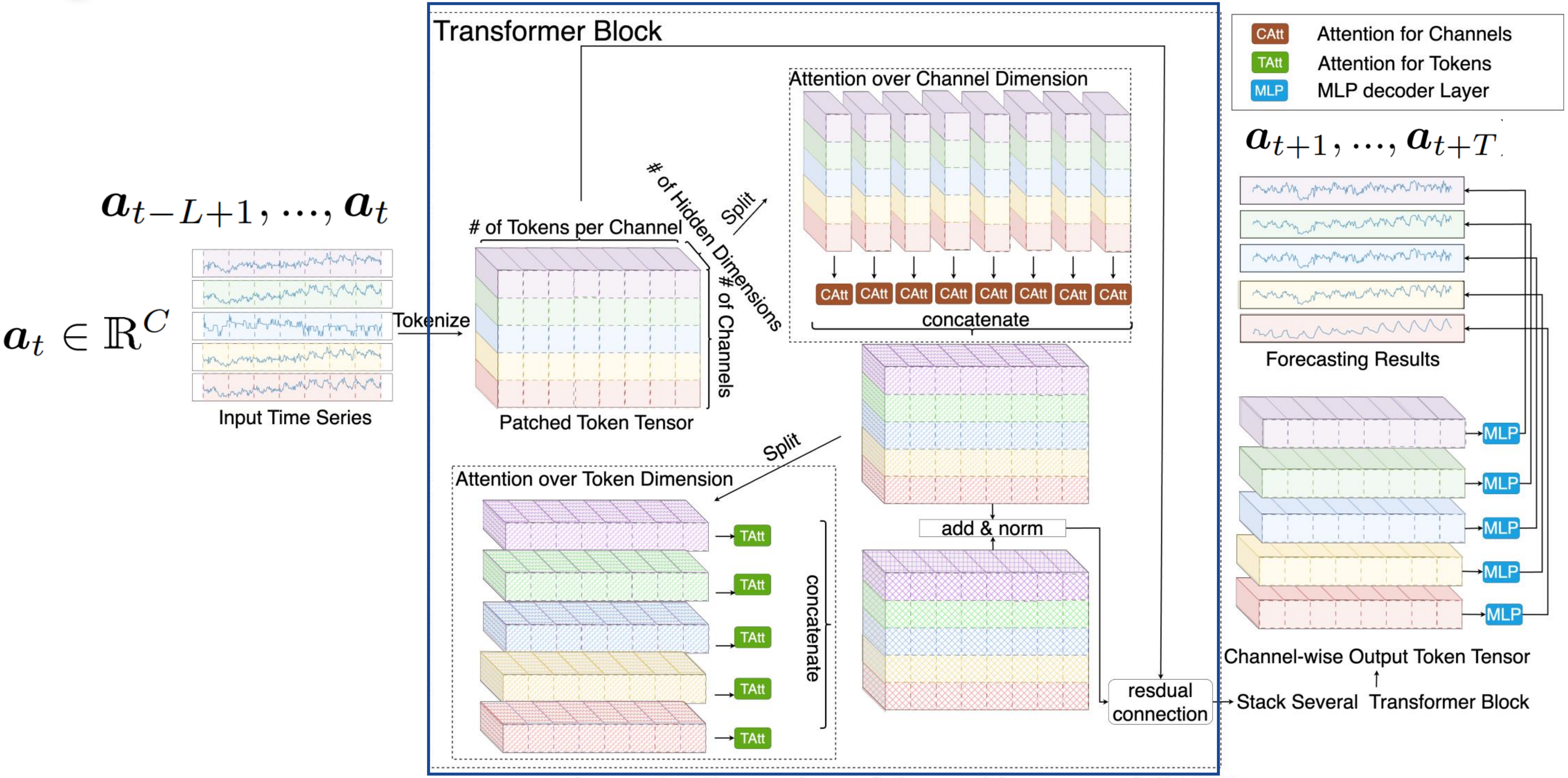


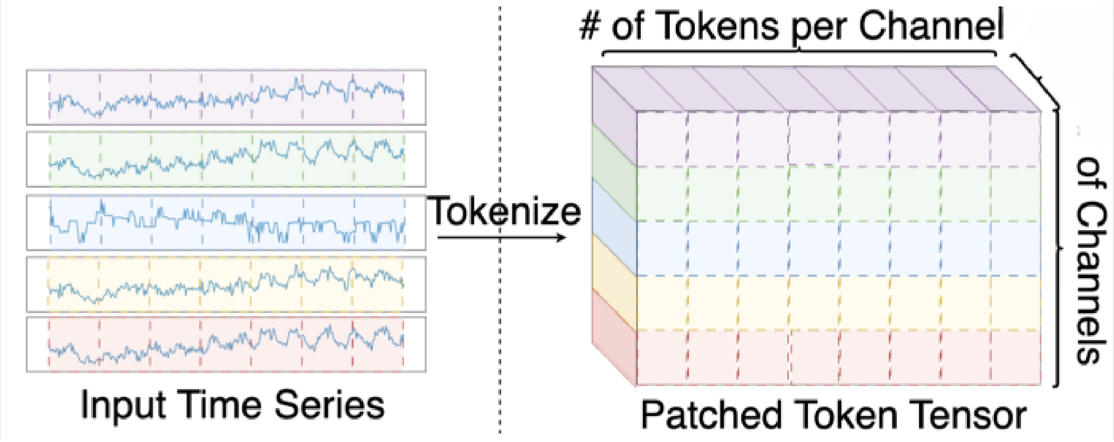
Figure 1: Illustration of the architecture of CARD.

04 模型结构：Tokenization

$$\boldsymbol{A} = [\boldsymbol{a}_{t-L+1}, ..., \boldsymbol{a}_t] \in \mathbb{R}^{C\times L} \xrightarrow{\textbf{Patching}} \tilde{X} \in \mathbb{R}^{C\times N\times P}$$ （L序列长度 → N个P长度）

$$\Rightarrow \boldsymbol{X} = [\mathbf{T_0}, F_1(\tilde{\boldsymbol{X}}) + \boldsymbol{E}],\ \boldsymbol{X} \in \mathbb{R}^{C\times(N+1)\times d}$$

$$\begin{cases} \text{MLP layer } F_1 : P \to d \\ \text{positional embedding } \boldsymbol{E} \in \mathbb{R}^{C\times N\times d} \\ \text{extra token } \mathbf{T}_0 \in \mathbb{R}^{C\times d} \end{cases}$$



$$\Rightarrow \boldsymbol{Q} = F_q(\boldsymbol{X}),\ \boldsymbol{K} = F_k(\boldsymbol{X}),\ \boldsymbol{V} = F_v(\boldsymbol{X}),$$

$$\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i \in \mathbb{R}^{C\times(N+1)\times d_{\text{head}}},\ i = 1, 2, ..., H.$$

04 模型结构：CARD Attention Over Tokens

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

$$\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i \in \mathbb{R}^{C\times(N+1)\times d_{\text{head}}},\ i = 1,2,...,H.$$

- 把每个变量分开： $Q_i^{c:}, K_i^{c:}, V_i^{c:} \in \mathbb{R}^{(N+1)\times d_{\text{head}}}$ and $c = 1,2,...,C.$
  - Patch之间：

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

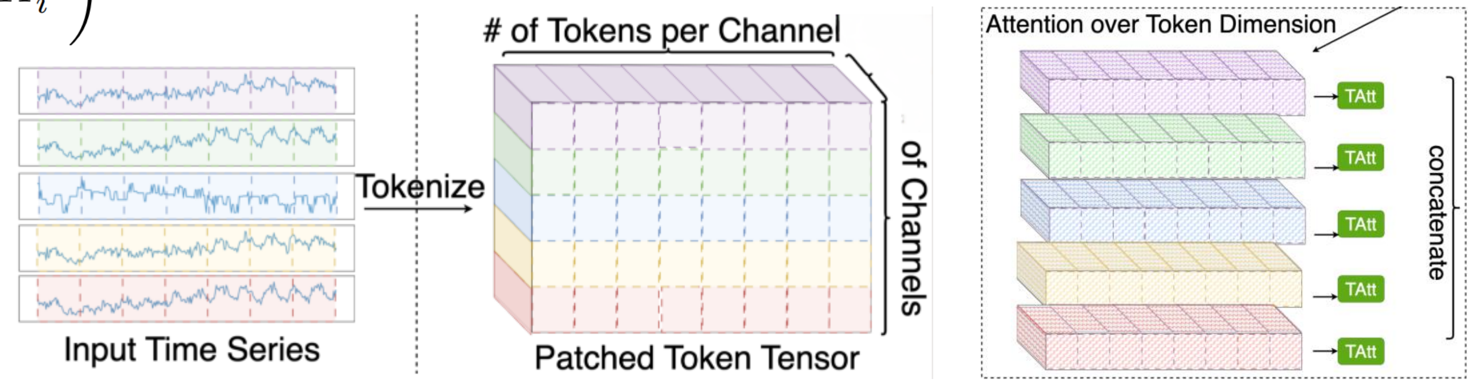
指数滑动平均：对前t个时刻的数据加权平均，时间越近权重越大

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

  - 隐藏维度之间：

$$\boldsymbol{A}_{i2}^{c:} = \text{softmax}\left(\frac{1}{\sqrt{N}} \cdot (Q_i^{c:})^\top K_i^{c:}\right)$$

$$\boldsymbol{A}_{i2}^{c:} \in \mathbb{R}^{d_{\text{head}}\times d_{\text{head}}}$$



04 模型结构：CARD Attention Over Tokens

$$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}})V$$

$$\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i\in\mathbb{R}^{C\times(N+1)\times d_{\text{head}}},\ i=1,2,...,H.$$

- 按每个变量分片（时序内相关性）： $Q_i^{c:},K_i^{c:},V_i^{c:}\in\mathbb{R}^{(N+1)\times d_{\text{head}}}$ and $c=1,2,...,C.$
  - Patch之间：

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

  - 隐藏维度之间：

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

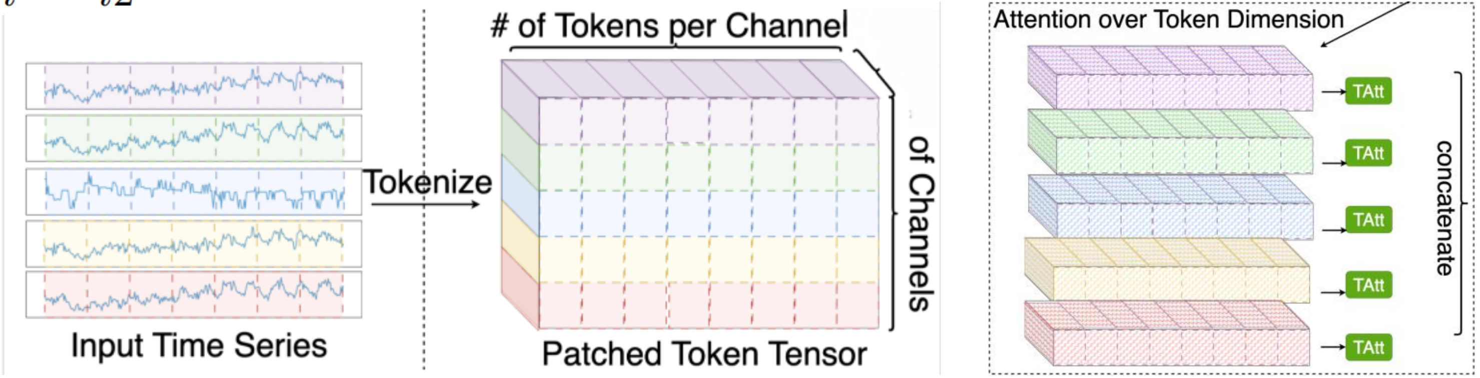
  - 输出：

$$\boldsymbol{O}_{i1}^{c:}=\boldsymbol{A}_{i1}^{c:}\boldsymbol{V}_i^{c},\quad \boldsymbol{O}_{i2}^{c:}=\boldsymbol{V}_i^{c:}\boldsymbol{A}_{i2}^{c:}$$

- 时间复杂度

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

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



04 模型结构：CARD Attention Over Channels

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

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

- 按Patch分片（时序间相关性）： $Q_i^{:n}, K_i^{:n}, V_i^{:n} \in \mathbb{R}^{C\times d_{\text{head}}}$ and $n = 1, 2, ..., N+1.$
  - 基于动态投影计算K和V：利用低秩矩阵将节点数量维度 C 投影至更低维 r

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

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

$$d_{head} \to r << C,$$

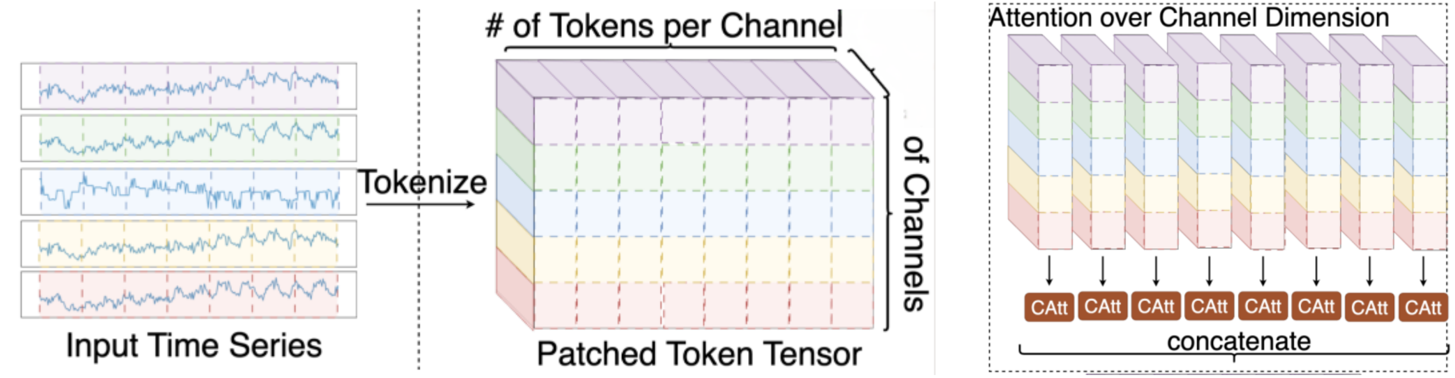
$$P_{ki}^{:n}, P_{vi}^{:n} \in \mathbb{R}^{C\times r}$$

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

- 时间复杂度

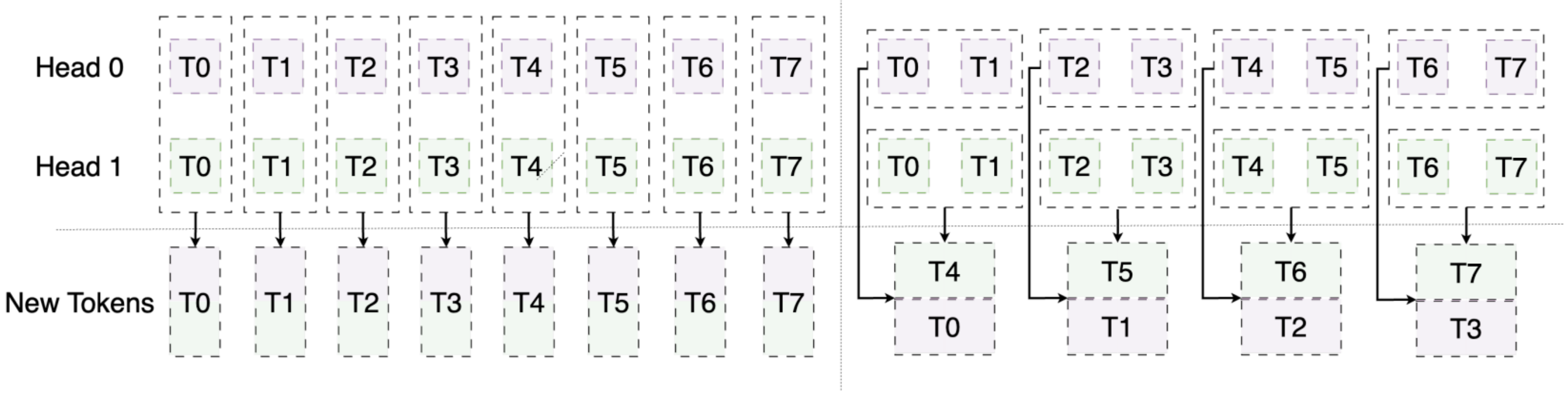
$$\mathcal{O}(L/S\cdot C^2\cdot d^2)$$

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



04 模型结构：Token Blend Module

- Token混合多尺度信息



Standard Token Construction

Token Blend

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

$$\Downarrow$$

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

← Blend size

04 模型结构：损失函数

- **基于信号衰减的损失函数**
  - 历史信息与远未来观测值的相关性 < 与近未来观测值的相关性
  - 远未来观测值具有更高的方差 $\mathrm{var}(\boldsymbol{a}_{t+l}) \preceq l\sigma^2 I$
  - 近期损失比远期损失对泛化改进的贡献更大

$$\min \quad \mathbb{E}_{\boldsymbol{A}}\left[\frac{1}{L}\sum_{l=1}^{L}\|\hat{\boldsymbol{a}}_{t+l}(\boldsymbol{A})-\boldsymbol{a}_{t+l}(\boldsymbol{A})\|_2^2\right]. \Rightarrow \min \mathbb{E}_{\boldsymbol{A}}\left[\frac{1}{L}\sum_{l=1}^{L}\underline{l^{-1/2}}\|\hat{\boldsymbol{a}}_{t+l}(\boldsymbol{A})-\boldsymbol{a}_{t+l}(\boldsymbol{A})\|_1\right]$$

$$l \in [t, t+L]$$

05

实验1：对比实验（回望窗口96）

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

05

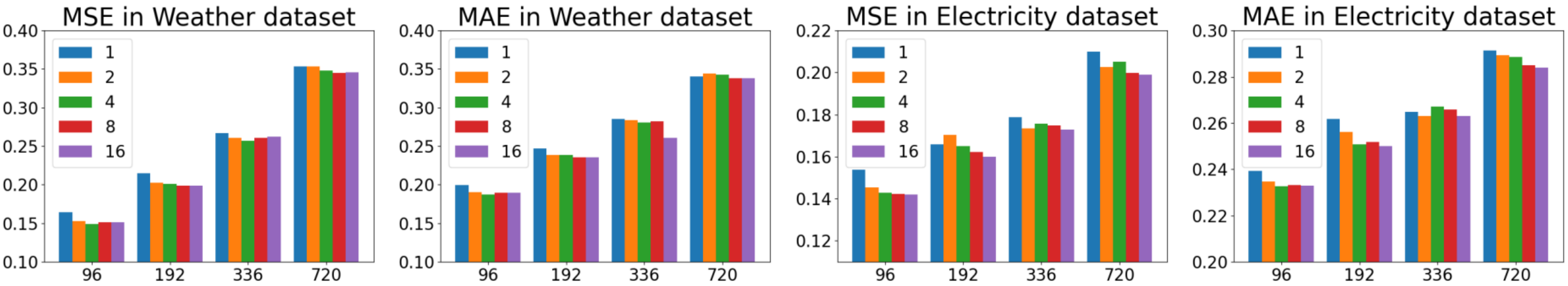
实验1：对比实验（回望窗口720）

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

05 实验2：信号衰减损失函数

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**

05 实验3：token的blend size



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

05 实验4：注意力图

- 注意力图是平滑的
- 注意力得分总和与DTW得分呈正相关

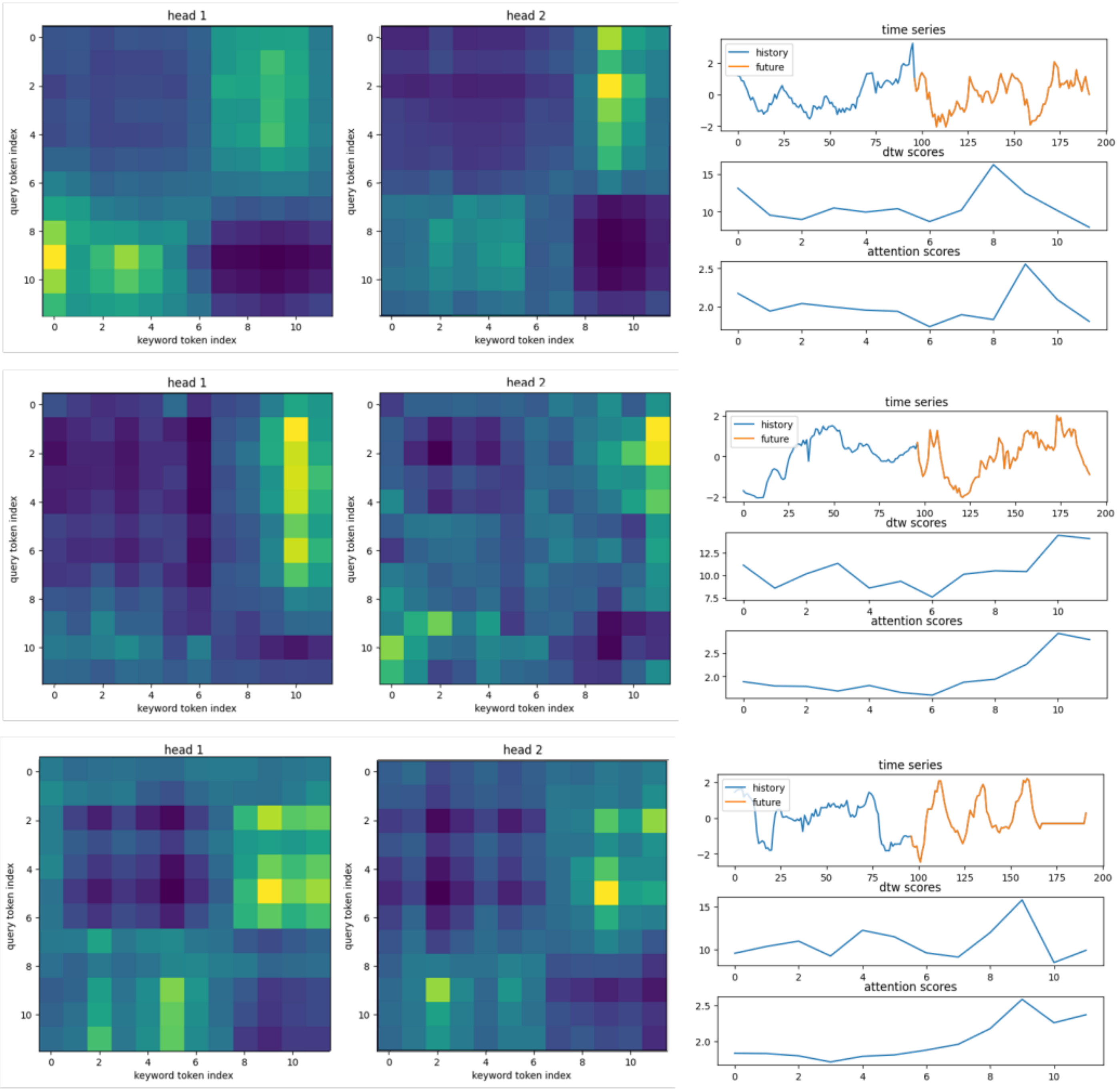
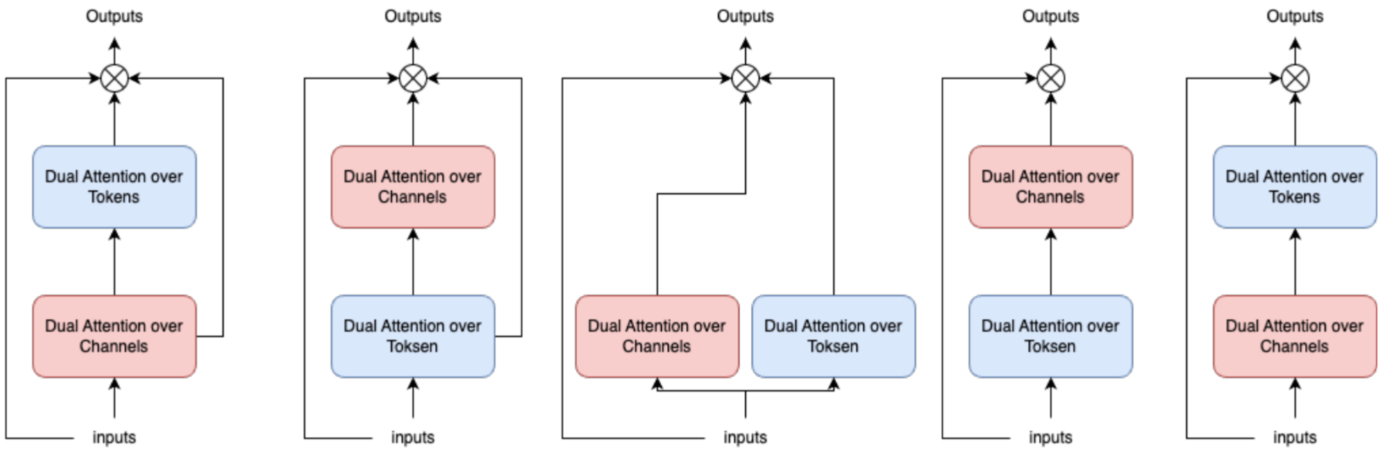


Figure 36: Attention Map Samples of ETTh1 task.

实验5：消融实验——注意力层的设计



C->T+C　　T->C+T　　T->C+T　　T->C　　C->T

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

05 实验5：消融实验——动态投影维度的选择

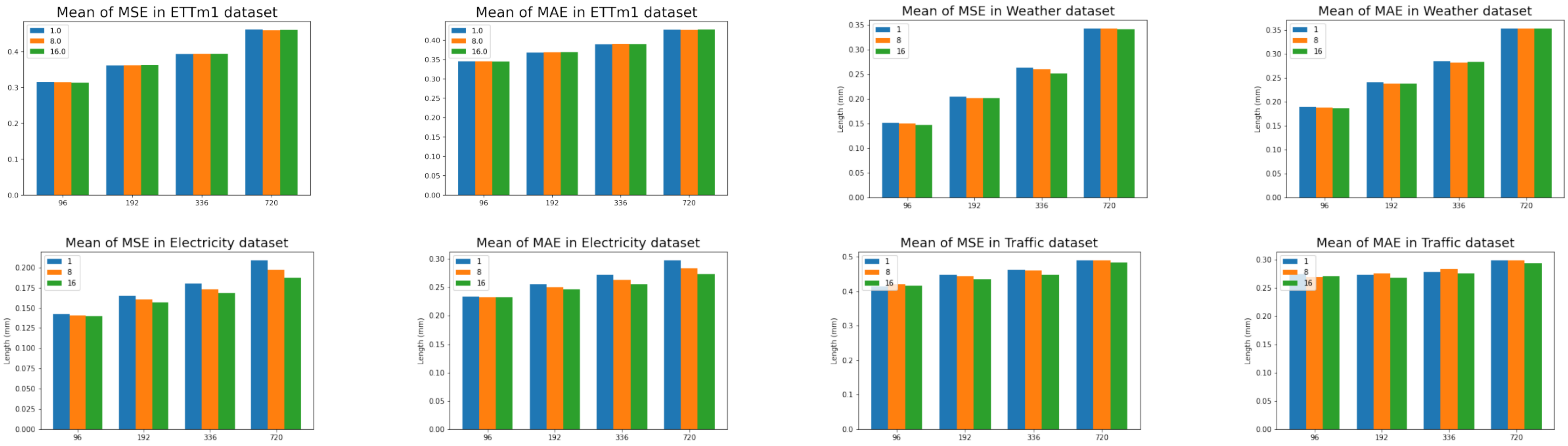
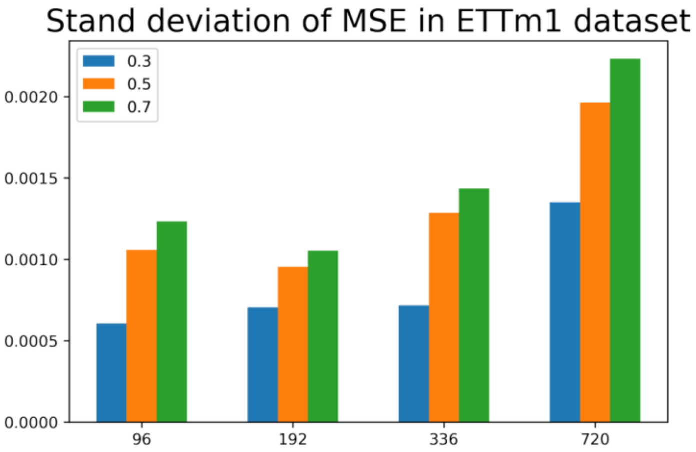
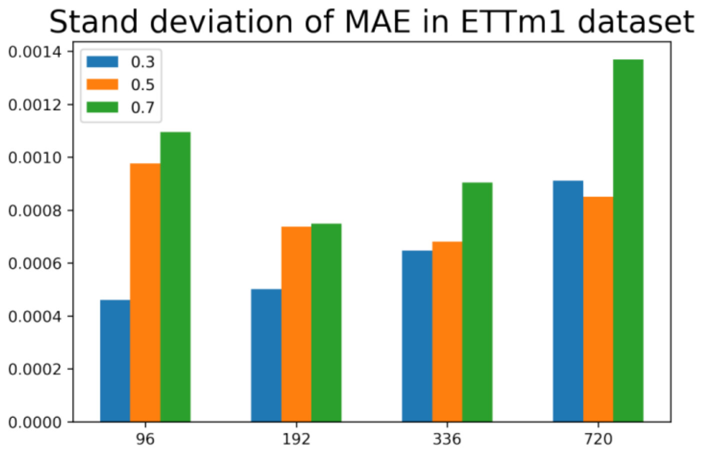
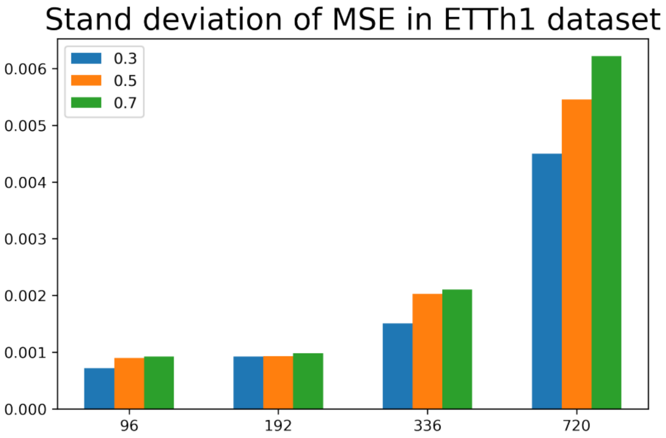


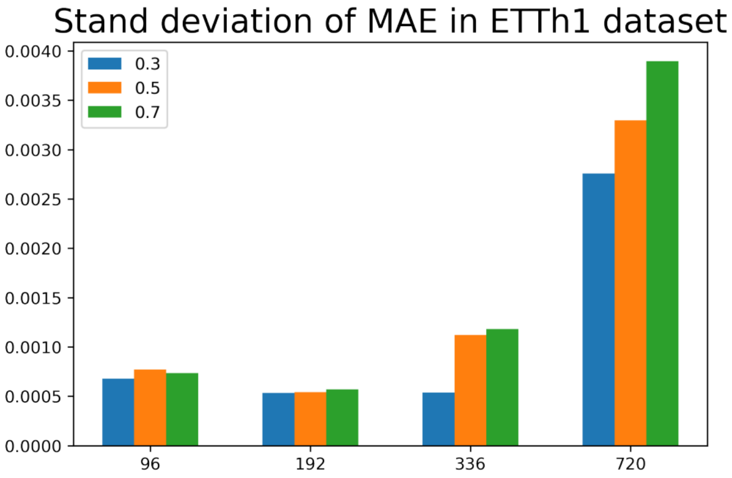
Figure 39: Experiments on dynamic projection dimensions. The projection dimension is varying in 1, 8, and 16.

05 实验5：消融实验——EMA的平滑参数









指数滑动平均：对前t个时刻的数据加权平均，时间越近权重越大

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

谢谢观看

MANY THANKS !

24.4.30