An Inverse-Distance Weighting Alignment-Diversity Attention Transformer for Days Ahead PM Concentration Forecasting

Supplementary Material

Yue Yu[1], Pavel Loskot[2], Yu Lin[2], Yu Gao[3]

[1] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China

[2] Zhejiang University - University of Illinois Urbana-Champaign Institute, Zhejiang University, Haining 314400, China

[3] AI Research Center, Midea Group, Shanghai 201103, China

S1. Kernel attention

Table S1 compares the complexity of the standard attention, and the proposed attention employing the CAK and DCK kernels. In the standard attention, the main computational cost arises from computing the scaled dot-product, $\mathbf{QK}^\top$, and the subsequent multiplication with $\mathbf{V}$, so the overall complexity (Big-O notation) is, $O(Bn_{\text{head}}N^2d_{\text{head}}) = O(BN^2D)$. In the kernel based attention, the CAK replaces the dot-product with a kernel-based similarity, which requires $O(Bn_{\text{head}}N^2)$ operations. The subsequent multiplication with $\mathbf{V}$ requires $O(Bn_{\text{head}}N^2d_{\text{head}})$ operations. Since $D \gg n_{\text{head}}$, the overall complexity remains $O(BN^2D)$.

The DCK constructs a richer composite kernel, which has the complexity, $O(Bn_{\text{head}}N^2d_{\text{head}})$. The multiplication with $\mathbf{V}$ requires $O(Bn_{\text{head}}N^2d_{\text{head}})$ operations, so the complexity of the DCK is also $O(BN^2D)$. Consequently, the overall asymptotic complexity of the proposed kernel-based attention is comparable to the complexity of the standard attention, whereas the former provides enhanced accuracy for predicting the multivariate time series data.

Table S1. The Big-O complexity of the standard attention, and of the kernel-based attention.

Method	**Main steps**	**Complexity**
Standard Attention	$\mathbf{QK}^\top$: $O(Bn_{\text{head}}N^2d_{\text{head}})$ Multiplication with $\mathbf{V}$: $O(Bn_{\text{head}}N^2d_{\text{head}})$	$O(BN^2D)$
CAK	Building CAK: $O(Bn_{\text{head}}N^2)$ Multiplication with $\mathbf{V}$: $O(Bn_{\text{head}}N^2d_{\text{head}})$	$O(BN^2D)$ $(D \gg n_{\text{head}})$
DCK	Building DCK: $O(Bn_{\text{head}}N^2d_{\text{head}})$ Multiplication with $\mathbf{V}$: $O(Bn_{\text{head}}N^2d_{\text{head}})$	$O(BN^2D)$

S2. Average concentrations and corresponding AQI categories

Table S2. The average $PM_{2.5}$ and PM_{10} concentrations over 24-hours, and the corresponding AQI categories[1].

Level	**$PM_{2.5}$ ($\mu g/m^3$)**	**PM_{10} ($\mu g/m^3$)**	**AQI**	**Category**	**Health impact**
1	0.0 – 9.0	0 – 54	0 – 50	Good	
2	9.1 – 35.4	55 – 154	51 – 100	Moderate	
3	35.5 – 55.4	155 – 254	101 – 150	Unhealthy for Sensitive Groups	
4	55.5 – 125.4	255 – 354	151 – 200	Unhealthy	
5	125.5 – 225.4	355 – 424	201 – 300	Very Unhealthy	
6	225.5+	425+	301+	Hazardous	

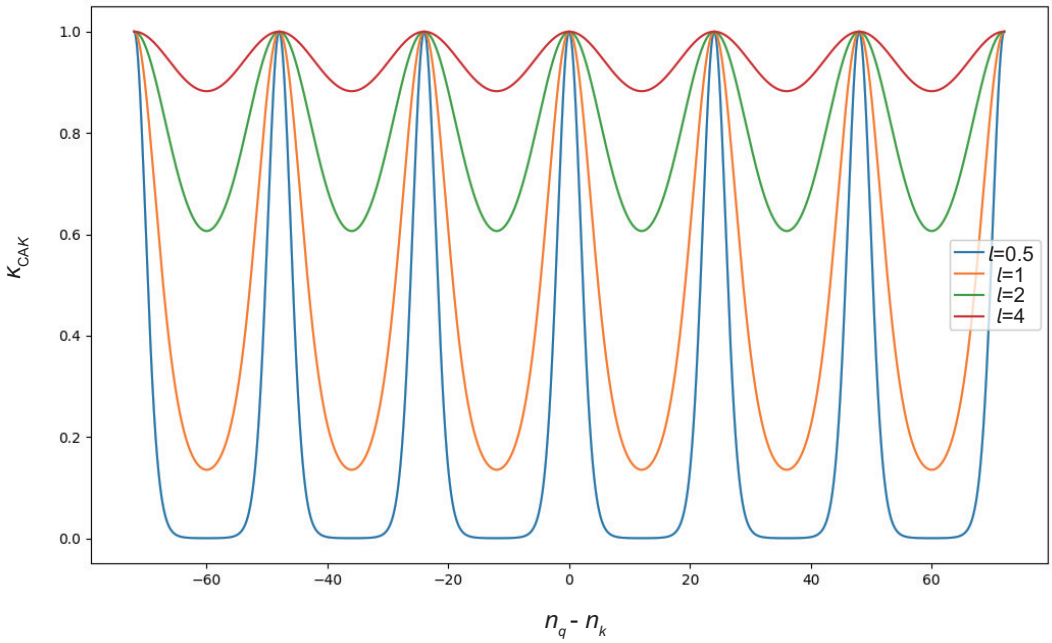

Figure S1. The cross-sensor alignment kernel function, $\kappa_{\mathrm{CAK}}(\mathbf{q}, \mathbf{k})$, assuming $l = 24$ and $p \in \{0.5, 1, 2, 4\}$.

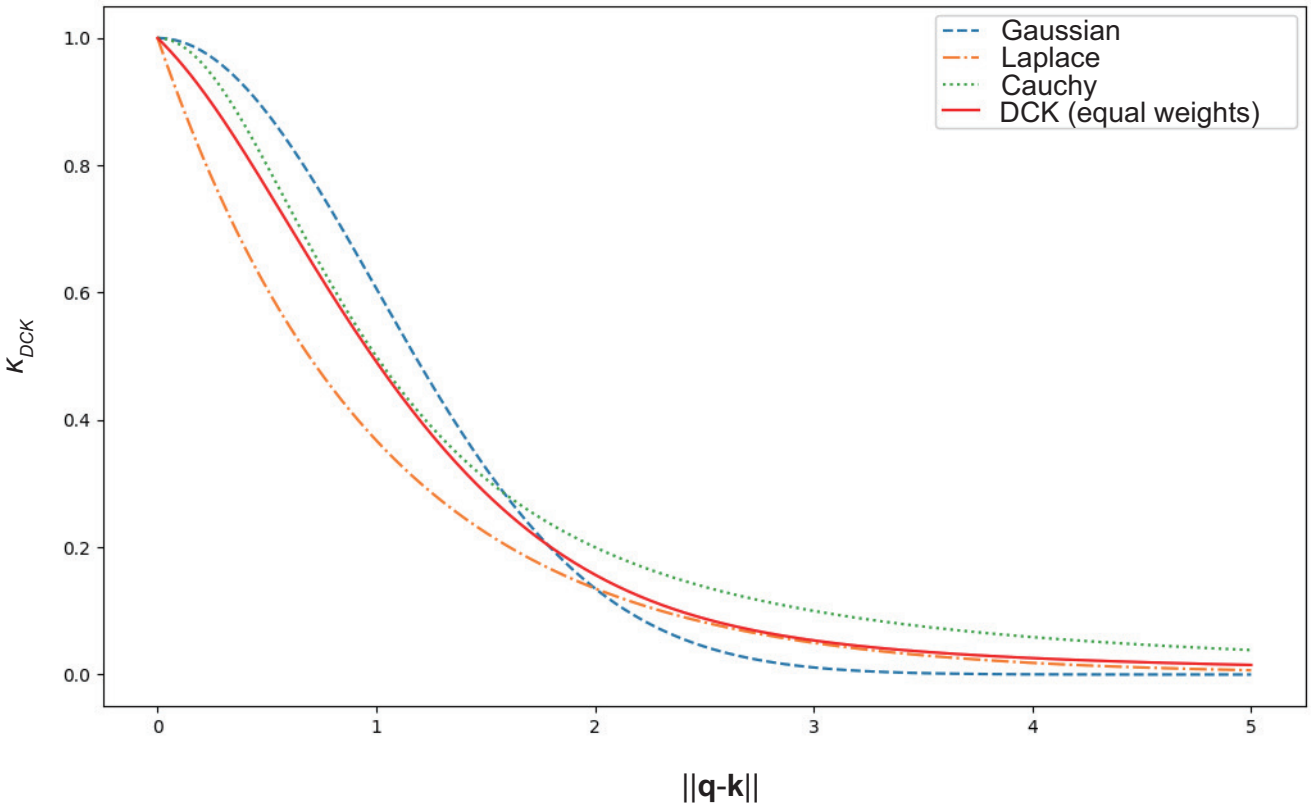

Figure S2. The Gaussian, Laplace, and Cauchy functions, and the diversity-correlation kernel function, $\kappa_{\mathrm{DCK}}(\mathbf{q}, \mathbf{k})$.

S3. Additional Numerical Results

Table S3. The $PM_{2.5}$ and PM_{10} data statistics considered in numerical experiments.

Pollutant	Mean	Standard Deviation	Min	Max	Median
$PM_{2.5}$	4.9510	7.3900	0	333.1170	2.4765
PM_{10}	5.7660	8.5740	0	455.2520	2.9760

Table S4. The baseline models considered for performance and complexity comparisons.

Model	Source	Architecture
iTransformer	[2]	Transformer encoder
CrossFormer	[3]	Transformer with temporal attention and cross-channel attention
DLinear	[4]	Moving average decomposition + MLP
FreTS	[5]	FFT + MLP
TimeMixer	[6]	Seasonal-trend decomposition in multi-scale + MLP
ETSformer	[7]	FFT + Transformer encoder
BigST	[8]	Linearized spatial attention with graph structure
PatchSTG	[9]	Hierarchical spatial patch Transformer

Table S5 and Table S6 present the ablation results for predicting the $PM_{2.5}$ and PM_{10} concentrations under different configurations of the main attention $\mathbf{A}_{\text{main}}$, the IDW attention $\mathbf{A}_{\text{IDW}}$, and the residual attention $\mathbf{A}_{\text{res}}$, which are shown in the block diagram of the proposed IDW-ADAFormer in Figure 1 in the main text.

For both pollutants, and for all forecasting experiment configurations, the complete attention model with all three attention modules included achieves the best or almost the best MAE, RMSE and sMAPE values, and competitive AQI-Acc scores. It implies that all three components are necessary for a good performance. As shown in Table S5, the full attention model achieves the most balanced performance for the $PM_{2.5}$ dataset, and yields small RMSE, relatively large AQI-Acc, and reduced sMAPE for all lengths of the prediction horizon. For $(L_p, L_h) = (96, 48)$ and $(L_p, L_h) = (120, 72)$, the configuration without any attention modules exhibits slightly smaller MAE values than the full model. However, the configuration without attention underperforms in other metrics; for example, RMSE, sMAPE, and AQI-Acc values deteriorate, for $(L_p, L_h) = (96, 48)$, and for $(L_p, L_h) = (120, 72)$. The configurations, which exclude some, but not all modules generally yield an intermediate performance.

In summary, the ablation experiments suggest that removing attention may slightly reduce MAE at the expense of diminished ability to capture the extreme and high-variability samples in the time series, resulting in larger RMSE and sMAPE and smaller AQI-Acc.

Table S5. Ablation study for predicting $PM_{2.5}$ values under different configurations of the attention module.

Modules			$(L_p, L_h) = (72, 24)$				$(L_p, L_h) = (96, 48)$				$(L_p, L_h) = (120, 72)$			
$\mathbf{A}_{\text{main}}$	$\mathbf{A}_{\text{IDW}}$	$\mathbf{A}_{\text{res}}$	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE
✓	✓	✓	**1.8610**	**5.8997**	0.9737	**77.8353**	1.9863	**6.0054**	**0.9751**	**80.9647**	2.0333	**6.0554**	**0.9763**	**82.7519**
	✓	✓	1.8906	5.9364	0.9714	78.9329	2.0417	6.0724	0.9740	82.0893	2.0362	6.0603	0.9760	82.9795
✓		✓	1.8636	5.9025	0.9731	77.8563	1.9943	6.0072	0.9744	81.0487	2.0356	6.0562	0.9762	82.7941
✓	✓		1.8626	5.9046	**0.9738**	77.8952	1.9940	6.0089	0.9745	81.0368	2.0368	6.0591	**0.9763**	82.8310
✓			1.8649	5.9075	0.9733	77.9012	1.9955	6.0100	0.9744	81.0642	2.0356	6.0577	0.9762	82.7877
	✓		1.8736	5.9157	0.9736	78.6098	1.9821	6.0220	0.9758	81.3762	2.0408	6.0672	0.9762	82.9785
		✓	1.9315	5.9954	0.9707	79.0240	2.0290	6.0615	0.9749	81.9343	2.0532	6.0870	0.9750	83.1060
			1.8898	5.9479	0.9723	78.5888	**1.9770**	6.0174	0.9745	81.0812	**2.0241**	6.0724	0.9752	82.7697

As reported in Table S6, the full model that contains all three attention modules offers the best performance for predicting the PM_{10} data. For all prediction horizons, it achieves the smallest MAE and RMSE, the largest AQI-Acc, and the competitive sMAPE values. For short prediction horizons, the configuration without any attention modules has similar MAE performance as the full attention model. However, completely removing the attention consistently underperforms in other performance metrics including RMSE, sMAPE, and AQI-Acc. Removing some, but not all attention modules again provides an intermediate performance. The results suggest that removing the attention may occasionally yield a comparable or slightly lower MAE

values, however, it negatively affects the model ability to capture the extreme and high-variability samples in time series data, which leads to larger RMSE, sMAPE, and reduced AQI-Acc values.

Table S6. Ablation study for predicting PM_{10} values under different configurations of the attention module.

Modules			$(L_p, L_h) = (72, 24)$				$(L_p, L_h) = (96, 48)$				$(L_p, L_h) = (120, 72)$			
$\mathbf{A}_{\text{main}}$	$\mathbf{A}_{\text{IDW}}$	$\mathbf{A}_{\text{res}}$	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE
✓	✓	✓	**2.1952**	**7.0003**	**0.9997**	69.6904	**2.3244**	**7.1119**	**0.9998**	**72.8543**	**2.3657**	**7.1879**	**0.9997**	**74.2665**
	✓	✓	2.1931	7.0133	0.9994	69.9618	2.3293	7.1198	0.9997	73.1932	2.3870	7.2056	0.9992	74.6408
✓		✓	2.1983	7.0052	0.9996	69.7407	2.3257	7.1130	**0.9998**	72.8722	2.3703	7.1905	**0.9997**	74.3177
✓	✓		2.1961	7.0048	0.9996	**69.6724**	2.3263	7.1150	**0.9998**	72.8722	2.3672	7.1898	**0.9997**	74.2865
✓			2.1992	7.0098	0.9995	69.7285	2.3277	7.1163	**0.9998**	72.8902	2.3719	7.1926	**0.9997**	74.3385
	✓		2.2004	7.0144	**0.9997**	70.1851	2.3230	7.1265	**0.9998**	72.9267	2.3814	7.1967	0.9995	74.6722
		✓	2.2106	7.0166	0.9995	69.9177	2.3929	7.1968	0.9996	73.9745	2.4884	7.3324	0.9989	75.8611
			2.2025	7.0166	**0.9997**	69.7548	2.3375	7.1334	**0.9998**	73.1500	2.3874	7.2062	**0.9997**	74.6848

Figure S3 illustrates the prediction stability of different forecasting methods for the $PM_{2.5}$ and PM_{10} time series under the MSE training and the proposed AQI-aware training. The error bars represent the half-widths of the corresponding 95% confidence intervals (CIs). Overall, the proposed method consistently achieves relatively small confidence interval half-widths across MAE, RMSE, and sMAPE under both loss functions indicating the stable forecasting performance across different evaluation data samples. In particular, despite obtaining the best or near-the best average prediction accuracy, the proposed method does not exhibit substantially larger uncertainty ranges in comparison with the competing models. For the $PM_{2.5}$ dataset, most methods exhibit comparable 95% CI half-widths for both MAE and RMSE, indicating similar levels of statistical variability across repeated evaluations. The Crossformer shows slightly larger uncertainty ranges for MAE compared with the other methods. More importantly, the proposed loss function consistently reduces the forecasting errors across all three performance metrics while producing similar CI half-widths to those obtained under the MSE loss, suggesting that the performance improvements are not achieved at the expense of increased predictive uncertainty. A similar trend can be also observed for the PM_{10} dataset.

Figure S4 presents the row-normalized misclassification transition distributions across discretized pollution levels under different prediction configurations. For $PM_{2.5}$, assuming the MSE loss, the prediction errors are mainly concentrated at level 2 and level 3, accounting for 0.42% and 0.54% of all misclassified samples, respectively. Most classification errors are level 2 misclassified as level 3, while those from level 3 are primarily assigned to level 2, indicating the strong bidirectional confusion between these adjacent levels. The errors at level 4 account for only 0.04% of cases. Under the proposed loss function, a similar, but slightly more dispersed pattern is observed, where the correspondent errors are distributed across levels 2, 3, and 4 with comparable proportions of 0.41%, 0.54%, and 0.05%, respectively. For PM_{10}, both loss functions exhibit a strong concentration of classification errors at level 2, with most of these misclassifications being assigned to level 1. Overall, the resulting misclassification matrices exhibit highly sparse transition patterns, where only a limited number of entries contain the dominant probabilities under both loss functions. This suggests that the forecasting errors are structurally distributed rather than randomly scattered across the AQI levels.

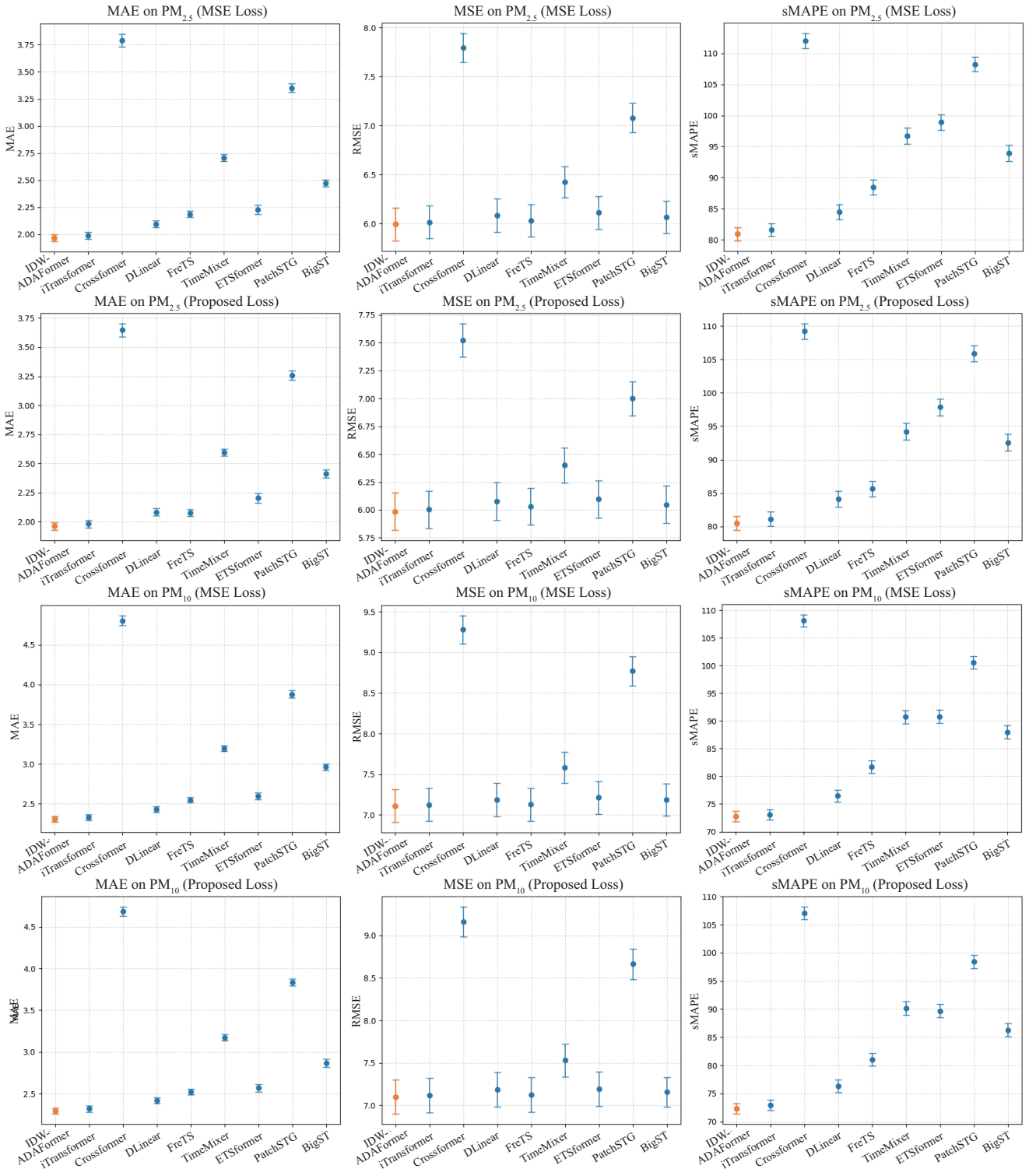

Figure S3. The error bars denoting the 95% confidence interval half-widths.

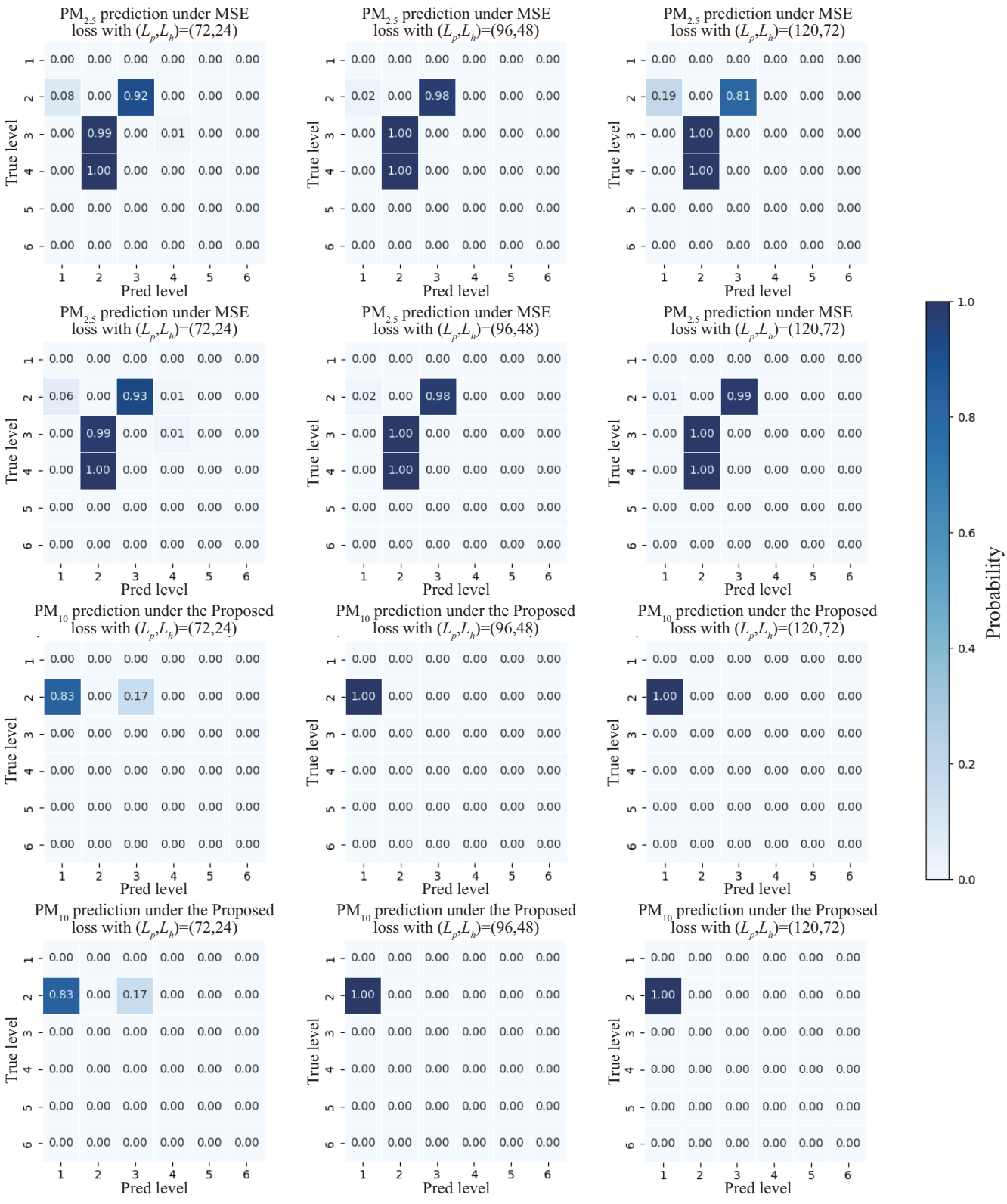

Figure S4. The row-normalized confusion matrix of AQI-level misclassifications.

Figure S5 illustrates the distribution of the optimal neighborhood sizes k for both $PM_{2.5}$ and PM_{10} datasets. These results reveal a clear bimodal pattern, where small and large values of k dominate, while the intermediate values occur less frequently. This indicates that the optimal neighborhood selection tends to favor either highly local or global spatial dependencies, rather than the interactions at intermediate scales. Moreover, the distributions of k for $PM_{2.5}$ and PM_{10} have similar shapes, so the adaptive neighborhood mechanism is consistent across different datasets. Such consistency implies that the spatial dependency structure is primarily governed by the underlying sensor network geometry and by the environmental diffusion characteristics, rather than the specific pollutant properties.

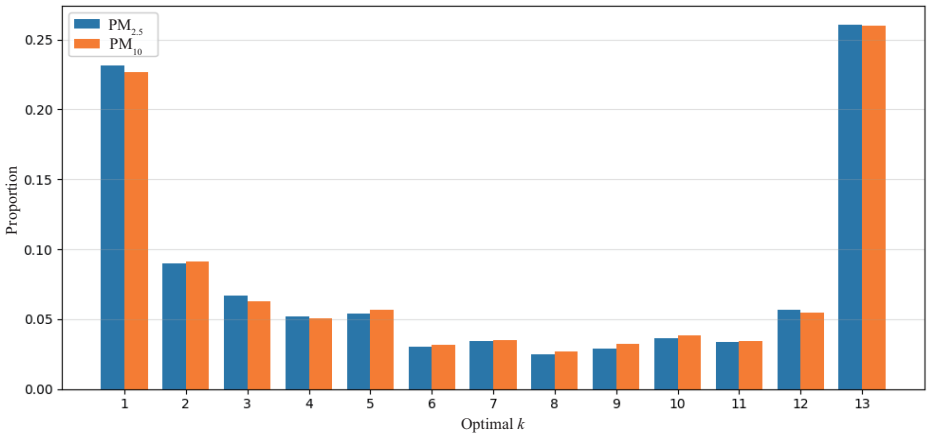

Figure S5. The probability distributions of the optimal k values for the $PM_{2.5}$ and PM_{10} datasets.

Figure S6 presents the box plot of the optimal neighborhood sizes k across different sensors. The results show a substantial variability among different sensors, pointing to the spatial heterogeneity in the adaptive neighborhood selection process. This suggests that different sensor locations require distinct neighborhood scales to achieve the optimal interpolation performance, reflecting a non-uniform spatial dependence within the monitoring network. Since the k values are discrete, the median denoted by the yellow line may coincide with the upper or the lower quartile, especially when the distribution is highly concentrated or skewed. Such a behavior does not indicate an anomaly but rather reflects the discrete and often concentrated nature of the optimal selection of k. Furthermore, the results for both $PM_{2.5}$ and PM_{10} exhibit similar patterns, which is consistent with the observations in Figure S5.

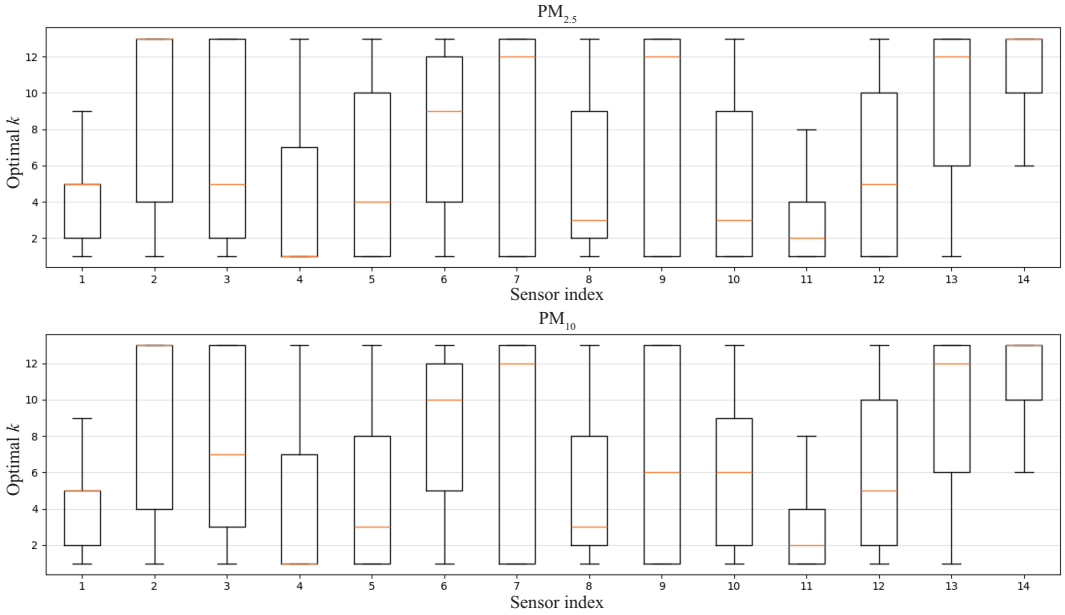

Figure S6. The spatial distribution of the optimal neighborhood sizes k.

Figure S7 illustrates the averaged FLOPs and the number of parameters of different models under their optimal configurations. These results show a substantial variation in the computational complexity across different models, which is primarily driven by differences in their architectural designs. Among all methods, DLinear demonstrates a clear advantage in the computational efficiency due to its extremely simple linear structure containing only a small number of learnable parameters. Consequently, DLinear achieves the smallest FLOPs and have the smallest parameter count among all other models considered. In contrast, PatchSTG has the largest computational cost, which can be attributed to its patch-based spatio-temporal embedding and to using multi-stage feature interaction modules, which significantly increases the model size.

The proposed IDW-ADAFormer maintains a moderate level of the computational complexity. Its FLOPs and the parameter count are primarily dictated by the ADA module. The dominant computational cost arises from the kernel-based multi-head attention operations within each ADA layer. These operations perform pairwise interactions across sensors, leading to a quadratic dependence on the number of sensors, i.e., $O(B \cdot L_p \cdot N^2 \cdot D)$, where N is the number of sensors (variates), L_p is the lookback length, B is the batch size, and D is the embedding dimension. In the standard Transformer-based architectures, the self-attention is typically employed along the temporal dimension for each time series independently, and $N \ll L_p$, resulting in a computational complexity of $O(B \cdot N \cdot L_p^2 \cdot D)$. In contrast, the proposed model replaces the temporal self-attention with cross-sensor attention, thereby reducing the asymptotic dependence on the sequence length. However, this theoretical reduction does not directly translate into a proportional decrease in FLOPs in practice. This is because the proposed architecture introduces multiple kernel-based attention branches (periodic and heterogeneous kernels), the cross-branch fusion mechanism, and feed-forward networks in each layer. In particular, each ADA layer consists of multiple parallel attention computations followed by nonlinear kernel evaluations and the feature fusion operations. Moreover, the use of multi-head projections across three interaction data flows (main, IDW, and residual) further increases the computational density per layer, leading to a moderate overall FLOPs. The overall parameter count of IDW-ADAFormer is primarily contributed by (i) the feed-forward networks within each ADA layer, which are comparable to the standard Transformer-based architectures, and (ii) the linear embedding layers shared across the input sources. Due to the multi-branch design (main, IDW, and residual data flows), the embedding layers introduce additional parameter overheads compared to single-flow architectures.

In concussion, although the proposed IDW-ADAFormer introduces additional computations due to its multi-branch processing, it also achieves a good efficiency-performance trade-off with only moderate increases in overall computational cost.

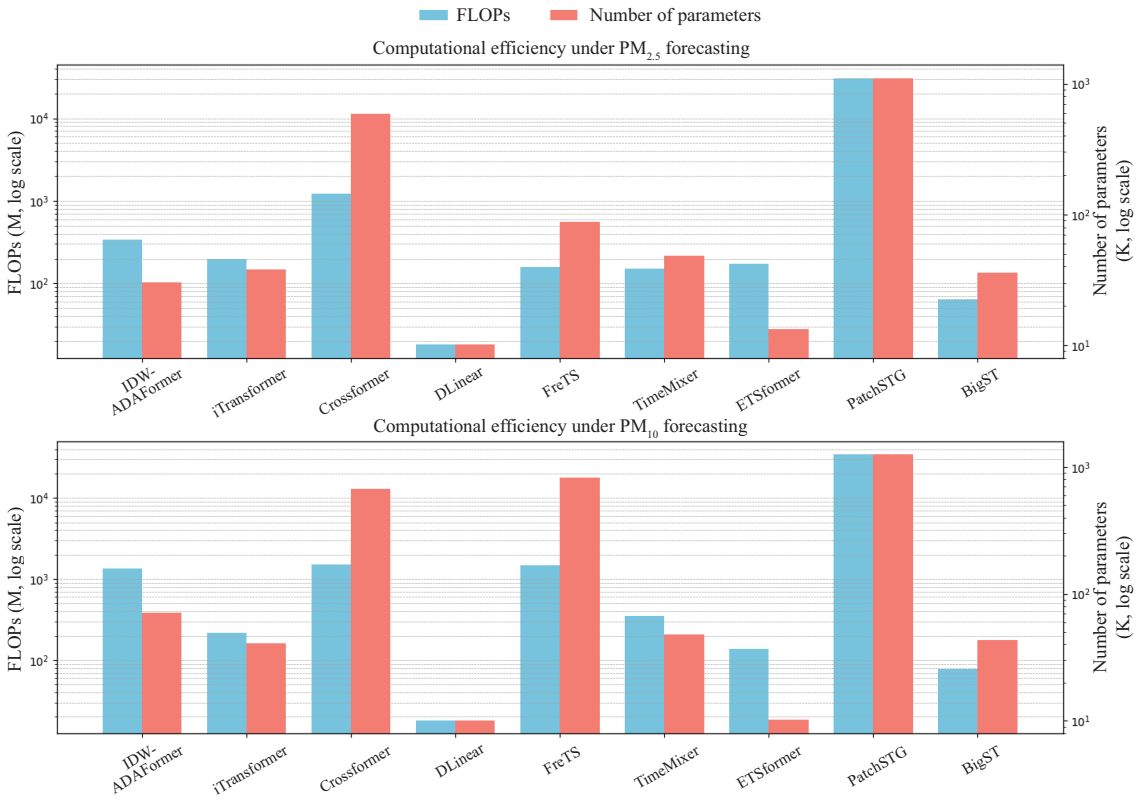

Figure S7. The comparison of the computational complexity in terms of FLOPs and the parameter counts.

References

1. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Air Quality Assessment Division. Technical assistance document for the reporting of daily air quality – the air quality index (AQI). Tech. Rep., U.S. Environmental Protection Agency, Research Triangle Park, NC (2024).

2. Liu, Y. *et al.* iTransformer: Inverted transformers are effective for time series forecasting. In *Proceedings of ICLR* (2024).

3. Zhang, Y. & Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *Proceedings of ICLR* (2023).

4. Zeng, A., Chen, M., Zhang, L. & Xu, Q. Are transformers effective for time series forecasting? In *Proceedings of AAAI* (2023). DOI 10.1609/aaai.v37i9.26317.

5. Yi, K. *et al.* Frequency-domain MLPs are more effective learners in time series forecasting. In *Proceedings of NIPS* (Red Hook, NY, USA, 2023).

6. Wang, S. *et al.* TimeMixer: Decomposable multiscale mixing for time series forecasting. In *Proceedings of ICLR* (2024).

7. Woo, G., Liu, C., Sahoo, D., Kumar, A. & Hoi, S. ETSformer: Exponential smoothing transformers for time-series forecasting (2022). ArXiv:2202.01381 [cs.LG].

8. Han, J. *et al.* BigST: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proc. VLDB Endow.* **17**, 1081–1090 (2024). DOI 10.14778/3641204.3641217.

9. Fang, Y. *et al.* Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of KDD*, 307–317 (2025). DOI 10.1145/3690624.3709177.