ExtrapAir: Air Quality Inference at Unmonitored Locations via Weather-Bridged Spatial Attention

Gaoyun Lin[1], Binwu Wang[1,2]*, Zhengyang Zhou[1,2], Yang Wang[1,2]*
[1]University of Science and Technology of China (USTC), Hefei, Anhui, China
[2]Suzhou Institute for Advanced Research, USTC, Suzhou, Jiangsu, China
gaoyunlin@mail.ustc.edu.cn
{wbw2024, zzy0929, angyan}@ustc.edu.cn

Abstract

Air quality prediction at unmonitored locations is challenging because historical AQI observations are unavailable, while conventional spatiotemporal models largely rely on target histories to learn spatial dependencies. Weather covariates, in contrast, are widely available and provide station-specific evidence for target-absent inference. In this paper, we propose **ExtrapAir**, a lightweight weather-bridged framework for air quality inference at unmonitored locations. ExtrapAir preserves variable-specific weather patterns through per-variable weather encoding and learns adaptive weather and AQI correlations for spatial transfer. We further introduce **Activation Attention with Prior Inductive Biases**, which enables non-competitive spatial aggregation, incorporates geographic and semantic priors for station-pair guidance, and avoids full station-to-station attention computation for scalability. Experiments on real-world air-quality datasets against **32** baselines show that ExtrapAir achieves better accuracy at both overall and unmonitored stations, with lower training time and memory cost. Our code is available at this repository.

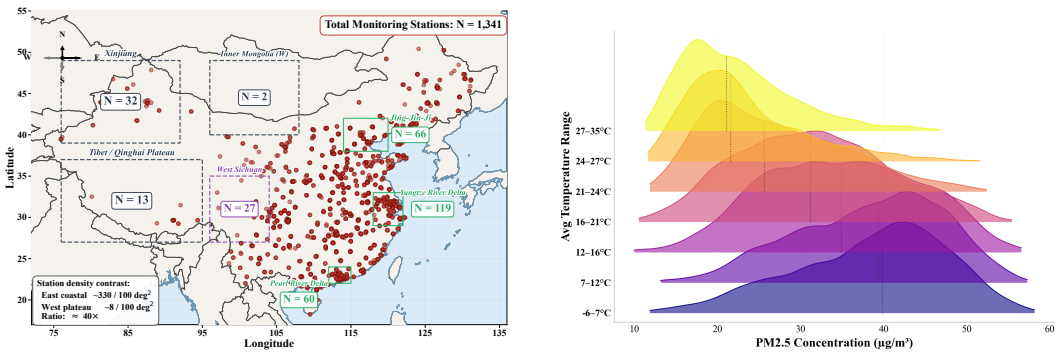
1 Introduction

Air quality prediction is vital for smart cities, environmental governance, and public health protection [1, 2]. Existing graph- and Transformer-based spatiotemporal models have advanced monitored-station forecasting, but typically assume that historical AQI observations are available at every target location [3, 4]. In practice, monitoring stations are costly and unevenly distributed, leaving many suburban, rural, and underdeveloped regions weakly covered or unmonitored, as shown in Figure 1(a). We therefore study *air quality inference at unmonitored locations*, where future AQI must be predicted without station-specific historical AQI records. This absence of target histories makes AQI-based spatial correlations unreliable, limiting both GCN-based signal propagation and Transformer-based content attention for spatial transfer.

Although AQI observations are spatially sparse, weather covariates are widely available from numerical weather prediction and reanalysis products. These factors influence pollutant formation, accumulation, dispersion, and removal, making them a natural bridge for inferring AQI at unmonitored locations. As shown in Figure 1(b) and Figure 16, our lag-wise Granger causality analysis reveals variable-specific and lag-dependent effects: temperature shows the strongest short-term influence, followed by humidity and the north–south wind component, while wind speed contributes through sustained dispersion. However, existing methods often ignore such heterogeneity by discarding external factors or simply concatenating them with AQI features. Therefore, reliable unmonitored-location inference requires explicit meteorology-aware modeling that captures variable-specific AQI–meteorology dependencies and transfers them from monitored to unmonitored areas.

*Binwu Wang and Yang Wang are corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Station coverage.

(b) Lag-wise weather–AQI dependence.

Figure 1: The sparse distribution of air quality monitoring stations in mainland China and the influence of weather covariates on AQI.

Another challenge lies in scalable spatial aggregation. In unmonitored-location inference, an invisible station often needs to collect weak but complementary evidence from multiple visible stations with similar weather, geographic, or functional contexts. However, standard softmax attention imposes row-wise normalization, forcing stations to compete for a fixed attention budget and often concentrating weights on a few dominant neighbors. This competitive behavior is undesirable for distributed evidence aggregation. Moreover, its quadratic complexity in the number of stations limits scalability on large monitoring networks. Therefore, air quality inference at scale requires a spatial aggregation mechanism that is both non-competitive and computationally efficient.

To address these challenges, we propose **ExtrapAir**, a lightweight and effective framework for air quality inference at unmonitored locations. ExtrapAir treats weather as a central information bridge rather than a peripheral covariate. It separately encodes AQI and weather variables, preserves variable-specific weather patterns through per-variable weather encoders, and learns adaptive weather and AQI correlations through gated fusion blocks. To transfer the learned weather and AQI dependencies across stations, ExtrapAir introduces **Activation Attention with Prior Inductive Biases**, a spatial aggregation module designed for target-absent inference. It avoids softmax-induced competition so that unmonitored locations can aggregate complementary evidence from multiple monitored stations, and injects geographic and semantic priors into query and key representations to provide reliable station-pair guidance when historical AQI similarity is unavailable. Its linear computation further improves scalability on large monitoring networks.

Our contributions are four-fold: ❶ **Practical Problem Focus.** We revisit air quality prediction from the perspective of target-absent spatial inference, emphasizing robust generalization to unmonitored locations. ❷ **Weather-Bridged Framework.** We propose **ExtrapAir**, which treats weather covariates as station-specific information gain, preserves variable-specific weather patterns, and learns adaptive weather and AQI correlations for transfer to unmonitored locations. ❸ **Prior-Guided Efficient Aggregation.** We design Activation Attention with Prior Inductive Biases to enable non-competitive spatial aggregation, inject geographic and semantic priors for reliable station-pair guidance, and avoid full station-to-station attention computation for scalability. ❹ **Extensive Evaluation.** Experiments against 32 baselines on seven datasets demonstrate that ExtrapAir not only improves MAE by up to **15.24%** overall and **16.01%** on unmonitored stations, but also reduces training time by **84.9%** and memory usage by **76.23%**, achieving a strong accuracy–efficiency trade-off.

2 Preliminary

Notation 1. AQI and Weather Variables: Let $\mathbf{x}_t \in \mathbb{R}^c$ denote the air-quality index (AQI) collected at one spatial location at time step t. Let $\mathbf{z}_t \in \mathbb{R}^f$ denote the weather variables, such as temperature, humidity, and wind speed. $\mathbf{z}_t^p$ and $\mathbf{z}_t^f$ denote past observed and future predicted variables, respectively.

Notation 2. Sparse Spatial Locations: Given N locations of interest, we divide the stations into two disjoint subsets: **monitored stations** $\mathcal{S}_{\text{vis}}$ and **unmonitored stations** $\mathcal{S}_{\text{inv}}$. For monitored stations in $\mathcal{S}_{\text{vis}}$, both historical AQI records and weather covariates are available. For unmonitored stations

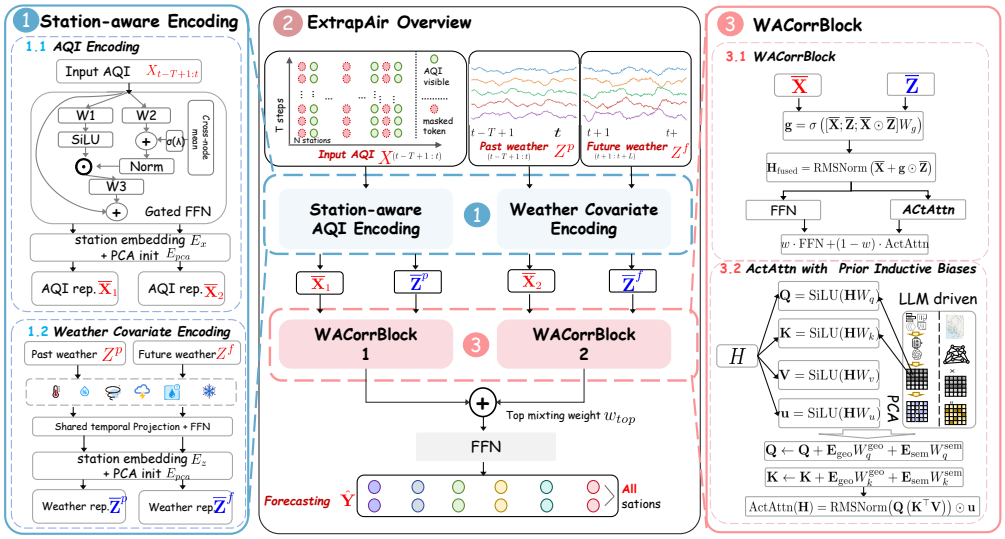


Figure 2: Overview of the proposed ExtrapAir.

in $\mathcal{S}_{\text{inv}}$, historical AQI records are inaccessible and are replaced with a learnable token; only the corresponding weather features are available.

Task Formulation: AQI Inference. Given the (partially masked) past AQI $\mathbf{X}_{t-T+1:t} \in \mathbb{R}^{T\times N\times c}$, past covariates $\mathbf{Z}^p_{t-T+1:t} \in \mathbb{R}^{T\times N\times f}$, and future covariates $\mathbf{Z}^f_{t+1:t+L} \in \mathbb{R}^{L\times N\times f}$, the goal is to learn a function $\mathcal{F}$ that predicts future AQI for *all* N stations: $\mathbf{Y}_{t+1:t+L} \in \mathbb{R}^{L\times N\times c}$.

3 Methodology

In this section, we present the details of ExtrapAir. The model first encodes AQI and weather covariates through independent branches, then models their interactions via WACorrBlocks equipped with ActAttn and dual bias injection. An overview is shown in Figure 2.

3.1 Station-Aware AQI and Weather Encoding

Weather provides crucial exogenous signals for AQI prediction, but weather variables such as temperature, humidity, wind speed, and precipitation differ substantially in value ranges and temporal dynamics. Directly mixing them at the input level may obscure variable-specific patterns and introduce cross-variable interference. We therefore introduce a Per-Variable Encoder that first models each weather variable independently and then fuses them into a unified weather representation.

Weather Covariate Encoding. For each weather variable v, we extract its temporal sequence $\mathbf{z}^{(v)}$, encode it using a shared temporal projection followed by an FFN, and concatenate the resulting variable-specific representations for fusion:

$$\overline{\mathbf{Z}} = \left[\text{FFN}(\mathbf{z}^{(1)}W_{\text{proj}}); \cdots ; \text{FFN}(\mathbf{z}^{(V)}W_{\text{proj}})\right] W_{\text{fuse}} + \mathbf{E}_z. \tag{1}$$

where the parameter W_{proj} is shared across variables. The same encoder is applied to past observations and future forecasts of the weather variables, and the outputs are denoted as $\overline{\mathbf{Z}}^p$ and $\overline{\mathbf{Z}}^f$. $\mathbf{E}_z$ represents the learnable station embedding for the weather stream.

Input AQI Encoding. We further map the input AQI $\mathbf{X}_{t-T+1:t}$ to a high-dimensional representation:

$$\widetilde{\mathbf{X}} = \text{FFN}(\mathbf{X}_{t-T+1:t}) + \mathbf{E}_x, \tag{2}$$

where $\mathbf{E}_x$ is the learnable positional embedding of the air monitoring station. To obtain a good initialization, especially for unmonitored stations, $\mathbf{E}_x$ and $\mathbf{E}_z$ are combined with a shared PCA-initialized embedding $\mathbf{E}_{\text{pca}}$ derived from high-dimensional LLM station embeddings (details in **Section 3.4**), providing each station with a semantically meaningful starting representation. Subsequently, $\widetilde{\mathbf{X}}$ is fed

into a Gated-FFN to further capture complex spatiotemporal dynamics:

$$\text{Gated-FFN}(\widetilde{\mathbf{X}}) = \left(\text{SiLU}(\widetilde{\mathbf{X}}W_1) \odot \text{RMSNorm}\left(\widetilde{\mathbf{X}}W_2 + \sigma(\lambda)\hat{\mathbf{X}}W_2\right)\right) W_3, \tag{3}$$

where W_1, W_2, and W_3 are learnable parameters. $\hat{\mathbf{X}} = \frac{1}{N}\sum_i \widetilde{\mathbf{x}}_i$ is the station-wise mean representation, $\text{RMSNorm}(\cdot)$ denotes root mean square layer normalization, and λ is a learnable parameter. The output is denoted as $\overline{\mathbf{X}} \in \mathbb{R}^{N\times d_h}$.

3.2 Weather Exogenous Signals Learning for AQI Inference

Given the AQI representation $\widetilde{\mathbf{X}}$ and the weather representation $\overline{\mathbf{Z}} \in \mathbb{R}^{N\times d_h}$, we next model their dependency through an adaptive weather–AQI interaction module. This is particularly crucial for invisible stations, where historical AQI is unavailable and weather covariates become the primary local signal. Because weather effects on AQI are station-specific and regime-dependent, simple concatenation cannot reliably capture their conditional relationship. We therefore design a **Weather–AQI Correlation Block (WACorrBlock)**, whose key is **Activation Attention with Prior Inductive Biases**. This allows weather–AQI correlations learned from visible stations to be transferred to invisible stations under the guidance of geographic and semantic priors.

Given an encoded AQI representation $\overline{\mathbf{X}} \in \mathbb{R}^{N\times d_h}$ and an encoded weather representation $\overline{\mathbf{Z}} \in \mathbb{R}^{N\times d_h}$, WACorrBlock first estimates a dimension-wise weather–AQI correlation gate and uses it to obtain a weather-aware AQI representation:

$$\begin{aligned} \mathbf{g} &= \sigma\left([\overline{\mathbf{X}}; \overline{\mathbf{Z}}; \overline{\mathbf{X}} \odot \overline{\mathbf{Z}}]W_g\right), \\ \mathbf{H}_{\text{fused}} &= \text{RMSNorm}\left(\overline{\mathbf{X}} + \mathbf{g} \odot \overline{\mathbf{Z}}\right), \end{aligned} \tag{4}$$

where $[\cdot;\cdot]$ denotes feature concatenation, $\odot$ is element-wise multiplication, and $\mathbf{g} \in \mathbb{R}^{N\times d_h}$ measures the relevance of weather signals to each AQI latent dimension. This residual fusion preserves AQI-derived information for visible stations, while allowing invisible stations to rely more on weather-driven evidence when historical AQI is unavailable. The fused representation is then processed by parallel local FFN and ActAttn modules for correlation modeling (detailed in **Section 3.4**):

$$\text{WACorrBlock}(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) = w \cdot \text{FFN}(\mathbf{H}_{\text{fused}}) + (1-w) \cdot \text{ActAttn}(\mathbf{H}_{\text{fused}}), \tag{5}$$

where $w = \sigma(\alpha)$ is a learnable scalar.

3.3 Weather Covariate-Driven AQI Inference

ExtrapAir deploys two WACorrBlocks with independent AQI branches to reflect the distinct roles of past and future weather covariates in AQI inference. The past-weather branch captures diagnostic weather–AQI associations from historical observations, while the future-weather branch incorporates future weather forecast covariates as forward-looking cues for the prediction horizon. This dual-branch design is particularly useful for invisible stations, where weather covariates serve as key evidence in the absence of historical AQI. Formally, the prediction is obtained as:

$$\hat{\mathbf{Y}} = \text{FFN}_{\text{dec}}\left(w_{\text{top}} \cdot f_0(\overline{\mathbf{X}}_1, \overline{\mathbf{Z}}^p) + (1-w_{\text{top}}) \cdot f_1(\overline{\mathbf{X}}_2, \overline{\mathbf{Z}}^f)\right), \tag{6}$$

where f_0 and f_1 denote two WACorrBlocks, $\overline{\mathbf{X}}_1$ and $\overline{\mathbf{X}}_2$ are independently encoded AQI representations, $\overline{\mathbf{Z}}^p$ and $\overline{\mathbf{Z}}^f$ denote past observed weather and future weather forecast encodings, and $w_{\text{top}} = \sigma(\alpha_{\text{top}})$ is a learnable top-level mixing weight.

3.4 Activation Attention with Prior Inductive Biases

Unmonitored-location air-quality inference requires invisible stations, which lack historical AQI observations, to aggregate complementary evidence from multiple monitored stations in a scalable and non-competitive manner. We address this challenge by developing ActAttn, an activation-based linear attention mechanism, and augmenting its query-key representations with geo-semantic inductive biases that provide explicit relational guidance for invisible stations.

3.4.1 ActAttn: Non-Competitive Linear Attention for Air Quality Inference

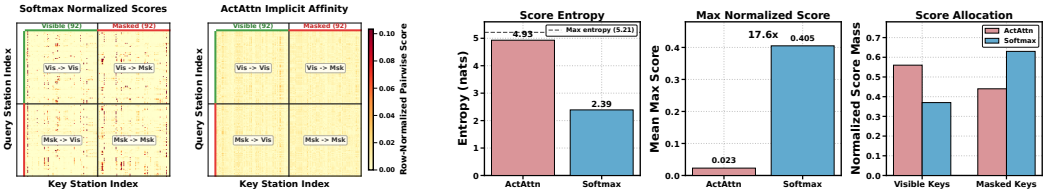
Spatial attention is effective for modeling interactions among air-quality monitoring stations [4, 5, 3], but standard mechanisms are not well suited to unmonitored-location inference. They are computationally expensive for large networks and rely on content similarity from historical AQI trajectories, which is unavailable at invisible stations. Therefore, such stations must infer future AQI by aggregating distributed evidence from multiple monitored stations through weather, geographic, and urban-functional cues.

Specifically, to improve computational efficiency, we adopt the linear attention formulation [6], which avoids constructing the full $N \times N$ attention matrix. Given a feature map $\phi(\cdot)$, linear attention rewrites spatial aggregation as

$$\mathrm{Attn}_{\mathrm{lin}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \phi(\mathbf{Q}) \left(\phi(\mathbf{K})^\top \mathbf{V}\right), \tag{7}$$

where $\phi(\mathbf{K})^\top \mathbf{V} \in \mathbb{R}^{d_h \times d_h}$ is computed first and then multiplied by $\phi(\mathbf{Q})$. This changes the complexity from $\mathcal{O}(N^2 d_h)$ to $\mathcal{O}(N d_h^2)$, making spatial aggregation linear in the number of stations.

However, efficiency alone is insufficient. Softmax attention normalizes station weights onto a probability simplex, inducing zero-sum competition among stations: assigning more weight to one station necessarily suppresses the others. As shown in Figure 3, this often leads to over-concentration on a few dominant stations. Such behavior is undesirable for unmonitored-location inference, where invisible stations lack reliable AQI-history-based queries and therefore require broader evidence aggregation from weather, geographic, and functional cues.



(a) Score maps (b) Masked-query score analysis

Figure 3: Attention characteristics analysis on KnowAir.

Therefore, we do not use linear attention as a mere approximation to Softmax attention. Instead, we revisit its feature-map design from the perspective of air quality inference. As shown in Table 8, among six candidate feature maps, SiLU achieves the best performance. We attribute this to its sign-preserving and unbounded nature, which enables more expressive station-wise interactions without imposing a fixed attention budget. This property is particularly suitable for unmonitored-location inference, where an unseen station needs to accumulate complementary evidence from multiple monitored stations. Accordingly, we instantiate **Activation Attention (ActAttn)** by directly applying SiLU to the query, key, and value projections:

$$\begin{aligned} \mathbf{Q} &= \mathrm{SiLU}(\mathbf{H}W_q), \quad \mathbf{K} = \mathrm{SiLU}(\mathbf{H}W_k), \\ \mathbf{V} &= \mathrm{SiLU}(\mathbf{H}W_v), \quad \mathbf{u} = \mathrm{SiLU}(\mathbf{H}W_u), \\ \mathrm{ActAttn}(\mathbf{H}) &= \mathrm{RMSNorm}\big(\mathbf{Q}\left(\mathbf{K}^\top \mathbf{V}\right)\big) \odot \mathbf{u}. \end{aligned} \tag{8}$$

where the parameters $W_q, W_k, W_v, W_u \in \mathbb{R}^{d_h \times d_h}$. The gate $\mathbf{u}$ can filter out irrelevant information.

Remark 1 (Our ActAttn vs SoftmaxAttn). As shown in Figure 3(a), SoftmaxAttn produces highly concentrated attention scores on few stations, whereas ActAttn yields a smoother and more balanced distribution. Figure 3(b) confirms this statistically: ActAttn has higher score entropy, indicating more dispersed attention, while SoftmaxAttn exhibits a much larger maximum weight (0.405 vs. 0.023), suggesting strong single-node dominance. Moreover, SoftmaxAttn assigns more attention to masked keys, i.e., unmonitored stations without historical AQI records, which may introduce unreliable dependencies. In contrast, ActAttn focuses more on visible keys and extracts information from stations with historical observations. Overall, ActAttn alleviates attention over-concentration, reduces reliance on unreliable unmonitored stations, and learns more robust spatial dependencies, thereby improving the accuracy and stability of air quality prediction at unmonitored locations. We provide a more detailed analysis of the characteristics of the two attention mechanisms in **Appendix G.6**.

3.4.2 Geo-Semantic Inductive Biases in Query-Key Representation

ActAttn achieves linear and non-competitive spatial aggregation, but it still depends on informative query/key representations. In spatial inference, invisible stations lack historical AQI observations, leaving their query/key features with limited pollution-pattern information and making content-similarity-based neighbor discovery unreliable. To address this challenge, we introduce **Geo-Semantic Inductive Biases** into ActAttn, enriching query/key features with geography-aware and semantics-aware priors. To preserve linear complexity, we avoid explicit $N \times N$ attention biases and instead inject these priors into low-dimensional query/key features for implicit pairwise guidance.

Geographic Inductive Bias. Geographic proximity is a natural prior for air-quality inference, as nearby stations often share weather conditions, emission patterns, and pollutant transport pathways. This prior is especially useful for invisible stations, whose AQI-history-based query/key representations are weak. We encode geographic relationships using an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where each entry reflects pairwise station proximity, and project each row into the hidden space:

$$\mathbf{E}_{\text{geo}} = \mathbf{A} W_{\text{geo}} \in \mathbb{R}^{N \times d_h}, \tag{9}$$

where $W_{\text{geo}} \in \mathbb{R}^{N \times d_h}$ is a learnable projection matrix.

Semantic Inductive Bias. Beyond geographic proximity, air-quality dynamics are also influenced by high-level functional factors such as urbanization, emission sources, terrain, and wind-related dispersion patterns. These factors are often implicit and heterogeneous, making them difficult to represent using standard numerical covariates alone. We therefore use an LLM to summarize station-level contextual knowledge into structured textual profiles, which serve as semantic descriptors for functional similarity. Specifically, for each station, we prompt an LLM to generate a structured profile covering location characteristics and $PM_{2.5}$-related influencing factors. The prompt template is provided in Appendix B. Let $\mathbf{E}_{\text{LLM}} \in \mathbb{R}^{N \times d_{\text{llm}}}$ denote the resulting LLM embeddings. We reduce their dimensionality by PCA and project them into the ActAttn hidden space:

$$\mathbf{E}_{\text{sem}} = \text{PCA}(\mathbf{E}_{\text{LLM}}) W_{\text{sem}} \in \mathbb{R}^{N \times d_h}, \tag{10}$$

where W_{sem} is a learnable projection matrix.

Query-Key Biasing with Geo-Semantic Priors. We incorporate the above two inductive biases into ActAttn by enriching the query and key representations before linear aggregation:

$$\mathbf{Q} \leftarrow \mathbf{Q} + \mathbf{E}_{\text{geo}} W_q^{\text{geo}} + \mathbf{E}_{\text{sem}} W_q^{\text{sem}}, \quad \mathbf{K} \leftarrow \mathbf{K} + \mathbf{E}_{\text{geo}} W_k^{\text{geo}} + \mathbf{E}_{\text{sem}} W_k^{\text{sem}}. \tag{11}$$

where $W_q^{\text{geo}}, W_k^{\text{geo}}, W_q^{\text{sem}}, W_k^{\text{sem}} \in \mathbb{R}^{d_h \times d_h}$ are learnable projection matrices. This enables ActAttn to aggregate information using both geographic proximity and functional similarity, providing reliable station-pair guidance for invisible stations with weak content-based representations.

3.5 Theoretical Analysis

From an information-theoretic perspective, we show that inference driven by weather covariates is effective for target-absent air-quality prediction.

Proposition (Information benefit of weather covariates). Let Y_s denote the future AQI at an invisible station s, $\mathbf{m}$ denote the shared mask token replacing its historical AQI input, $\mathbf{Z}_s$ denote its local weather covariates, and $\mathbf{X}_{\text{vis}}$ denote the AQI observations from visible stations. This remark explains why weather covariates provide necessary local evidence for target-absent inference. Since the mask token is deterministic, it contains no information about station-specific future AQI: $I(Y_s; \mathbf{m}) = 0$. By contrast, when local weather is statistically coupled with air-quality dynamics, the chain rule of mutual information gives

$$I(Y_s; \mathbf{Z}_s, \mathbf{X}_{\text{vis}}) = I(Y_s; \mathbf{Z}_s) + I(Y_s; \mathbf{X}_{\text{vis}} \mid \mathbf{Z}_s) \geq I(Y_s; \mathbf{Z}_s) > 0. \tag{12}$$

Therefore,

$$I(Y_s; \mathbf{Z}_s, \mathbf{X}_{\text{vis}}) > I(Y_s; \mathbf{m}), \tag{13}$$

showing that weather covariates provide strictly more informative evidence than the mask token for invisible-station AQI inference. The detailed proof is provided in Appendix F.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate ExtrapAir using five datasets that cover both nationwide and regional air-quality monitoring networks: LargeAQ, KnowAir, GAGNN, YRD, and BTHSA. We further include two cross-continent datasets, US-NAT and EU-193 [7], to examine whether the proposed weather-bridged inference mechanism generalizes beyond China. LargeAQ [4] consists of monitoring records collected from 1,341 stations across 343 major cities in China. It covers an eight-year period from January 1, 2016, to December 31, 2023, at an hourly temporal resolution. Detailed statistics for the remaining datasets are reported in **Appendix G.1.1**.

Baselines. We compare ExtrapAir with **32** representative baselines from four categories: ❶ **Time-series Models**, ❷ **Spatiotemporal Models**, ❸ **Air Quality Models**, and ❹ **Spatial Kriging Interpolation Models**. Details are provided in **Appendix G.1.2**.

Spatial Inference Setting. We randomly mask a subset of stations as unmonitored stations, with the default masking ratio set to $r = 0.5$. In addition, we evaluate ExtrapAir under different masking ratios, with $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The AQI values of the masked stations are unavailable by default. During training, these stations are excluded from the loss computation because their ground-truth values are not observed. During testing, their AQI values are predicted and evaluated.

Implementation Details. Models are implemented in PyTorch 2.2.0 on an NVIDIA A100 GPU (40 GB). We use AdamW optimizer with learning rate 0.01, weight decay 10^{-4}, and MultiStepLR scheduler. The hidden dimension is $d_h = 64$ and the number of attention heads is $N_h = 8$. The semantic embeddings are generated by Qwen3 [8] and reduced to $2d_h$=128 via PCA before projection. Results are averaged over 10 runs with seeds 42–51, each using a different visible/invisible partition. Both the historical window and the future window are set to $T = L = 48$. All models follow the same station masking protocol. Weather covariates are used whenever a model supports covariate inputs; univariate models receive them through their standard input interface.

Evaluation Metrics. We evaluate continuous-value prediction using MAE and RMSE, and further assess coarse-grained pollution-level classification using F1 scores for three AQI levels: *Good* (AQI < 35), *Moderate* ($35 \leq$ AQI < 75), and *Poor* (AQI ≥ 75). We report all metrics under two settings: *overall evaluation* across all stations and *masked evaluation* on unseen stations only.

4.2 Main Result Analysis

Tables 1 and 9 report the main results. Time-series models perform poorly, as unmonitored stations cannot borrow information from nearby visible stations. Spatiotemporal and air-quality-specific models achieve stronger results due to spatial propagation mechanisms or weather branches. In particular, GWNet benefits from multi-view convolution and captures richer spatial information. However, these methods are still limited by shallow use of weather information and content-based attention, leading to larger errors at unmonitored stations. Spatial kriging methods also perform poorly when adapted to forecasting at unmonitored locations, indicating that graph-based interpolation of observed signals is insufficient for future AQI inference. Compared with the best competing baseline, ExtrapAir achieves consistent improvements in both regression and classification metrics, and it reduces the overall error by up to **15.2%** and the error at unmonitored stations by up to **16.0%**. For AQI-level prediction, ExtrapAir achieves the largest improvement on the Polluted category, with a relative F1-score gain of **33.7%**.

We further evaluate robustness under different ratios of unmonitored stations, as shown in Figure 4. As available information decreases, the prediction errors of most models increase rapidly, especially for Transformer-based models such as AirFormer and CauAir. Existing models consistently exhibit larger errors at unmonitored stations. In contrast, ExtrapAir remains more stable than most baselines as the mask ratio increases, indicating that weather covariates help compensate for missing AQI histories at unmonitored stations.

4.3 Ablation Study

Figure 5 reports the overall and unmonitored-station performance of different ablation variants, which confirm the effectiveness of each component in ExtrapAir. Replacing ActAttn with softmax leads to

Table 1: Performance on **LargeAQ** under mask ratio r=0.5. The remaining four Chinese datasets are presented in **Appendix Table 9**, and the additional US-NAT and EU-193 datasets are presented in **Appendix Table 10**. 'Hx–y' indicates the forecasting interval from the x-th time step to the y-th time step. The **best**, second, and third results are highlighted.

Method	H1–24 (Overall)		H1–24 (Unmonitored)		H25–48 (Overall)		H25–48 (Unmonitored)		F1 Score (H1–48)		
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	Good	Moderate	Polluted
Time-Series Models											
CATS	13.68	18.52	17.20	21.28	14.57	19.58	17.20	21.29	0.8699	0.4205	0.1757
CrossGNN	13.16	18.05	17.45	21.45	14.54	19.39	17.61	21.58	0.8658	0.4817	0.1828
CycleNet	13.18	18.11	17.20	21.28	14.52	19.53	17.20	21.29	0.8731	0.4391	0.2333
DLinear	12.81	17.72	16.35	20.73	13.91	18.88	15.98	20.56	0.8708	0.4171	0.1825
DSformer	13.13	18.01	17.47	21.47	14.43	19.32	17.53	21.51	0.8696	0.4970	0.2559
SOFTS	12.88	17.86	16.83	21.03	14.23	19.22	16.83	21.04	0.8750	0.4531	0.2449
SparseTSF	13.80	18.61	17.21	21.28	14.79	19.77	17.21	21.29	0.8677	0.4356	0.2380
TimeMixer	13.10	18.05	17.20	21.28	14.44	19.45	17.21	21.29	0.8742	0.4382	0.2254
TimeXer	12.35	17.18	15.72	20.15	13.28	18.02	15.41	19.98	0.8812	0.5234	0.2187
Spatiotemporal Models											
AGCRN	15.87	20.36	17.14	21.36	15.90	20.40	17.18	21.41	0.8023	0.3091	0.0000
ASTGCN	12.47	16.89	15.80	19.68	13.11	17.44	15.08	18.95	0.8535	0.6216	0.1066
BigST	12.61	17.24	16.56	20.54	13.16	17.73	15.43	19.68	0.8569	0.5889	0.1301
D^{2}STGNN	12.91	17.53	17.51	21.25	14.00	18.42	17.54	21.19	0.7665	0.5528	0.2172
GWNet	11.37	16.13	13.82	18.37	12.38	17.07	13.43	17.99	0.8914	0.5936	0.1303
RPMixer	20.14	31.39	31.73	42.42	25.70	44.93	40.77	61.59	0.7087	0.5120	0.0941
STAEformer	12.59	17.38	17.19	21.25	13.47	18.15	16.90	21.02	0.8711	0.5046	0.2465
STGCN	12.48	17.28	16.86	21.03	13.50	18.22	16.76	20.98	0.8821	0.4923	0.2426
STGODE	13.52	17.99	17.15	21.26	14.07	18.60	17.01	21.16	0.8732	0.4489	0.2020
STID	12.65	17.37	16.65	20.67	13.71	18.24	16.47	20.43	0.8565	0.5772	0.1352
STNorm	12.08	16.87	16.04	20.19	12.64	17.65	15.08	19.90	0.8687	0.5972	0.2291
STTN	11.75	16.36	14.20	18.44	12.67	17.23	13.73	17.99	0.8925	0.6268	0.1130
Air Quality Models											
AirFormer	12.73	17.62	16.85	21.02	13.92	18.79	16.84	21.01	0.8780	0.4625	0.0947
AirPhyNet	12.99	17.97	16.47	20.81	14.09	19.10	16.48	20.83	0.8732	0.4450	0.0321
PCDCNet	12.60	17.25	15.93	19.85	13.15	17.65	15.50	19.30	0.8798	0.5830	0.1250
DeepAir	12.66	17.41	16.85	20.93	13.42	18.12	16.40	20.52	0.8750	0.5247	0.1785
GAGNN	17.25	21.40	17.21	21.28	17.26	21.41	17.21	21.29	0.8443	0.0000	0.0000
MGSFformer	13.64	18.46	17.43	21.44	14.75	19.75	17.42	21.45	0.8712	0.4273	0.2149
$PM_{2.5}$GNN	12.49	16.93	15.88	19.43	12.27	16.62	13.63	17.46	0.8761	0.6467	0.0899
CauAir	12.55	17.62	16.85	21.49	13.00	18.07	16.25	21.17	0.8418	0.4957	0.2309
Spatial Kriging Models (adapted to the forecasting setting)											
IGNNK	14.52	19.37	17.89	22.41	15.63	20.58	18.12	22.68	0.8312	0.3842	0.1205
INCREASE	13.95	18.82	17.21	21.56	14.88	19.75	17.38	21.82	0.8445	0.4128	0.1432
SATCN	13.42	18.25	16.58	20.89	14.21	19.12	16.72	21.15	0.8521	0.4356	0.1589
Ours											
Ours	**9.98**	**14.46**	**11.87**	**16.12**	**10.40**	**14.88**	**11.28**	**15.62**	**0.9110**	**0.6907**	**0.3422**

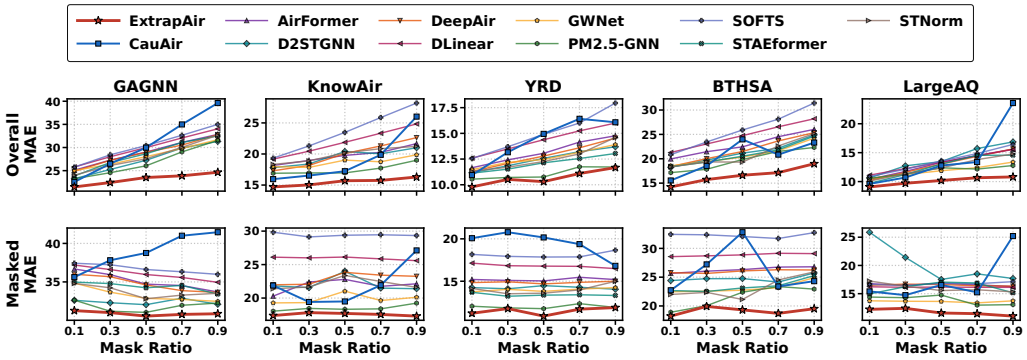


Figure 4: MAE across mask ratios. Top row: Overall; bottom row: Unmonitored-only.

the most significant degradation at masked stations, suggesting that the row-normalized competitive mechanism is particularly harmful in this setting. Removing either the spatial bias (w/o Sp.Bias) or the semantic bias (w/o Se.Bias) also degrades masked-station prediction, indicating the benefit of these prior signals. The variant w/o WeaEmb removes the station-independent encoding of weather variables in Eq. 1, causing the largest degradation in overall performance. The variant w/o PCA removes the PCA-based initialization of station embeddings in Eq. 2 and instead uses a fully random initialization, which increases the prediction error for extreme values.

4.4 Efficiency Study

Figure 6 compares ExtrapAir with representative baselines from four perspectives: prediction accuracy, training speed, memory usage, and model complexity. Graph-convolution-based architectures, such as GWNet, STNorm, and GAGNN, generally incur higher computational cost and model complexity. Transformer-based models require even longer training time, yet still produce larger

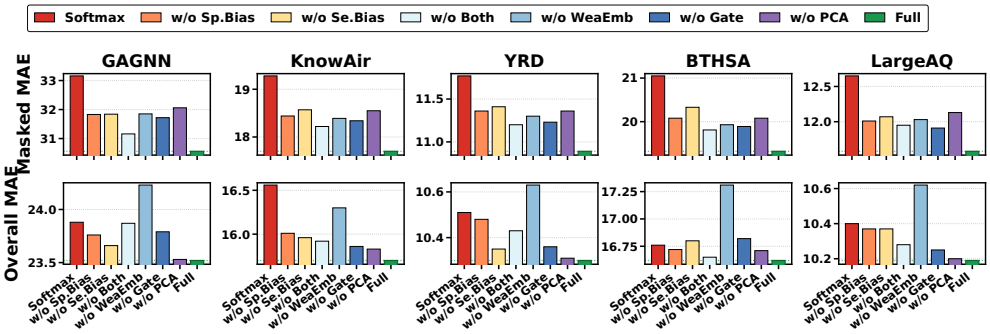


Figure 5: Ablation study on five datasets with r=0.5. Top: Masked MAE; bottom: Overall MAE.

prediction errors than ExtrapAir. Benefiting from its lightweight architecture, CauAir has the lowest complexity among the baselines while achieving competitive performance. Compared with the best-performing baseline GWNet, ExtrapAir reduces the training time by 84.9%, decreases the Overall MAE by 14.15%, and lowers memory usage by 76.23%. These results demonstrate that ExtrapAir achieves a more favorable accuracy–efficiency trade-off, improving prediction performance while substantially reducing computational and memory costs.

4.5 Case Study: Regional Robustness

We partition LargeAQ into four longitude-based regions and compare regional masked MAE under r=0.5. Figure 7 shows that baselines degrade sharply in sparse western regions, where nearby monitored stations are limited. ExtrapAir has the smallest west-to-east gap, indicating that weather bridging and non-competitive aggregation improve spatial transfer under low monitoring density.

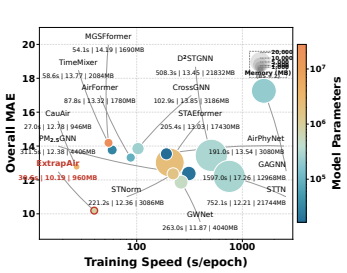


Figure 6: Efficiency study.

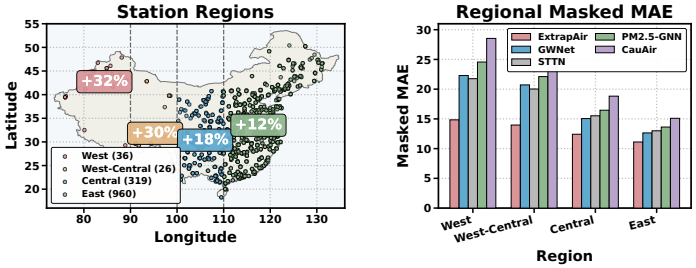


Figure 7: Regional robustness on LargeAQ.

5 Related Work

Research in air-quality prediction has shifted from physics-based dispersion systems to neural spatiotemporal models that learn pollutant dynamics from monitoring records [1, 2, 3, 9, 4, 10]. Although these methods differ in architecture, including recurrent networks, temporal convolutions, graph propagation, Transformer attention, physical priors, and multi-scale fusion, they fundamentally fail when applied to unmonitored stations without target histories. Kriging-based methods are closely related to our setting, as they estimate unknown locations through spatial interpolation or missing-value reconstruction. Classical kriging relies on geostatistical assumptions [11], while neural variants use graph encoders, temporal modules, or attention mechanisms [12, 13, 14, 15]. Recent neural kriging models such as IGNNK, INCREASE, and SATCN are inductive and generalize to nodes unseen during training. They aggregate the observed signals of neighboring sensors through graph convolution or relation-aware aggregation over spatial adjacency, and reconstruct values within the observed time window. In contrast, ExtrapAir learns weather-bridged station representations and propagates them with prior-guided, softmax-free linear attention to forecast future AQI trajectories. We discuss the relevant work in detail in **Appendix D**.

6 Conclusion

This work reframes air-quality prediction as target-absent spatial inference. Without AQI history, ExtrapAir uses weather as the key transfer signal, encoding variables separately and aligning them with AQI via dual fusion. ActAttn aggregates weak evidence from visible stations, while geographic and semantic biases preserve station-pair identity. Across seven datasets, 32 baselines, and mask ratios up to 90%, ExtrapAir achieves consistent gains with competitive efficiency, highlighting the value of physical covariates and explicit priors for sensor-sparse inference.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their constructive feedback. This paper is partially supported by the National Natural Science Foundation of China (No.12227901) and the Natural Science Foundation of Jiangsu Province (BK20250482). The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Science.

References

[1] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. Deep distributed fusion network for air quality prediction. In *Proc. of KDD*, 2018.

[2] Shuo Wang, Yanran Li, Jiang Zhang, Qingye Meng, Lingwei Meng, and Fei Gao. Pm2. 5-gnn: A domain knowledge enhanced graph neural network for pm2. 5 forecasting. In *Proceedings of the 28th international conference on advances in geographic information systems*, 2020.

[3] Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Proc. of AAAI*, 2023.

[4] Jiaming Ma, Zhiqing Cui, Binwu Wang, Pengkun Wang, Zhengyang Zhou, Zhe Zhao, and Yang Wang. Causal learning meet covariates: Empowering lightweight and effective nationwide air quality forecasting. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 3171–3179. International Joint Conferences on Artificial Intelligence Organization, 2025. doi: 10.24963/ijcai.2025/353.

[5] Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proc. of CIKM*, 2023.

[6] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proc. of ICML*, 2020.

[7] Xing Xu, Xu Wang, Yudong Zhang, Huilin Zhao, Zhengyang Zhou, and Yang Wang. AirQualityBench: A realistic evaluation benchmark for global air quality forecasting. *arXiv preprint arXiv:2605.05854*, 2026. URL https://huggingface.co/datasets/xuxing123/AirQualityBench.

[8] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[9] Kethmi Hirushini Hettige, Jiahao Ji, Shili Xiang, Cheng Long, Gao Cong, and Jingyuan Wang. Airphynet: Harnessing physics-guided neural networks for air quality prediction. In *Proc. of ICLR*, 2024.

[10] Chengqing Yu, Fei Wang, Yilun Wang, Zezhi Shao, Tao Sun, Di Yao, and Yongjun Xu. MGSFformer: A multi-granularity spatiotemporal fusion transformer for air quality prediction. *Information Fusion*, 2025.

[11] Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

[12] Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive graph neural networks for spatiotemporal kriging. In *Proc. of AAAI*, 2021.

[13] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, Jianzhong Qi, Chaochao Chen, and Longbiao Chen. INCREASE: Inductive graph representation learning for spatio-temporal kriging. In *Proc. of WWW*, pages 673–683, 2023. doi: 10.1145/3543507.3583525.

[14] Yuankai Wu, Dingyi Zhuang, Mengying Lei, Aurelie Labbe, and Lijun Sun. Spatial aggregation and temporal convolution networks for real-time kriging. *arXiv preprint arXiv:2109.12144*, 2021.

[15] Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In *Proc. of AAAI*, pages 2151–2158, 2018. doi: 10.1609/AAAI.V32I1.11871.

[16] Hua Wang, Jinghao Lu, and Fan Zhang. Eeo-tfv: Escape-explore optimizer for web-scale time-series forecasting and vision analysis. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 7271–7282, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792434. URL https://doi.org/10.1145/3774904.3792434.

[17] Fan Zhang, Shiming Fan, and Hua Wang. What if we let forecasting forget? A sparse bottleneck for cross-variable dependencies. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 155326–155347. PMLR, 06–11 Jul 2026. URL https://proceedings.mlr.press/v306/zhang26at.html.

[18] Fan Zhang, Shiming Fan, and Hua Wang. TimeSAF: Towards LLM-guided semantic asynchronous fusion for time series forecasting. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26267–26284, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1208. URL https://aclanthology.org/2026.acl-long.1208/.

[19] Fan Zhang, Shiming Fan, and Hua Wang. Rethinking activation function: A simple path to efficient and accurate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 38(10):6987–7004, 2026. doi: 10.1109/TKDE.2026.3716656.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proc. of NeurIPS*, 2017.

[21] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proc. of AAAI*, 2023.

[22] Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. Sparsetsf: Modeling long-term time series forecasting with 1k parameters. *arXiv preprint arXiv:2405.00946*, 2024.

[23] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*, 2024.

[24] Jiaming Ma, Qihe Huang, Haofeng Ma, Guanjun Wang, Sheng Huang, Zhengyang Zhou, Pengkun Wang, Xu Wang, Binwu Wang, and Yang Wang. PHAT: Modeling period heterogeneity for multivariate time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/dbb5180957513805ebeea787b8c66ac9-Abstract-Conference.html.

[25] Jiecheng Lu, Xu Han, Yan Sun, and Shihao Yang. Cats: Enhancing multivariate time series forecasting by constructing auxiliary time series as exogenous variables. *arXiv preprint arXiv:2403.01673*, 2024.

[26] Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. *In Proc. of NeurIPS*, 36:46885–46902, 2023.

[27] Chengqing Yu, Fei Wang, Zezhi Shao, Tao Sun, Lin Wu, and Yongjun Xu. DSformer: A double sampling transformer for multivariate time series long-term prediction. In *Proc. of CIKM*, 2023.

[28] Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. Softs: Efficient multivariate time series forecasting with series-core fusion. *arXiv preprint arXiv:2404.14197*, 2024.

[29] Jiaming Ma, Binwu Wang, Qihe Huang, Guanjun Wang, Pengkun Wang, Zhengyang Zhou, and Yang Wang. MoFo: Empowering long-term time series forecasting with periodic pattern modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2844.

[30] Binwu Wang, Gaoyun Lin, Jiaming Ma, Qihe Huang, Zhengyang Zhou, Xu Wang, Pengkun Wang, and Yang Wang. Crisp: A spectral-based interaction strategy for multivariate time series forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 128034–128050. PMLR, 2026. URL https://proceedings.mlr.press/v306/wang26ec.html.

[31] Zhipeng Liu, Peibo Duan, Xuan Tang, Baixin Li, Yongsheng Huang, Mingyang Geng, Changsheng Zhang, Bin Zhang, and Binwu Wang. TimeFormer: Transformer with attention modulation empowered by temporal characteristics for time series forecasting. *Expert Systems with Applications*, 307:131040, 2026. doi: 10.1016/j.eswa.2025.131040.

[32] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (18):18780–18788, 2025. doi: 10.1609/aaai.v39i18.34067.

[33] Chenxi Liu, Hao Miao, Qianxiong Xu, Shaowen Zhou, Cheng Long, Yan Zhao, Ziyue Li, and Rui Zhao. Efficient multivariate time series forecasting via calibrated language models with privileged knowledge distillation. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 3165–3178. IEEE, 2025. doi: 10.1109/ICDE65448.2025.00237.

[34] Yisong Fu, Zezhi Shao, Chengqing Yu, Yujie Li, Zhulin An, Qi Wang, Yongjun Xu, and Fei Wang. Selective learning for deep time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3277.

[35] Chenxi Liu, Shaowen Zhou, Qianxiong Xu, Hao Miao, Cheng Long, Ziyue Li, and Rui Zhao. Towards cross-modality modeling for time series analytics: A survey in the LLM era. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 10564–10572, 2025. doi: 10.24963/ijcai.2025/1173.

[36] Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, Gao Cong, Christian S. Jensen, and Xueqi Cheng. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):291–305, 2025. doi: 10.1109/TKDE.2024.3484454.

[37] Yisong Fu, Fei Wang, Zezhi Shao, Boyu Diao, Lin Wu, Zhulin An, Chengqing Yu, Yujie Li, and Yongjun Xu. On the integration of spatial-temporal knowledge: A lightweight approach to atmospheric time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2760.

[38] Binwu Wang, Jiaming Ma, Pengkun Wang, Xu Wang, Yudong Zhang, Zhengyang Zhou, and Yang Wang. STONE: A spatio-temporal OOD learning framework kills both spatial and temporal shifts. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2948–2959, 2024. doi: 10.1145/3637528.3671680.

[39] Jiaming Ma, Binwu Wang, Guanjun Wang, Kuo Yang, Zhengyang Zhou, Pengkun Wang, Xu Wang, and Yang Wang. Less but more: Linear adaptive graph learning empowering spatiotemporal forecasting. *Advances in Neural Information Processing Systems*, 38:58382–58418, 2025. doi: 10.52202/085713-1958.

[40] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[41] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

[42] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Proc. of NeurIPS*, 2020.

[43] Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S. Jensen. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. *Proc. VLDB Endow.*, 2022.

[44] Jindong Han, Weijia Zhang, Hao Liu, Tao Tao, Naiqiang Tan, and Hui Xiong. Bigst: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proceedings of the VLDB Endowment*, 2024.

[45] Guanjun Wang, Binwu Wang, Jiaming Ma, Zhengyang Zhou, Pengkun Wang, Xu Wang, and Yang Wang. One for two: A unified framework for imbalanced graph classification via dynamic balanced prototype. In *International Conference on Learning Representations*, volume 2026, pages 97210–97247, 2026.

[46] Binwu Wang, Yudong Zhang, Xu Wang, Pengkun Wang, Zhengyang Zhou, Lei Bai, and Yang Wang. Pattern expansion and consolidation on evolving graphs for continual traffic prediction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2223–2232, 2023. doi: 10.1145/3580305.3599463.

[47] Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. PromptST: Prompt-enhanced spatio-temporal multi-attribute prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3195–3205. ACM, 2023. doi: 10.1145/3583780.3615016.

[48] Zijian Zhang, Ze Huang, Zhiwei Hu, Xiangyu Zhao, Wanyu Wang, Zitao Liu, Junbo Zhang, S. Joe Qin, and Hongwei Zhao. MLPST: MLP is all you need for spatio-temporal prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3381–3390. ACM, 2023. doi: 10.1145/3583780.3614969.

[49] Minlan Shao, Zijian Zhang, Yili Wang, Yiwei Dai, Xu Shen, and Xin Wang. HyperD: Hybrid periodicity decoupling framework for traffic forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(18):15689–15697, 2026. doi: 10.1609/aaai.v40i18.38599.

[50] Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1567–1577. ACM, 2022. doi: 10.1145/3534678.3539396.

[51] Hao Miao, Yan Zhao, Chenjuan Guo, Bin Yang, Kai Zheng, Feiteng Huang, Jiandong Xie, and Christian S. Jensen. A unified replay-based continuous learning framework for spatio-temporal prediction on streaming data. In *2024 IEEE 40th International Conference on Data Engineering*, pages 1050–1062. IEEE, 2024. doi: 10.1109/ICDE60146.2024.00085.

[52] Chenxi Liu, Kethmi Hirushini Hettige, Qianxiong Xu, Cheng Long, Shili Xiang, Gao Cong, Ziyue Li, and Rui Zhao. ST-LLM+: Graph enhanced spatio-temporal large language models for traffic prediction. *IEEE Transactions on Knowledge and Data Engineering*, 37(8):4846–4859, 2025. doi: 10.1109/TKDE.2025.3570705.

[53] Ling Chen, Jiahui Xu, Binqing Wu, and Jianlong Huang. Group-aware graph neural network for nationwide city air quality forecasting. *ACM Transactions on Knowledge Discovery from Data*, 2023.

[54] Shuo Wang, Yun Cheng, Qingye Meng, Olga Saukh, Jiang Zhang, Jingfang Fan, Yuanting Zhang, Xingyuan Yuan, and Lothar Thiele. PCDCNet: A surrogate model for air quality forecasting with physical-chemical dynamics and constraints. *arXiv preprint arXiv:2505.19842*, 2025. doi: 10.48550/arXiv.2505.19842. URL https://arxiv.org/abs/2505.19842.

[55] Jiaming Ma, Guanjun Wang, Sheng Huang, Kuo Yang, Binwu Wang, Pengkun Wang, and Yang Wang. Spatiotemporal causal decoupling model for air quality forecasting. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025. URL https://ieeexplore.ieee.org/abstract/document/11099015/.

[56] Stefania Renna, Carlos Rodriguez-Pardo, and Lara Aleluia Reis. A dataset of harmonized global air quality monitoring metadata. *Scientific Data*, 13(1):466, 2026. doi: 10.1038/s41597-026-06797-0.

[57] Chengqing Yu, Fei Wang, Zezhi Shao, Tangwen Qian, Zhao Zhang, Wei Wei, and Yongjun Xu. GinAR: An end-to-end multivariate time series forecasting model suitable for variable missing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3989–4000. ACM, 2024. doi: 10.1145/3637528.3672055.

[58] Chengqing Yu, Fei Wang, Chuanguang Yang, Zezhi Shao, Tao Sun, Tangwen Qian, Wei Wei, Zhulin An, and Yongjun Xu. Merlin: Multi-view representation learning for robust multivariate time series forecasting with unfixed missing rates. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 3633–3644. ACM, 2025. doi: 10.1145/3711896.3737046.

[59] Yujie Li, Zezhi Shao, Chengqing Yu, Tangwen Qian, Zhao Zhang, Yifan Du, Shaoming He, Fei Wang, and Yongjun Xu. STA-GANN: A valid and generalizable spatio-temporal kriging approach. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 1726–1736. ACM, 2025. doi: 10.1145/3746252.3761045.

[60] Jiyue Li, Senzhang Wang, Jiaqiang Zhang, Hao Miao, Junbo Zhang, and Philip S. Yu. Fine-grained urban flow inference with incomplete data. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5851–5864, 2023. doi: 10.1109/TKDE.2022.3154163.

[61] Shuo Wang, Yun Cheng, Qingye Meng, Olga Saukh, Jiang Zhang, Jingfang Fan, Yuanting Zhang, Xingyuan Yuan, and Lothar Thiele. Knowair-v2: A benchmark dataset for air quality forecasting with pcdcnet, 2025. URL https://doi.org/10.5281/zenodo.15614907.

[62] U.S. Environmental Protection Agency. Air Quality System (AQS) AirData: Pre-Generated Data Files, 2026. URL https://aqs.epa.gov/aqsweb/airdata/download_files.html. Accessed: 2026-09-17.

[63] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, Andras Horanyi, Joaquin Munoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

[64] Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: enhancing time series forecasting through modeling periodic patterns. *arXiv preprint arXiv:2409.18479*, 2024.

[65] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *arXiv preprint arXiv:2402.19072*, 2024.

[66] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proc. of AAAI*, 2019.

[67] Chin-Chia Michael Yeh, Yujie Fan, Xin Dai, Uday Singh Saini, Vivian Lai, Prince Osei Aboagye, Junpeng Wang, Huiyuan Chen, Yan Zheng, Zhongfang Zhuang, et al. Rpmixer: Shaking up time series forecasting with random projections for large spatial-temporal data. In *Proc. of KDD*, 2024.

[68] Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proc. of KDD*, 2021.

[69] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proc. of CIKM*, 2022.

[70] Jinliang Deng, Xiusi Chen, Renhe Jiang, Xuan Song, and Ivor W Tsang. St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In *Proc. of KDD*, 2021.

[71] Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020.

A ExtrapAir Forward Pass

Algorithm 1: ExtrapAir Forward Pass

Input: Masked historical AQI $\widetilde{\mathbf{X}} \in \mathbb{R}^{B\times T\times N\times 1}$; past observed weather $\mathbf{Z}^p \in \mathbb{R}^{B\times T\times N\times V}$; future forecast weather $\mathbf{Z}^f \in \mathbb{R}^{B\times L\times N\times V}$; adjacency $\mathbf{A} \in \mathbb{R}^{N\times N}$; LLM station embeddings $\mathbf{E}_{\mathrm{LLM}} \in \mathbb{R}^{N\times d_{\mathrm{llm}}}$.

Output: Future AQI $\widehat{\mathbf{Y}} \in \mathbb{R}^{B\times L\times N\times 1}$.

$\mathbf{E}_{\mathrm{pca}} \leftarrow \mathrm{PCA}(\mathbf{E}_{\mathrm{LLM}})$, $\mathbf{E}_{\mathrm{geo}} \leftarrow \mathbf{A}\mathbf{W}_{\mathrm{geo}}$, $\mathbf{E}_{\mathrm{sem}} \leftarrow \mathrm{PCA}(\mathbf{E}_{\mathrm{LLM}})\mathbf{W}_{\mathrm{sem}}$; `// station priors`

$\widetilde{\mathbf{X}}_{\mathrm{aqi}} \leftarrow \mathrm{Permute}_{B,N,T}(\widetilde{\mathbf{X}})$; `// AQI stream`

$\mathbf{Z}^p_n \leftarrow \mathrm{Permute}_{B,N,T,V}(\mathbf{Z}^p)$, $\mathbf{Z}^f_n \leftarrow \mathrm{Permute}_{B,N,L,V}(\mathbf{Z}^f)$; `// weather streams`

$\widetilde{\mathbf{X}}_1 \leftarrow \mathrm{FFN}_{x_1}(\widetilde{\mathbf{X}}_{\mathrm{aqi}}) + \mathrm{RMSNorm}(\mathbf{E}_{x_1} + \mathbf{E}_{\mathrm{pca}})$,
$\widetilde{\mathbf{X}}_2 \leftarrow \mathrm{FFN}_{x_2}(\widetilde{\mathbf{X}}_{\mathrm{aqi}}) + \mathrm{RMSNorm}(\mathbf{E}_{x_2} + \mathbf{E}_{\mathrm{pca}})$;

$\overline{\mathbf{X}}_1 \leftarrow \text{Gated-FFN}_{x_1}(\widetilde{\mathbf{X}}_1)$, $\overline{\mathbf{X}}_2 \leftarrow \text{Gated-FFN}_{x_2}(\widetilde{\mathbf{X}}_2)$; `// two independent AQI branches`

$\overline{\mathbf{Z}}^p \leftarrow \mathrm{PerVarEnc}_p(\mathbf{Z}^p_n, \mathbf{E}_{z_p}, \mathbf{E}_{\mathrm{pca}})$, $\overline{\mathbf{Z}}^f \leftarrow \mathrm{PerVarEnc}_f(\mathbf{Z}^f_n, \mathbf{E}_{z_f}, \mathbf{E}_{\mathrm{pca}})$;

$\mathbf{O}_p \leftarrow \mathrm{WACorrBlock}_0(\overline{\mathbf{X}}_1, \overline{\mathbf{Z}}^p; \mathbf{E}_{\mathrm{geo}}, \mathbf{E}_{\mathrm{sem}})$; `// past-weather branch`

$\mathbf{O}_f \leftarrow \mathrm{WACorrBlock}_1(\overline{\mathbf{X}}_2, \overline{\mathbf{Z}}^f; \mathbf{E}_{\mathrm{geo}}, \mathbf{E}_{\mathrm{sem}})$; `// future-weather branch`

$w_{\mathrm{top}} \leftarrow \sigma(\alpha_{\mathrm{top}})$, $\mathbf{H}_{\mathrm{out}} \leftarrow w_{\mathrm{top}}\mathbf{O}_p + (1 - w_{\mathrm{top}})\mathbf{O}_f$;

$\widehat{\mathbf{Y}} \leftarrow \mathrm{Reshape}_{B,L,N,1}(\mathrm{FFN}_{\mathrm{dec}}(\mathbf{H}_{\mathrm{out}}))$;

`/* Operator definitions, with` $b \in \{0,1\}$ `for past/future branches */`

$\mathrm{PerVarEnc}_b(\mathbf{Z}, \mathbf{E}, \mathbf{E}_{\mathrm{pca}}) =$
$\big[\mathrm{FFN}_b(\mathbf{z}^{(1)}\mathbf{W}_{\mathrm{proj}}); \cdots; \mathrm{FFN}_b(\mathbf{z}^{(V)}\mathbf{W}_{\mathrm{proj}})\big]\mathbf{W}_{\mathrm{fuse}} + \mathrm{RMSNorm}(\mathbf{E} + \mathbf{E}_{\mathrm{pca}})$;

$\text{Gated-FFN}_b(\mathbf{H}) = (\mathrm{SiLU}(\mathbf{H}\mathbf{W}^b_1) \odot \mathrm{RMSNorm}(\mathbf{H}\mathbf{W}^b_2 + \sigma(\lambda_b)\bar{\mathbf{H}}\mathbf{W}^b_2))\mathbf{W}^b_3$, where $\bar{\mathbf{H}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{H}_i$;

$\mathbf{g}_b \leftarrow \sigma([\mathbf{x}; \mathbf{z}; \mathbf{x} \odot \mathbf{z}]\mathbf{W}^b_g)$, $\mathbf{H}^b_{\mathrm{fused}} \leftarrow \mathrm{RMSNorm}(\mathbf{x} + \mathbf{g}_b \odot \mathbf{z})$;

$\mathrm{WACorrBlock}_b(\mathbf{x}, \mathbf{z}; \mathbf{E}_{\mathrm{geo}}, \mathbf{E}_{\mathrm{sem}}) =$
$\sigma(\alpha_b)\,\mathrm{FFN}_b(\mathbf{H}^b_{\mathrm{fused}}) + (1 - \sigma(\alpha_b))\,\mathrm{ActAttn}_b(\mathbf{H}^b_{\mathrm{fused}}; \mathbf{E}_{\mathrm{geo}}, \mathbf{E}_{\mathrm{sem}})$;

$\mathbf{Q}_b \leftarrow \mathrm{SiLU}(\mathbf{H}^b_{\mathrm{fused}}\mathbf{W}^b_q)$, $\mathbf{K}_b \leftarrow \mathrm{SiLU}(\mathbf{H}^b_{\mathrm{fused}}\mathbf{W}^b_k)$, $\mathbf{V}_b \leftarrow \mathrm{SiLU}(\mathbf{H}^b_{\mathrm{fused}}\mathbf{W}^b_v)$,
$\mathbf{u}_b \leftarrow \mathrm{SiLU}(\mathbf{H}^b_{\mathrm{fused}}\mathbf{W}^b_u)$;

$\mathbf{Q}_b \leftarrow \mathbf{Q}_b + \mathbf{E}_{\mathrm{geo}}\mathbf{W}^b_{q,\mathrm{geo}} + \mathbf{E}_{\mathrm{sem}}\mathbf{W}^b_{q,\mathrm{sem}}$, $\mathbf{K}_b \leftarrow \mathbf{K}_b + \mathbf{E}_{\mathrm{geo}}\mathbf{W}^b_{k,\mathrm{geo}} + \mathbf{E}_{\mathrm{sem}}\mathbf{W}^b_{k,\mathrm{sem}}$;

$\mathrm{ActAttn}_b(\mathbf{H}^b_{\mathrm{fused}}; \mathbf{E}_{\mathrm{geo}}, \mathbf{E}_{\mathrm{sem}}) = \mathrm{RMSNorm}\big(\mathbf{Q}_b(\mathbf{K}_b^\top\mathbf{V}_b)\big) \odot \mathbf{u}_b$;

return $\widehat{\mathbf{Y}}$;

B LLM Prompt Template

The following structured prompt is used to generate station-level semantic embeddings via Qwen-3. For each monitoring station, we provide the city name and request a JSON-structured profile:

```
You are an air quality expert. For the city
"{city_name}", generate a structured JSON:
```

```
{
  "location_profile": {
    "geographic_env": "terrain, basin/coastal,
      ventilation conditions",
    "regional_location": "macro region,
      transport corridors",
    "urban_characteristics": "population density,
      urbanization rate, green coverage"
  },
  "pm25_factors": {
    "emission_sources": "energy structure,
      industry, transportation, residential",
    "meteorological": "wind patterns,
      climate type, inversion frequency",
    "natural_sources": "dust, biomass burning"
  }
}
```

The resulting LLM-derived station embedding serves as input to both the semantic bias projection and PCA-initialized position embeddings.

C LLM-derived Semantic Embedding Checks

This appendix details the leakage boundary and reproducibility checks for the LLM-derived semantic embeddings. The prompt does not include AQI/$PM_{2.5}$ measurements, monitoring time series, train/validation/test split information, future observations, forecasting windows, or target labels. It only asks for static contextual information such as geography, terrain, urban structure, meteorological background, and potential emission-source types. Thus, the semantic-prior pipeline does not use explicit dataset observations or future labels. Table 2 gives an abridged example of a generated station profile, and Table 3 summarizes the leakage boundary of the generated station profiles. City names may still activate general world knowledge, so we further evaluate city-name-free, shuffled-semantic, and no-semantic controls.

To evaluate profile-generation reproducibility, we fix the prompt template, independently generate five station-profile sets with GPT-5.6 (sol), encode each profile set using Qwen3-1.7B, and apply the same PCA transformation and downstream ExtrapAir pipeline. Data splits, station masks, model configuration, and training protocol are unchanged. Table 4 shows that the downstream variation is small, with the largest standard deviation being 0.0263.

We further replace Qwen3-1.7B with several public embedding backbones while keeping the generated profiles, PCA dimension, station masks, and downstream ExtrapAir pipeline fixed. As shown in Table 5, the maximum absolute difference from Qwen3-1.7B is 0.07 across the reported metrics, indicating that the semantic-prior conclusions are not tied to a particular embedding backbone.

Finally, we compare the full semantic prior with city-name-free, shuffled-semantic, and no-semantic variants. In the city-name-free setting, the LLM receives no explicit city identity and only uses non-identity structured information, including coordinates, station type, urban/rural category, land-cover characteristics, terrain, and meteorological context. Table 6 shows that removing city identity increases Overall MAE by only 0.08 on KnowAir and 0.03 on YRD, while still outperforming shuffled-semantic and no-semantic controls. This indicates that the semantic gain mainly comes from identity-free geographic and environmental context rather than memorized city-specific pollution knowledge.

D Related Work

D.1 Time Series Prediction for Air Pollution Forecasting

$PM_{2.5}$ concentration forecasting can be viewed as a time-series prediction problem, and prior air-quality forecasting studies show that temporal patterns must be modeled together with weather and spatial context [1, 2, 3, 16, 17, 18, 19]. Research on long-term time-series forecasting has also provided useful techniques for air quality prediction. Transformer-based architectures [20] have

Table 2: Representative generated station profile used for semantic embedding.

Category	Field	Generated content
Location profile: geographic environment	Terrain	Located on the northern edge of the North China Plain, with mountainous areas to the north and northwest and relatively flat plains toward the southeast.
	Basin/coastal	An inland continental city, not located in a coastal zone or enclosed basin.
	Ventilation conditions	Atmospheric circulation is influenced by surrounding mountain ranges and open plain areas, with seasonal variations in regional air exchange conditions.
Location profile: regional location	Macro region	Located in northern China within the Beijing–Tianjin–Hebei metropolitan region.
	Transport corridors	Connected with major national transportation networks, including highways, railways, and urban transportation corridors linking northern China.
Location profile: urban characteristics	Population density	A highly populated metropolitan area with concentrated residential, commercial, and administrative activities.
	Urbanization rate	A highly urbanized city with extensive built-up areas and developed urban infrastructure.
	Green coverage	Contains extensive urban parks, ecological spaces, roadside vegetation, and urban greening projects.
$PM_{2.5}$ factors: emission sources	Energy structure	The city has undergone long-term transitions in residential and industrial energy systems, with increasing adoption of cleaner energy technologies.
	Industry	The surrounding metropolitan region includes diverse economic activities such as manufacturing, services, and technology-related industries.
	Transportation	The city has intensive transportation activity with a large network of private vehicles, public transit systems, and logistics infrastructure.
	Residential	Residential activities include dense housing areas, commercial zones, and diverse energy consumption patterns associated with urban living.
$PM_{2.5}$ factors: meteorological	Wind patterns	The region experiences seasonal wind variations associated with the East Asian monsoon system, with different prevailing wind conditions across seasons.
	Climate type	Characterized by a temperate continental monsoon climate, with distinct seasonal changes including cold dry winters and hot humid summers.
	Inversion frequency	Atmospheric stability conditions vary seasonally, with more frequent stable boundary-layer conditions during colder months.
$PM_{2.5}$ factors: natural sources	Dust	The city can be influenced by natural dust activity from northern arid and semi-arid regions, particularly during spring seasons.
	Biomass burning	The surrounding agricultural regions experience seasonal changes in land management activities.
	Other sources	Natural environmental factors include seasonal vegetation changes, soil characteristics, and regional ecological conditions.

Table 3: Leakage-boundary check for generated station profiles.

Check item	Included in profile?	Implication
AQI/$PM_{2.5}$ measurements	No	No direct dataset-label leakage
Monitoring time series	No	No temporal leakage
Train/validation/test split	No	No split leakage
Future observations or prediction labels	No	No future-label leakage
General city-level background knowledge	Yes	Evaluated via city-name-free control

Table 4: Repeated LLM profile generation with fixed prompt and Qwen3-1.7B embeddings.

Setting	KnowAir				YRD			
	Overall MAE	Overall RMSE	Masked MAE	Masked RMSE	Overall MAE	Overall RMSE	Masked MAE	Masked RMSE
GPT-5.6 (sol) profiles + Qwen3-1.7B embedding, 5 generations	15.7241±0.0187	24.0036±0.0214	17.7328±0.0226	26.5917±0.0263	10.3134±0.0126	14.9729±0.0178	10.9086±0.0152	15.7394±0.0207

Table 5: Reproducibility across embedding backbones. Values in parentheses denote the absolute difference from Qwen3-1.7B.

Embedding backbone	KnowAir MAE/RMSE	YRD MAE/RMSE
Qwen3-1.7B	15.70/23.98	10.30/14.96
Qwen3-8B	15.72 (+0.02) / 23.97 (-0.01)	10.29 (-0.01) / 14.98 (+0.02)
Qwen3-14B	15.67 (-0.03) / 23.94 (-0.04)	10.26 (-0.04) / 14.93 (-0.03)
Llama-3.1-8B-Instruct	15.74 (+0.04) / 24.00 (+0.02)	10.32 (+0.02) / 15.01 (+0.05)
Gemma-2-9B-It	15.76 (+0.06) / 24.02 (+0.04)	10.33 (+0.03) / 15.03 (+0.07)

become influential in long-sequence modeling, while recent efficient forecasting models improve temporal representation through decomposition, sparse structure, channel mixing, or covariate-aware design [21, 22, 23, 24, 25, 26, 27, 28, 29]. Recent multivariate forecasting methods further exploit

Table 6: City-name-free, shuffled-semantic, and no-semantic controls.

Setting	KnowAir MAE/RMSE	YRD MAE/RMSE
City-name-free	15.78/24.06	10.33/14.99
Shuffled semantic	15.94/24.31	10.34/15.04
No semantic	15.96/24.34	10.35/15.05
Full semantic prior	15.70/23.98	10.30/14.96

structural priors: Crisp [30] selects interactions using spectral compatibility. TimeFormer [31] incorporates temporal characteristics into self-attention. Recent studies have also investigated representation learning, knowledge distillation, and selective training for multivariate forecasting [32, 33, 34], with surveys of cross-modal time-series modeling [35] and benchmarking across heterogeneous time series [36]. STELLA [37] integrates geographic and temporal information in a lightweight model for atmospheric time-series forecasting. Nevertheless, models that only focus on temporal dependencies usually overlook spatial interactions among monitoring stations, which are crucial for pollutant transport and diffusion. This limitation motivates unified spatiotemporal forecasting frameworks.

D.2 Spatiotemporal Graph Convolutional Network for Air Pollution Forecasting

Air pollution at one monitoring site is not only affected by local emissions and weather conditions, but also by pollutant transport from nearby regions [38, 39]. Therefore, modeling spatiotemporal dependencies is essential for accurate air quality forecasting. Compared with grid-based or implicit spatial modeling strategies, graph neural networks provide a flexible representation by treating monitoring stations as nodes and their relationships as edges [40, 41]. In early GNN-based forecasting studies, the spatial graph was usually predefined according to geographic distance or neighborhood proximity. Although these fixed graphs are simple and interpretable, they are unable to fully represent the complex and dynamic movement of pollutants. To overcome this limitation, adaptive and dynamic graph learning methods allow spatial correlations to vary across stations or over time [42, 43, 5, 44]. Recent graph learning work also addresses imbalanced supervision with dynamic balanced prototypes [45], which is complementary to our target-absent inference focus. Beyond air-quality forecasting, continual learning on evolving graphs has been used to accommodate expanding traffic networks [46]. Related approaches explore multi-attribute, lightweight, and periodicity-aware forecasting [47, 48, 49], as well as pretraining and continual learning for spatiotemporal prediction [50, 51]. Language-model-based approaches have also been explored for spatiotemporal traffic prediction [52]. These methods offer complementary strategies for spatiotemporal prediction; our work focuses on transferring information to locations without historical target AQI observations.

D.3 Dedicated Air Pollution Forecasting Model

Early air pollution forecasting systems were mainly based on physical simulation models and expert-designed domain knowledge. As urban sensor networks have expanded and machine learning techniques have developed rapidly, data-driven forecasting methods have become an increasingly important alternative. In particular, deep learning models have achieved substantial improvements in air quality prediction accuracy [1, 2, 3]. Motivated by the success of spatiotemporal graph convolutional networks, many studies have designed air-pollution-specific forecasting frameworks that formulate air quality observations as graph-structured data, where temporal dependencies are modeled by RNNs or Transformers and spatial interactions are captured through graph convolution or attention mechanisms [53, 54, 10, 4]. Another important research direction is physics-informed or causality-aware air quality forecasting, where physical constraints, differential-equation-inspired modules, or causal decoupling mechanisms are introduced to describe pollutant evolution and meteorology–AQI interactions [9, 55]. Although these approaches can improve interpretability, their additional modeling assumptions and computation make scalability on large monitoring networks a practical concern.

Kriging and target-absent inference. Spatial inference is commonly approached through kriging or missing-value reconstruction. Classical kriging estimates unknown values from nearby observations under geostatistical assumptions [11], while neural variants replace hand-crafted interpolation with graph encoders, temporal modules, or attention [12, 13, 14, 15]. Broader spatiotemporal backbones

also learn spatial dependence through adaptive graphs, diffusion, or self-attention [40, 41, 42, 43, 5, 44]. Among neural kriging methods, IGNNK [12], INCREASE [13], and SATCN [14] already adopt inductive training and generalize to nodes unseen during training without retraining. They differ from ExtrapAir mainly in method design: they propagate observed signals over a spatial graph, e.g., diffusion graph convolution in IGNNK, multi-relation aggregation in INCREASE, and spatial aggregation with temporal convolution in SATCN, and output reconstructed values for the observed period. ExtrapAir instead couples weather-bridged station representations with prior-guided, softmax-free spatial aggregation, so that each unmonitored location collects complementary evidence from multiple visible stations, and decodes future AQI trajectories. Recent harmonized station-metadata resources also improve the consistency and reproducibility of global monitoring descriptors [56]. Such structured metadata can complement, and in some cases partially replace, LLM-derived semantic profiles when sufficiently standardized station-level fields are available.

Related work also studies incomplete sensing in other settings: GinAR [57] combines interpolation attention and graph convolution for forecasting with missing variables, with related work considering varying missing rates [58]. STA-GANN [59] uses dynamic graph modeling and adversarial learning for spatiotemporal kriging at unknown sensors; other work considers urban-flow inference from incomplete observations [60].

E Analysis of Prior-guided Station-pair Discriminability

This section explains why injecting geographic and semantic priors into ActAttn improves station-pair discriminability, especially for invisible stations whose AQI-history-based representations collapse to a shared mask token. We show that prior injection changes the effective attention kernel from a purely content-based similarity to a prior-guided station-pair kernel with content, geographic, semantic, and interaction components.

Proposition 1 (Prior-guided station-pair discriminability). Let $\mathbf{c}_i$ denote the content-based query/key representation of station i, and define the prior-enhanced representation as

$$\widetilde{\mathbf{c}}_i = \mathbf{c}_i + \mathbf{b}_i^{\mathrm{geo}} + \mathbf{b}_i^{\mathrm{sem}}. \tag{14}$$

Then, the ActAttn station-pair kernel

$$\kappa(i,j) = \phi(\widetilde{\mathbf{c}}_i)^\top \phi(\widetilde{\mathbf{c}}_j) \tag{15}$$

can be decomposed as

$$\kappa(i,j) = \kappa_{\mathrm{cnt}}(i,j) + \kappa_{\mathrm{geo}}(i,j) + \kappa_{\mathrm{sem}}(i,j) + \kappa_{\mathrm{cross}}(i,j) + \mathcal{O}(\|\mathbf{b}\|^2). \tag{16}$$

This decomposition explains why prior injection can restore station-pair discriminability when AQI-history-based content similarity becomes unreliable at invisible stations.

Proof. Let $\mathbf{c}_i$ denote the content-based query/key representation of station i. For invisible stations, historical AQI inputs are replaced by a shared mask token, making $\mathbf{c}_i$ weakly discriminative. We enhance the representation by adding geographic and semantic priors:

$$\widetilde{\mathbf{c}}_i = \mathbf{c}_i + \mathbf{b}_i^{\mathrm{geo}} + \mathbf{b}_i^{\mathrm{sem}} = \mathbf{c}_i + \mathbf{b}_i, \tag{17}$$

where

$$\mathbf{b}_i = \mathbf{b}_i^{\mathrm{geo}} + \mathbf{b}_i^{\mathrm{sem}}. \tag{18}$$

The prior-guided station-pair kernel is then defined as

$$\kappa(i,j) = \phi(\widetilde{\mathbf{c}}_i)^\top \phi(\widetilde{\mathbf{c}}_j) = \phi(\mathbf{c}_i + \mathbf{b}_i)^\top \phi(\mathbf{c}_j + \mathbf{b}_j), \tag{19}$$

where $\phi(\cdot)$ denotes the activation function used in ActAttn, e.g., $\mathrm{SiLU}(\cdot)$.

We apply a first-order Taylor expansion of ϕ around $\mathbf{c}_i$:

$$\phi(\mathbf{c}_i + \mathbf{b}_i) = \phi(\mathbf{c}_i) + \phi'(\mathbf{c}_i) \odot \mathbf{b}_i + \mathcal{O}(\|\mathbf{b}_i\|^2), \tag{20}$$

and similarly,

$$\phi(\mathbf{c}_j + \mathbf{b}_j) = \phi(\mathbf{c}_j) + \phi'(\mathbf{c}_j) \odot \mathbf{b}_j + \mathcal{O}(\|\mathbf{b}_j\|^2). \tag{21}$$

Substituting the two expansions into the prior-guided kernel gives

$$\begin{aligned}\kappa(i,j) =& \left(\phi(\mathbf{c}_i) + \phi'(\mathbf{c}_i) \odot \mathbf{b}_i + \mathcal{O}(\|\mathbf{b}_i\|^2)\right)^\top \left(\phi(\mathbf{c}_j) + \phi'(\mathbf{c}_j) \odot \mathbf{b}_j + \mathcal{O}(\|\mathbf{b}_j\|^2)\right) \\ =& \phi(\mathbf{c}_i)^\top \phi(\mathbf{c}_j) + \left(\phi'(\mathbf{c}_i) \odot \mathbf{b}_i\right)^\top \phi(\mathbf{c}_j) + \phi(\mathbf{c}_i)^\top \left(\phi'(\mathbf{c}_j) \odot \mathbf{b}_j\right) \\ &+ \left(\phi'(\mathbf{c}_i) \odot \mathbf{b}_i\right)^\top \left(\phi'(\mathbf{c}_j) \odot \mathbf{b}_j\right) + \mathcal{O}(\|\mathbf{b}\|^2). \end{aligned} \tag{22}$$

Since $\mathbf{b}_i = \mathbf{b}_i^{\text{geo}} + \mathbf{b}_i^{\text{sem}}$, the first-order bias terms can be decomposed into geographic and semantic components:

$$\left(\phi'(\mathbf{c}_i) \odot \mathbf{b}_i\right)^\top \phi(\mathbf{c}_j) = \left(\phi'(\mathbf{c}_i) \odot \mathbf{b}_i^{\text{geo}}\right)^\top \phi(\mathbf{c}_j) + \left(\phi'(\mathbf{c}_i) \odot \mathbf{b}_i^{\text{sem}}\right)^\top \phi(\mathbf{c}_j), \tag{23}$$

$$\phi(\mathbf{c}_i)^\top \left(\phi'(\mathbf{c}_j) \odot \mathbf{b}_j\right) = \phi(\mathbf{c}_i)^\top \left(\phi'(\mathbf{c}_j) \odot \mathbf{b}_j^{\text{geo}}\right) + \phi(\mathbf{c}_i)^\top \left(\phi'(\mathbf{c}_j) \odot \mathbf{b}_j^{\text{sem}}\right). \tag{24}$$

We therefore define the content component as

$$\kappa_{\text{cnt}}(i,j) = \phi(\mathbf{c}_i)^\top \phi(\mathbf{c}_j), \tag{25}$$

the geographic component as

$$\kappa_{\text{geo}}(i,j) = \left(\phi'(\mathbf{c}_i) \odot \mathbf{b}_i^{\text{geo}}\right)^\top \phi(\mathbf{c}_j) + \phi(\mathbf{c}_i)^\top \left(\phi'(\mathbf{c}_j) \odot \mathbf{b}_j^{\text{geo}}\right), \tag{26}$$

and the semantic component as

$$\kappa_{\text{sem}}(i,j) = \left(\phi'(\mathbf{c}_i) \odot \mathbf{b}_i^{\text{sem}}\right)^\top \phi(\mathbf{c}_j) + \phi(\mathbf{c}_i)^\top \left(\phi'(\mathbf{c}_j) \odot \mathbf{b}_j^{\text{sem}}\right). \tag{27}$$

The remaining second-order and mixed terms are collected into

$$\kappa_{\text{cross}}(i,j) = \left(\phi'(\mathbf{c}_i) \odot \mathbf{b}_i\right)^\top \left(\phi'(\mathbf{c}_j) \odot \mathbf{b}_j\right). \tag{28}$$

Thus, the prior-guided kernel can be written as

$$\kappa(i,j) = \kappa_{\text{cnt}}(i,j) + \kappa_{\text{geo}}(i,j) + \kappa_{\text{sem}}(i,j) + \kappa_{\text{cross}}(i,j) + \mathcal{O}(\|\mathbf{b}\|^2). \tag{29}$$

This decomposition shows that geo-semantic prior injection does not merely append auxiliary station features. Instead, it changes the effective station-pair kernel used by ActAttn. When the AQI-history-based content term $\kappa_{\text{cnt}}(i,j)$ becomes weak or mask-dominated for invisible stations, the geographic term $\kappa_{\text{geo}}(i,j)$ and semantic term $\kappa_{\text{sem}}(i,j)$ remain station-pair dependent. Therefore, invisible stations can still obtain discriminative guidance from spatial proximity and functional similarity, enabling more reliable transfer of weather–AQI dependencies from visible stations. □

F Proof of the Information Benefit of Weather Covariates

Since the mask token $\mathbf{m}$ is a fixed learnable vector shared by invisible stations, it is deterministic with respect to the target variable. Therefore, conditioning on $\mathbf{m}$ does not reduce the uncertainty of Y_s:

$$I(Y_s; \mathbf{m}) = H(Y_s) - H(Y_s \mid \mathbf{m}) = H(Y_s) - H(Y_s) = 0. \tag{30}$$

Next, by the chain rule of mutual information, we have

$$I(Y_s; \mathbf{Z}_s, \mathbf{X}_{\text{vis}}) = I(Y_s; \mathbf{Z}_s) + I(Y_s; \mathbf{X}_{\text{vis}} \mid \mathbf{Z}_s). \tag{31}$$

Since conditional mutual information is non-negative,

$$I(Y_s; \mathbf{X}_{\text{vis}} \mid \mathbf{Z}_s) \geq 0, \tag{32}$$

which gives

$$I(Y_s; \mathbf{Z}_s, \mathbf{X}_{\text{vis}}) \geq I(Y_s; \mathbf{Z}_s). \tag{33}$$

Finally, if future AQI is not independent of local weather conditions, i.e.,

$$Y_s \not\perp\!\!\!\perp \mathbf{Z}_s, \tag{34}$$

then

$$I(Y_s; \mathbf{Z}_s) > 0. \tag{35}$$

This assumption is physically reasonable because wind affects pollutant dispersion, precipitation removes particulate matter, and temperature or humidity can influence chemical transformation and accumulation. Combining the above results yields

$$I(Y_s; \mathbf{Z}_s, \mathbf{X}_{\text{vis}}) \geq I(Y_s; \mathbf{Z}_s) > I(Y_s; \mathbf{m}) = 0. \tag{36}$$

Thus, weather covariates provide positive information about future AQI at invisible stations, whereas the shared mask token does not.

G Experiment

G.1 Experiment Setup

G.1.1 Dataset Details

The statistical characteristics of the datasets are shown in Table 7. LargeAQ [4] is the largest open-source dataset to date specifically designed for long-term air pollution forecasting. Compared with existing datasets, LargeAQ expands the spatiotemporal scale by nearly an order of magnitude and comprehensively covers two key types of information: $PM_{2.5}$ monitoring data and weather variables.

❶ **$PM_{2.5}$ Data.** The $PM_{2.5}$ data used in this study are obtained from the real-time air quality monitoring platform maintained by the China National Environmental Monitoring Centre (CNEMC). This platform provides hourly pollutant monitoring data from urban monitoring stations and weather stations across China. The dataset covers 343 major cities in China and 1,341 monitoring stations in total. The spatial distribution of these stations is shown in Figure 1(a). Temporally, the dataset spans eight years, from January 1, 2016 to December 31, 2023, with an hourly sampling frequency, resulting in more than 70,000 discrete time steps. As is commonly observed in environmental datasets, $PM_{2.5}$ concentrations exhibit pronounced seasonal patterns.

❷ **Weather Data.** The weather data are primarily provided by the China Meteorological Administration and Nanjing University of Information Science and Technology. In addition, we supplement the data with ERA5 climate reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and publicly available data from the National Climatic Data Center (NCDC) to fill missing values. The final dataset is systematically extracted and refined from a large-scale archive, which contains over ten years of continuous observations from 2,000 real geographically distributed monitoring stations. This process yields a high-quality dataset with complete observational records for all measured variables. LargeAQ integrates several key weather variables, including temperature, humidity, precipitation, wind direction, and wind speed. Given the significant influence of these external factors on air quality, we process these variables as follows to effectively utilize their information:

- **Temperature and Humidity:** These variables are directly used as fundamental indicators that reflect atmospheric stability and diffusion conditions.
- **Precipitation Processing:** We compute the accumulated precipitation over the preceding 24 hours to provide an accurate indicator for long-term forecasting, characterizing rainfall intensity and its associated wet deposition effects.
- **Conversion of Wind Direction and Wind Speed into UV Components:** Wind direction and wind speed are converted into u east–west and v north–south components, thereby eliminating angular discontinuities and providing a continuous wind-field representation for pollutant dispersion modeling.

Table 7: Summary of the seven evaluation datasets. Volume = # Steps × # Stations × Dim.

Dataset	# Stations	Region	# Steps	Dim	Volume
LargeAQ [4]	1,341	Nationwide China	70,128	8	752.1M
KnowAir [2]	184	China	11,688	13	27.9M
GAGNN [53]	209	China	20,373	10	42.6M
YRD [61]	127	Yangtze River Delta	70,128	9	80.1M
BTHSA [61]	228	Bohai Sea Area	70,128	9	143.9M
US-NAT [62]	392	United States	17,520	7	48.1M
EU-193 [7]	193	Europe	17,544	6	20.3M

❸ **Additional Cross-Continent Datasets.** We additionally evaluate on US-NAT and EU-193 to assess transfer across monitoring systems outside China. US-NAT combines $PM_{2.5}$ observations from the U.S. Environmental Protection Agency Air Quality System (AQS) AirData archive [62] with ERA5 meteorological reanalysis fields [63]. EU-193 uses European $PM_{2.5}$ records from AirQualityBench [7], with weather covariates derived from ERA5 reanalysis.

G.1.2 Baselines

We compare ExtrapAir with **32** baselines from four categories: ❶ **Time-series Models**: CATS [25], CrossGNN [26], CycleNet [64], DLinear [21], DSformer [27], SOFTS [28], SparseTSF [22], TimeMixer [23], TimeXer [65]. ❷ **Spatiotemporal models**: AGCRN [42], ASTGCN [66], BigST [44], D^{2}STGNN [43], GWNet [41], RPMixer [67], STAEformer [5], STGCN [40], STGODE [68], STID [69], STNorm [70], STTN [71]. ❸ **Air Quality Models**: AirFormer [3], AirPhyNet [9], DeepAir [1], GAGNN [53], MGSFformer [10], PCDCNet [54], PM$_{2.5}$GNN [2], CauAir [4]. ❹ **Spatial Kriging Models**: IGNNK [12], INCREASE [13], SATCN [14].

G.2 Efficiency Analysis

Figure 6 compares ExtrapAir with representative baselines in terms of prediction accuracy, training speed, and model complexity. ExtrapAir achieves the lowest Overall MAE of 10.19 while maintaining moderate training cost, requiring only 39.6 s/epoch and 860.5K parameters.

Compared with high-accuracy baselines such as GWNet and STNorm, ExtrapAir significantly reduces the MAE while using a comparable or even smaller number of parameters. Meanwhile, although lightweight models such as PM$_{2.5}$GNN and AirPhyNet contain fewer parameters, their prediction errors are substantially higher, indicating that extreme model compactness alone is insufficient for accurate long-term air-quality forecasting.

ExtrapAir also demonstrates a favorable balance between efficiency and accuracy when compared with larger or slower models. For example, methods such as D^{2}STGNN, STTN, and GAGNN require much longer training time, yet still produce higher errors than ExtrapAir. Compared with the average of all baselines, ExtrapAir reduces the training time from 336.9 s/epoch to 39.6 s/epoch, corresponding to an 88.2% reduction. Its parameter count is also 69.8% smaller than the baseline average, indicating substantially lower model complexity.

These results show that ExtrapAir does not obtain its performance gains by simply increasing model size or computational cost. Instead, its weather-centered design and explicit spatial priors enable more effective information transfer, leading to a superior accuracy–efficiency trade-off for long-term air-quality forecasting.

G.3 Horizon-wise Performance Analysis

Figure 8 shows per-horizon MAE for seven representative models on LargeAQ (r=0.5).

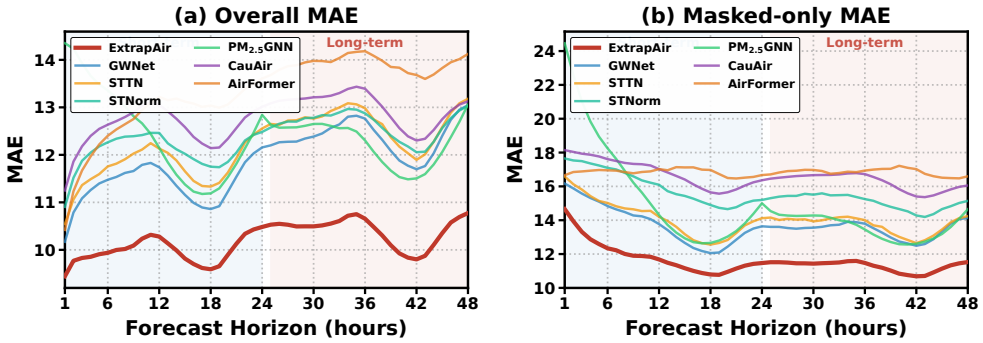


Figure 8: Per-horizon MAE on LargeAQ. ExtrapAir remains lowest across horizons.

G.4 Visible–Masked Performance Gap

Figure 9 quantifies the performance gap between visible and masked stations. The *degradation ratio* (Masked MAE / Visible MAE) measures how much a model's accuracy deteriorates when moving from well-observed to unmonitored locations. ExtrapAir achieves the lowest degradation ratio (1.32×), significantly lower than CauAir (1.84×), AirFormer (1.72×), and STNorm (1.70×). This indicates that ExtrapAir's weather-bridged attention mechanism effectively transfers learned

patterns from visible to invisible stations, whereas models relying primarily on AQI content similarity (softmax attention) or fixed graph topology suffer disproportionately at masked locations.

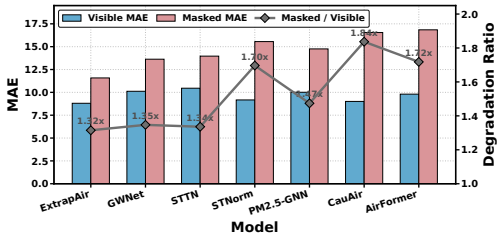


Figure 9: Visible–masked MAE gap on LargeAQ. ExtrapAir has the smallest degradation ratio.

G.5 Cross-Dataset Mask Ratio Sensitivity

Figure 10 visualizes ExtrapAir's performance across all five datasets and five mask ratios.

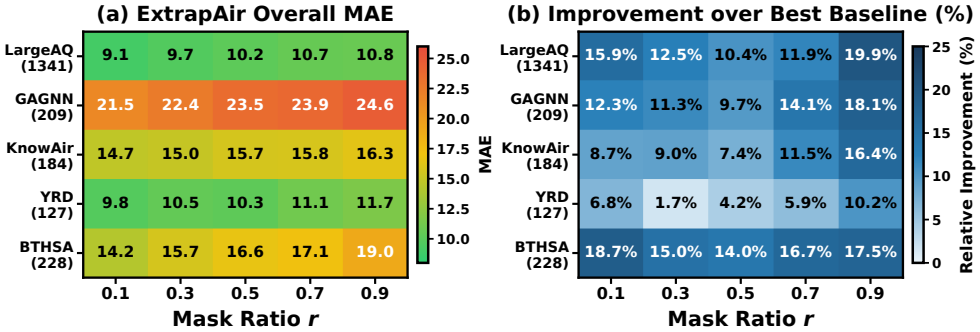


Figure 10: ExtrapAir performance and relative improvement across datasets and mask ratios.

G.6 Analysis of Attention Characteristics of Our ActAttn and Softmax

G.6.1 Additional Pairwise Score Analysis

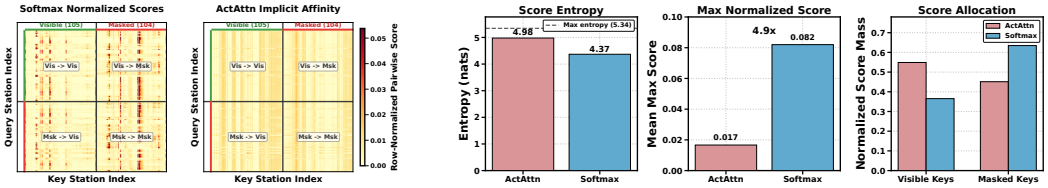
Figures 11–13 provide further evidence that ActAttn and SoftmaxAttn exhibit substantially different attention behaviors across different datasets. From the score maps, SoftmaxAttn tends to produce sparse and highly peaked attention patterns, where only a small number of stations receive dominant attention scores. In contrast, ActAttn generates smoother and more evenly distributed affinity maps, suggesting that it captures broader spatial correlations rather than relying on a few dominant nodes.

The statistical results further support this observation. On GAGNN, ActAttn achieves a higher entropy than SoftmaxAttn, i.e., 4.95 vs. 4.37, indicating that its attention distribution is more dispersed. Meanwhile, the mean maximum normalized score of SoftmaxAttn is much larger than that of ActAttn, 0.082 vs. 0.017, showing a 4.9× stronger concentration on a single key station. Similar trends are more pronounced on YRD and BTHSA. On YRD, the maximum normalized score of SoftmaxAttn reaches 0.485, which is 14.8× larger than that of ActAttn. On BTHSA, this gap further increases to 21.3× with scores of 0.298 and 0.014, respectively. These results indicate that SoftmaxAttn is more likely to suffer from attention over-concentration, especially when modeling complex regional dependencies.

In addition, the score allocation results show that SoftmaxAttn assigns a larger proportion of attention mass to masked keys, i.e., unmonitored stations without historical AQI observations. For example, on GAGNN, SoftmaxAttn allocates about 0.67 of its score mass to masked keys, while ActAttn assigns less attention to masked keys and places more emphasis on visible stations. Similar tendencies can also be observed on YRD and BTHSA. This suggests that SoftmaxAttn may introduce unreliable

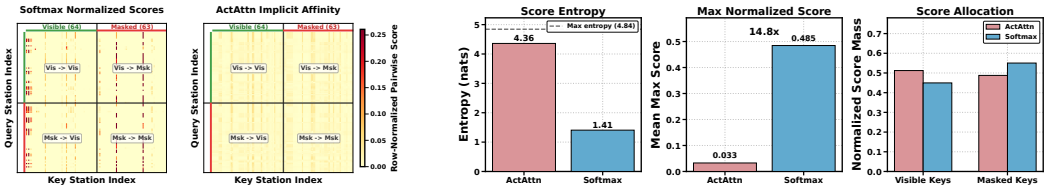
dependencies from unobserved nodes, whereas ActAttn makes better use of stations with available historical information.

Overall, ActAttn consistently alleviates the over-concentration problem of SoftmaxAttn, produces smoother spatial dependency patterns, and reduces excessive reliance on masked or unmonitored stations. These properties enable ActAttn to learn more robust and reliable spatial representations, which is particularly beneficial for forecasting air quality at unmonitored locations.



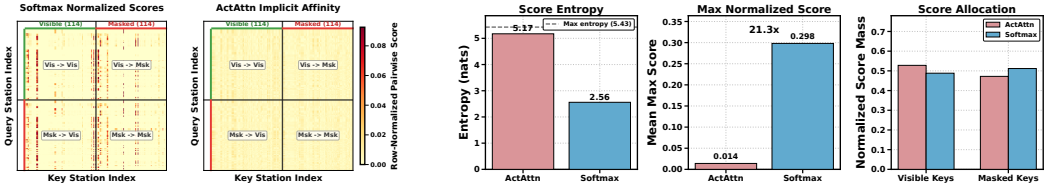
(a) Score maps (b) Masked-query score analysis

Figure 11: Additional attention analysis on GAGNN. Score maps compare Softmax (left) and ActAttn (right); masked-query statistics quantify entropy, concentration, and key allocation.



(a) Score maps (b) Masked-query score analysis

Figure 12: Additional attention analysis on YRD. Score maps compare Softmax (left) and ActAttn (right); masked-query statistics quantify entropy, concentration, and key allocation.



(a) Score maps (b) Masked-query score analysis

Figure 13: Additional attention analysis on BTHSA. Score maps compare Softmax (left) and ActAttn (right); masked-query statistics quantify entropy, concentration, and key allocation.

G.6.2 Pairwise Score Distribution Across Mask Ratios

Figure 14 presents the overall variation of masked-query scores under different mask ratios, while Figure 15 further reports the corresponding distributional patterns on KnowAir, including normalized-score entropy in the top row and maximum normalized score in the bottom row. For ActAttn, we use the implicit affinity matrix $\phi(Q)\phi(K)^\top$ solely for diagnostic analysis.

The results show that ActAttn maintains a consistently high-entropy score distribution across all mask ratios. Its entropy remains close to the theoretical upper bound, i.e., $\ln N{=}5.21$ nats, as indicated by the dashed reference line. In contrast, the entropy of Softmax gradually decreases and becomes more concentrated as the mask ratio increases, suggesting that its attention distribution becomes increasingly uneven under severe masking. This phenomenon is further verified by the max normalized score distributions. Softmax exhibits a clear growth in large maximum scores, especially at high mask ratios, indicating stronger attention concentration on a small number of keys. By comparison, ActAttn keeps its maximum scores concentrated near zero, demonstrating that it avoids excessive score dominance and preserves a more stable and balanced affinity pattern even when only limited historical observations are available.

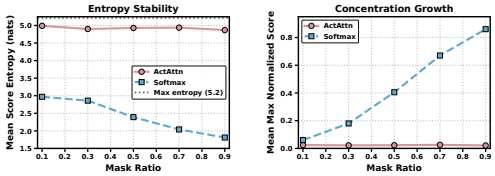


Figure 14: Masked-query score trend across mask ratios on KnowAir.

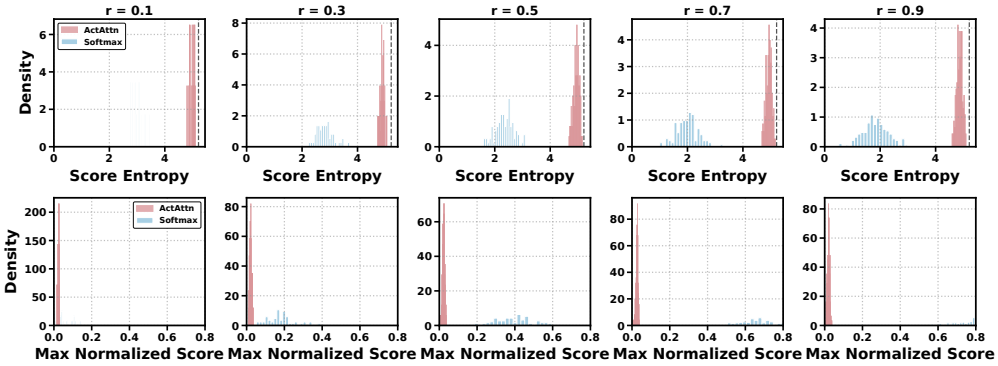


Figure 15: Normalized-score distributions on KnowAir across mask ratios. ActAttn remains broad, while Softmax becomes concentrated.

G.7 Feature Map Comparison

Table 8 reports a systematic comparison of six feature maps for ActAttn on LargeAQ under r=0.5. SiLU consistently achieves the lowest error. The elu+1 map, designed to approximate softmax, retains partial competitive behavior. ReLU suffers from dead neurons (zero gradients for negative inputs). The bounded $1+\tanh$ saturates for large inputs. Identity degenerates to a bilinear form without nonlinear feature interaction.

Table 8: Feature map comparison on LargeAQ (r=0.5, H1–48). SiLU consistently achieves the lowest error across both evaluation scopes.

Feature Map $\phi(\cdot)$	**Overall MAE**	**Masked MAE**	**Overall RMSE**	**Masked RMSE**
SiLU (ours)	**10.19**	**11.57**	**14.67**	**15.87**
elu(·)+1 [6]	10.82	12.45	15.28	16.72
ReLU	10.95	12.68	15.42	16.95
$1+\tanh(\cdot)$	10.71	12.28	15.15	16.58
Softmax (standard)	11.28	13.52	15.85	17.68
Identity (no activation)	11.45	13.85	16.02	17.92

H Clustered Geographic Holdout

Random station masking is useful for controlled evaluation, but real unmonitored regions are often geographically clustered rather than independently scattered. To better approximate this deployment regime, we construct a stricter clustered holdout on LargeAQ by masking two contiguous western subregions, denoted *West* and *West-Central*. The evaluation protocol remains target-absent: all held-out stations are invisible during training and only evaluated at test time. Table 11 shows that ExtrapAir remains the best-performing method under this harder setting, improving both overall and masked metrics over the strongest compared baselines.

I Future Weather Covariates

Table 9: Performance on the remaining four datasets under mask ratio r=0.5. 'Overall' and 'Masked' report MAE and RMSE, and $F1_\mu$ denotes micro-averaged AQI-level classification. The **best**, second, and third results are highlighted.

Method	Overall		Masked		$F1_\mu$	Overall		Masked		$F1_\mu$	Overall		Masked		$F1_\mu$	Overall		Masked		$F1_\mu$
	MAE	RMSE	MAE	RMSE		MAE	RMSE	MAE	RMSE		MAE	RMSE	MAE	RMSE		MAE	RMSE	MAE	RMSE	
	GAGNN					KnowAir					YRD					BTHSA				
Time-Series Models																				
CATS	31.84	48.99	36.81	54.75	0.4861	24.12	33.22	29.67	37.36	0.4502	15.71	21.06	18.61	23.13	0.4026	27.36	38.56	32.63	41.61	0.4391
CrossGNN	30.65	47.06	37.62	54.46	0.5466	24.40	33.09	31.26	38.31	0.4634	16.27	21.49	20.93	25.02	0.4185	27.01	37.37	34.37	42.48	0.4833
CycleNet	30.94	47.98	36.83	54.74	0.5135	23.77	32.80	29.71	37.39	0.4650	15.44	20.79	18.66	23.17	0.4126	27.02	38.08	32.80	41.71	0.4475
DLinear	29.98	47.94	35.96	55.61	0.4745	21.88	31.16	26.13	35.16	0.4468	14.38	19.94	16.78	22.06	0.4501	24.74	36.36	28.87	39.84	0.4314
DSformer	30.66	47.70	36.76	54.76	0.5176	23.76	32.71	29.89	37.51	0.4808	14.54	19.98	17.27	22.19	0.5117	26.95	37.76	33.86	42.33	0.4736
SOFTS	30.42	47.68	36.61	54.85	0.5226	23.47	32.51	29.39	37.16	0.4695	14.91	20.34	17.85	22.64	0.4818	25.90	36.98	32.06	41.28	0.4740
SparseTSF	32.31	49.19	36.83	54.74	0.4990	24.55	33.42	29.72	37.39	0.4665	15.97	21.23	18.67	23.17	0.4097	28.27	39.52	32.81	41.71	0.4387
TimeMixer	30.50	47.62	36.83	54.74	0.5244	22.25	31.70	26.96	35.66	0.4696	15.38	20.74	18.64	23.16	0.4101	26.73	37.81	32.78	41.70	0.4548
TimeXer	28.42	45.85	34.18	52.45	0.5542	18.52	28.12	20.85	30.52	0.5825	12.45	17.28	14.12	18.85	0.5952	21.25	32.15	24.52	34.85	0.5685
Spatiotemporal Models																				
AGCRN	34.63	55.68	37.49	58.94	0.3039	22.27	32.38	26.73	37.65	0.5211	16.79	22.37	17.64	23.10	0.3436	28.64	40.49	29.64	41.36	0.2315
ASTGCN	28.90	46.60	33.65	52.84	0.5317	19.63	29.01	21.60	31.44	0.5555	13.04	18.17	14.72	19.70	0.5803	22.69	33.39	26.25	36.30	0.5375
BigST	28.10	45.44	33.57	52.42	0.5621	20.28	29.52	23.09	32.59	0.5149	12.34	17.72	13.52	18.86	0.5500	21.01	32.08	23.56	34.31	0.5872
D^2STGNN	27.24	44.66	32.04	50.82	0.5860	20.40	29.45	24.00	33.07	0.5241	12.57	17.50	14.48	19.14	0.5912	21.15	31.36	24.75	34.25	0.5796
GWNet	28.02	46.13	32.86	52.77	0.5487	19.01	28.65	21.03	31.13	0.5576	12.45	17.49	13.80	18.71	0.5824	20.34	31.22	22.73	33.38	0.5999
RPMixer	33.60	52.29	44.14	63.47	0.5488	25.93	37.84	35.58	47.39	0.5352	33.93	66.09	57.23	92.33	0.4371	25.98	40.02	35.49	49.33	0.5764
STAEformer	28.65	46.86	34.24	53.88	0.4959	18.98	29.03	20.09	30.38	0.5655	12.14	17.50	13.36	18.86	0.5739	20.52	31.60	23.12	33.98	0.5961
STGCN	28.01	45.74	33.13	52.72	0.5146	20.67	29.62	24.22	33.27	0.4902	15.01	20.05	19.02	23.39	0.4527	22.01	32.59	26.13	35.85	0.5441
STGODE	28.33	45.26	33.89	52.10	0.5627	19.97	29.18	23.13	32.55	0.4866	14.25	19.26	17.55	21.99	0.5604	25.49	35.92	32.69	41.63	0.4881
STID	30.67	47.86	38.15	55.93	0.5241	19.02	28.59	21.60	31.71	0.5199	13.29	18.33	15.17	19.99	0.5511	22.92	33.57	27.19	36.89	0.5170
STNorm	27.56	45.44	32.81	52.06	0.5698	20.18	28.93	23.38	32.21	0.5336	12.26	17.36	13.59	18.56	0.5734	19.32	30.95	21.12	32.54	0.6316
STTN	26.53	43.37	30.63	48.91	0.6017	19.06	28.18	21.00	30.71	0.5789	12.10	16.94	13.20	17.89	0.6074	19.66	30.38	21.84	32.32	0.6148
Air Quality Models																				
AirFormer	29.01	46.76	34.71	53.80	0.5090	19.77	29.02	22.79	32.10	0.4791	13.06	18.24	15.04	19.98	0.5540	22.44	33.43	26.31	36.58	0.5024
AirPhyNet	30.43	48.48	36.05	55.56	0.5108	21.94	31.30	25.94	35.10	0.4504	14.43	20.12	16.67	22.01	0.4210	25.32	36.79	29.41	40.06	0.4371
PCDCNet	27.15	44.85	32.10	51.20	0.5780	17.85	27.10	19.90	29.75	0.6150	11.50	16.20	12.80	17.40	0.6350	20.10	30.65	22.50	33.10	0.6020
DeepAir	28.97	45.68	34.60	52.49	0.5389	20.07	28.82	23.84	32.45	0.5299	12.79	17.69	14.69	19.36	0.5994	21.49	31.83	26.09	35.55	0.5806
GAGNN	34.97	51.42	36.74	54.50	0.3873	29.65	36.45	29.72	37.39	0.2044	18.74	23.26	18.72	23.31	0.1144	32.39	40.20	32.84	41.03	0.2940
MGSFformer	31.94	48.12	38.31	54.68	0.4938	24.05	33.08	29.91	37.50	0.4585	15.78	21.10	18.77	23.24	0.4079	27.92	38.92	33.54	42.15	0.4388
$PM_{2.5}$GNN	26.05	44.13	31.04	50.13	0.6024	16.95	25.40	18.32	27.07	0.6703	10.75	15.23	11.77	16.05	0.6699	19.84	29.29	22.38	31.93	0.6048
CauAir	29.98	45.64	38.76	53.88	0.5874	17.23	26.37	19.49	29.57	0.6555	14.95	20.29	20.16	24.45	0.5814	23.98	35.29	32.81	43.45	0.5870
Spatial Kriging Models (adapted)																				
IGNNK	32.18	49.52	38.45	56.21	0.4523	22.85	32.14	27.62	36.48	0.4812	14.82	20.15	17.35	22.18	0.4685	25.72	37.15	30.48	40.82	0.4625
INCREASE	31.25	48.38	37.12	55.08	0.4782	21.52	31.05	25.88	35.12	0.5045	14.15	19.52	16.42	21.35	0.4912	24.85	36.28	29.52	39.95	0.4852
SATCN	30.42	47.25	35.85	53.72	0.4958	20.88	30.42	24.52	34.18	0.5218	13.58	18.85	15.72	20.52	0.5125	23.95	35.42	28.15	38.82	0.5042
Ours																				
ExtrapAir	**23.52**	**38.18**	**30.56**	**48.80**	**0.7338**	**15.70**	**23.98**	**17.70**	**26.56**	**0.7258**	**10.30**	**14.96**	**10.89**	**15.72**	**0.7206**	**16.62**	**27.87**	**19.32**	**31.56**	**0.6879**

We directly examine whether ExtrapAir depends excessively on highly accurate future-weather inputs. Two complementary settings are considered: (1) removing the future-weather WACorrBlock entirely, and (2) replacing fixed Gaussian weather perturbations with horizon-dependent errors. These experiments show that future weather provides complementary gains, while ExtrapAir's main advantage does not depend solely on near-ground-truth future-weather inputs.

The future-weather-free variant still outperforms all compared baselines that use future-weather covariates, indicating that the future-weather branch improves performance but is not the sole source of the model's gains. We also test horizon-dependent degradation. For normalized future-weather covariates at forecasting horizon $\ell \in \{1, \dots, L\}$, we use

$$\widetilde{\mathbf{Z}}^f_{t+\ell} = \mathbf{Z}^f_{t+\ell} + \boldsymbol{\epsilon}_\ell, \qquad \boldsymbol{\epsilon}_\ell \sim \mathcal{N}(0, \sigma_\ell^2), \qquad \sigma_\ell = 0.10 + 0.10\frac{\ell - 1}{L - 1}, \tag{37}$$

so long-range weather inputs are less accurate than short-range inputs.

Under horizon-dependent errors, ExtrapAir's MAE increases by 1.57% on LargeAQ and 0.97% on YRD. It remains the best-performing method on both datasets, reducing MAE relative to the strongest baseline by 14.11% on LargeAQ and 4.85% on YRD. Together, these results show that ExtrapAir remains effective both when future-weather information is unavailable and when weather errors increase with the forecasting horizon.

J Effects of Weather Variables on $PM_{2.5}$ Concentration

Figure 16 illustrates the nonlinear associations between $PM_{2.5}$ concentration and three weather factors: temperature, relative humidity, and wind speed. The colored curves with circular markers denote the observed response patterns, while the gray dashed curves represent smoothed fitted trends. $PM_{2.5}$ exhibits a clear inverted-U relationship with temperature, reaching its highest level near low-temperature conditions and declining as temperature increases. A similar nonlinear pattern is observed for humidity, where $PM_{2.5}$ increases with humidity up to a moderate range and decreases under high-humidity conditions. In contrast, wind speed shows a generally negative association with $PM_{2.5}$, indicating that stronger wind conditions facilitate pollutant dispersion and reduce particulate concentration. These results highlight the heterogeneous and nonlinear impacts of weather variables on air pollution dynamics.

Table 10: Performance on the additional US-NAT and EU-193 datasets under mask ratio r=0.5. The **best**, second, and third results are highlighted.

Method	Overall		Masked		Overall		Masked	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
	US-NAT				**EU-193**			
Time-Series Models								
CATS	4.18	7.77	4.42	7.90	8.40	12.76	9.63	12.99
CrossGNN	3.94	7.84	4.31	8.17	7.64	12.46	7.94	12.74
CycleNet	3.91	7.53	4.30	7.89	6.94	11.62	7.22	12.11
DLinear	3.92	7.51	4.28	7.89	6.93	12.17	7.09	12.81
DSformer	4.35	7.77	4.85	8.01	9.47	13.58	11.62	14.37
SOFTS	4.33	7.66	5.00	8.07	9.48	13.48	12.01	14.67
SparseTSF	4.21	7.75	4.51	7.91	8.80	13.21	9.99	13.22
TimeMixer	4.12	7.63	4.53	7.91	8.68	13.00	10.02	13.23
TimeXer	4.17	8.31	4.11	7.99	7.07	12.11	6.99	12.05
Spatiotemporal Models								
AGCRN	3.91	7.61	4.39	7.99	10.39	14.00	10.30	13.96
ASTGCN	3.77	7.54	4.10	7.89	7.03	11.56	7.40	12.18
BigST	3.97	7.52	4.44	7.92	7.01	11.61	7.21	12.24
D^{2}STGNN	4.06	8.23	4.20	8.16	7.06	12.35	6.97	12.29
GWNet	4.26	7.59	5.12	8.07	6.93	11.36	7.29	11.80
RPMixer	3.75	7.35	4.06	7.61	10.51	17.85	9.95	16.10
STAEformer	3.82	7.53	4.16	7.80	7.25	12.52	7.15	12.46
STGCN	3.78	7.50	4.07	7.72	6.91	11.47	6.87	11.48
STGODE	3.91	7.64	4.12	7.87	7.20	12.08	7.33	12.17
STID	3.83	7.46	4.21	7.84	7.14	11.40	7.53	11.98
STNorm	3.81	7.54	4.13	7.78	7.16	12.13	8.25	13.49
STTN	4.24	7.70	4.93	8.12	7.41	11.83	7.73	12.31
Air Quality Models								
AirFormer	4.03	7.60	4.50	7.85	6.69	11.31	7.06	12.24
AirPhyNet	3.96	7.71	4.29	7.89	6.83	11.53	7.26	12.10
DeepAir	3.89	7.66	4.14	7.89	7.10	12.01	7.15	12.25
GAGNN	3.94	7.57	4.28	7.79	7.05	12.46	6.95	12.40
MGSFformer	4.03	7.69	4.24	7.92	8.83	13.04	10.53	13.57
PCDCNet	3.88	7.51	4.29	7.77	7.06	12.19	7.14	12.44
PM$_{2.5}$GNN	3.73	7.48	4.02	7.57	7.35	12.19	7.86	12.89
CauAir	3.64	7.25	4.06	7.61	6.90	12.29	8.16	14.18
Spatial Kriging Models (adapted)								
IGNNK	3.90	7.67	4.19	7.89	7.05	11.89	7.22	12.00
INCREASE	3.83	7.52	4.19	7.89	7.38	11.82	7.63	12.36
SATCN	4.54	8.12	4.67	8.27	7.70	12.52	7.91	12.94
Ours								
ExtrapAir	**3.43**	**7.10**	**3.78**	**7.39**	**6.40**	**10.53**	**6.75**	**10.86**

K Technical Novelty Discussion

ExtrapAir is not a simple combination of weather covariates, linear attention, and spatial priors. Its novelty lies in redesigning these techniques for target-absent air-quality inference, where unmonitored stations have no historical AQI observations or direct supervision.

Table 11: Clustered geographic holdout on LargeAQ. Overall metrics and masked metrics denote overall and masked evaluation results, respectively.

Model	O-MAE	O-RMSE	M-MAE	M-RMSE
$PM_{2.5}$GNN	10.44	15.53	17.53	28.82
DLinear	11.01	15.98	19.90	25.58
AirFormer	10.37	15.30	21.98	26.78
GWNet	10.41	15.17	16.68	24.64
CauAir	9.31	13.95	16.69	25.77
ExtrapAir	**8.92**	**13.45**	**15.88**	**23.72**

Table 12: Performance when future-weather information is removed. Baselines use future weather with fixed Gaussian perturbation $\sigma = 0.10$.

Model / Setting	Weather setting	LargeAQ MAE	LargeAQ RMSE	YRD MAE	YRD RMSE
$PM_{2.5}$GNN	Future weather + Gaussian noise ($\sigma = 0.10$)	12.38	16.78	10.75	15.23
AirFormer	Future weather + Gaussian noise ($\sigma = 0.10$)	13.32	18.21	13.06	18.24
GWNet	Future weather + Gaussian noise ($\sigma = 0.10$)	11.87	16.60	12.45	17.49
CauAir	Future weather + Gaussian noise ($\sigma = 0.10$)	12.78	17.85	14.95	20.29
ExtrapAir	Future weather + Gaussian noise ($\sigma = 0.10$)	**10.19**	**14.67**	**10.30**	**14.96**
ExtrapAir w/o future weather	Future-weather branch removed	10.42	14.93	10.43	15.10
ExtrapAir w/o past weather	Past-weather branch removed	10.51	15.04	10.50	15.18

Table 13: Performance under fixed and horizon-growing weather noise.

Model	LargeAQ				YRD			
	Fixed MAE	Fixed RMSE	Growing MAE	Growing RMSE	Fixed MAE	Fixed RMSE	Growing MAE	Growing RMSE
$PM_{2.5}$GNN	12.38	16.78	12.56	16.97	10.75	15.23	10.93	15.44
AirFormer	13.32	18.21	13.51	18.42	13.06	18.24	13.24	18.45
GWNet	11.87	16.60	12.05	16.79	12.45	17.49	12.63	17.69
CauAir	12.78	17.85	12.96	18.04	14.95	20.29	15.14	20.51
ExtrapAir	**10.19**	**14.67**	**10.35**	**14.86**	**10.30**	**14.96**	**10.40**	**15.07**

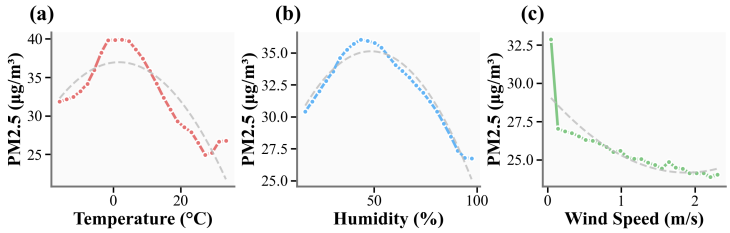


Figure 16: Nonlinear relationships between primary weather variables and $PM_{2.5}$. (a) Temperature exhibits an inverted-U relationship with peak $PM_{2.5}$ around 0°C. (b) Humidity shows a similar nonlinear pattern with maximum pollution at 50–60% relative humidity. (c) Wind speed demonstrates a strong negative correlation, with $PM_{2.5}$ decreasing sharply as wind speed increases.

First, weather covariates are treated as transferable local evidence rather than auxiliary inputs. ExtrapAir preserves variable-specific meteorological patterns and learns weather–AQI correlation gates, enabling weather information to compensate for missing AQI histories and transfer dependencies from monitored to unmonitored stations.

Second, ActAttn is designed for non-competitive spatial transfer rather than only for efficiency. Unlike softmax attention, which imposes row-wise competition and may suppress weak but useful signals, ActAttn allows unmonitored stations to aggregate complementary evidence from multiple monitored stations. Operating on fused weather–AQI representations, it further captures cross-station dependencies between meteorological conditions and AQI dynamics.

Third, geo-semantic priors are injected into query/key representations to recover station-pair discriminability when AQI-history-based similarity is unavailable. This provides relational guidance for invisible stations while preserving the linear computation of ActAttn.

Together, these designs form a task-driven mechanism: weather supplies target-side evidence, ActAttn performs non-competitive dependency transfer, and geo-semantic biasing guides reliable station-pair aggregation. Thus, the gains arise from addressing the coupled failures caused by missing target histories, rather than from stacking existing components.

L Limitations and Broader Impact

Geographic scope. The main benchmark covers five Chinese datasets spanning nationwide, regional, and city-cluster scales, Table 10 further evaluates US-NAT and EU-193, and Appendix H adds a stricter clustered geographic holdout. These additions strengthen geographic generalization, but broader deployment still requires validation on more continents, station-density regimes, and operational data ecosystems.

Weather forecast uncertainty. Appendix I evaluates both removing future-weather inputs and adding horizon-dependent weather errors. These tests show that ExtrapAir is not driven solely by near-perfect future weather. Nevertheless, the perturbations are still synthetic; future work should evaluate operational forecast products with realistic bias, resolution mismatch, and horizon-dependent uncertainty.

LLM embedding reproducibility. Appendix C specifies the leakage boundary and reports repeated-generation, cross-backbone, city-name-free, shuffled-semantic, and no-semantic controls. The remaining limitation is the lack of a fully hand-crafted metadata baseline, because the current public datasets do not provide sufficiently standardized station-level metadata. Recent harmonized metadata resources [56] could provide a reproducible and auditable complement, or partial replacement, for LLM-derived semantic profiles when such fields are available.

Extreme mask ratios. At $r{=}0.9$, the cross-node gating mechanism computes a global mean dominated by mask tokens (90% of stations). While ExtrapAir still outperforms all baselines at this extreme, exploring visible-station-only aggregation for the gating branch could further improve robustness.

Broader impact. Accurate air quality inference at unmonitored locations can directly benefit public health protection in underserved and under-monitored communities, where fixed monitoring infrastructure is sparse. At the same time, the most socially consequential scenarios often involve extreme pollution events, which remain harder to predict and can amplify the harm of overconfident model outputs. We therefore recommend that unmonitored-location predictions be used as supplementary guidance rather than definitive regulatory assessments, and that future deployments report calibrated prediction intervals and uncertainty estimates alongside point forecasts, especially when supporting health alerts or environmental decision-making.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the spatial-inference setting, the ExtrapAir components, and the empirical scope; the claims are supported by Sections 4.3 and the experimental results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and broader impacts are discussed in Appendix L, including geographic scope, weather forecast uncertainty, LLM embedding reproducibility, and extreme mask ratios.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 3.5 gives the theoretical summary, and the appendix provides derivations for prior-guided station-pair discriminability and the information benefit of meteorological covariates.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections on methodology and experiments specify the model architecture, masking protocol, datasets, baselines, metrics, hyperparameters, and random seeds; Algorithm 1 summarizes the forward computation.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The abstract provides an anonymized code link, and the manuscript describes the datasets, masking protocol, implementation details, and evaluation settings needed to reproduce the main experiments.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup specifies the mask ratios, input/output windows, optimizer, learning rate, hidden dimension, attention heads, semantic embedding generation, and 10-run evaluation protocol.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Results are averaged over 10 runs with seeds 42–51, each using a different visible/invisible station partition (Section 4.1); the source of variability is the station mask partition. Table 4 reports mean±std over repeated LLM profile generation.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The implementation details report the GPU type, and the efficiency study reports training speed, memory, parameter count, and accuracy trade-offs on LargeAQ.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses public or previously released air-quality datasets and does not involve deceptive data collection, human intervention, or unsafe model release.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix L discusses positive public-health use cases and the risk of over-reliance on predictions at unmonitored locations without ground-truth validation.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk generative models, scraped media datasets, or systems with an obvious dual-use misuse pathway requiring special safeguards.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets, baselines, and Qwen3 are cited in the manuscript; dataset statistics are summarized in Appendix G.1.1. Any released package should preserve the original license and terms of use.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The current paper introduces a method but does not introduce a new dataset or separately released asset in the manuscript.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research does not involve crowdsourcing, newly recruited participants, or new human-subject experiments.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper uses existing air-quality and weather datasets rather than newly collected human-subject data, so IRB approval is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM-generated station profiles are a methodological component: Qwen3 is used to derive semantic station embeddings, as described in the methodology, implementation details, and Appendix B.