KG-STMI: A Knowledge-Driven MultiModal Framework with Triple-Stage Knowledge Injection for Air Quality Forecasting in Environmental Sensing Networks

Jiachen Yuan[0009−0000−8763−0440], Ning Xin[0009−0003−7718−6552], Wenzheng Huang[0009−0004−0969−7331], Qiujin Wang[0009−0001−4682−0392], Lingxiao Zhao(✉)[0000−0002−1952−3026], and Md Maruf Hasan(✉)[0000−0001−9809−9376]

School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China
Jiachen.Yuan23@student.xjtlu.edu.cn, Ning.Xin21@student.xjtlu.edu.cn, Wenzheng.Huang23@student.xjtlu.edu.cn, Qiujin.Wang23@student.xjtlu.edu.cn, Lingxiao.Zhao@xjtlu.edu.cn, MdMaruf.Hasan@xjtlu.edu.cn

Abstract. Air quality forecasting is challenged by sudden pollutant fluctuations driven by external events such as public holidays and industrial activities. Current data-driven models often struggle to capture these dynamics due to technical limitations. We propose KG-STMI, a novel Knowledge-Driven MultiModal Framework with Triple-Stage Knowledge Injection for air quality forecasting in dense environmental sensing networks. Our architecture integrates a Prior Knowledge Reasoner (PKR) to extract event-aware features from Knowledge Graphs, a Knowledge-Driven Mamba architecture for long-range sequential modeling, and a Triple-Stage Knowledge Injection mechanism for multivariate feature fusion. Evaluated on a high-resolution air quality dataset (2020-2024), KG-STMI achieves state-of-the-art (SOTA) performance with a physical MAE of 6.26, outperforming leading baselines like SimVP (6.72 of MAE) and KST-GCN (7.04 of MAE). Ablation studies confirm the necessity of each processing step. Removing the Triple-Stage Injection causes the largest degradation, increasing MAE by 10.7%; replacing relational reasoning with simple embeddings raises MAE by 6.7%; and substituting the Mamba backbone with a Transformer increases MAE by 8.8%. These results demonstrate the value of integrating structured knowledge reasoning into Spatio-Temporal sensing networks.

Keywords: Knowledge-driven Learning · Multimodal Knowledge Graphs · Knowledge Representation and Reasoning · Temporal Knowledge Modeling · Air Quality Forecasting · Spatio-Temporal Prediction · Knowledge-based Sensing Systems

1 Introduction

Air quality monitoring is essential for public health management and urban planning [1]. Accurate forecasting requires not only modeling historical numerical patterns but also integrating structured knowledge about external events that can significantly disrupt pollutant dynamics. Large-scale high-resolution environmental datasets, provided by increasingly dense monitoring networks, pose significant challenges to current predictive models [2].

Existing prediction methods can be generally classified into three categories: CNN-based models for capturing spatial correlations, RNN-based models for modeling temporal patterns, and Transformer architectures for extracting global features. While these approaches have demonstrated improved accuracy on standard benchmarks, they primarily rely on learning from historical numerical patterns and may struggle to capture sudden fluctuations triggered by external environmental events such as public holidays or industrial adjustments. Moreover, as input sequences grow longer, Transformer-based methods face more computational complexity, reducing their efficiency in large-scale sensing networks.

To overcome these limitations, advanced models such as SimVP [7], Earthformer [8], and ClimaX [4] have been proposed, demonstrating strong performance on standard benchmarks. Recently, knowledge-enhanced frameworks including KST-GCN [21], UniST [22], and STKGformer [23] have explored the integration of prior information into Spatio-Temporal modeling. However, despite these advances, a critical gap remains. Existing approaches typically incorporate structured knowledge at the node level or as shallow input features, without systematically propagating semantic context across all stages of the Spatio-Temporal model. These models struggle to capture sudden fluctuations [3].

In this paper, we propose Knowledge Graph enhanced Spatio-Temporal Mamba with Triple-Stage Injection(KG-STMI), a knowledge-driven framework designed to bridge the semantic gap between numerical patterns and the real world. Our model incorporates three key components to address the limitations. First, to capture the structured relationships between environmental events and pollutant dynamics, we develop the Prior Knowledge Reasoner (PKR), which employs a Relational Graph Convolutional Network (RGCN) to perform relational reasoning over event-centric knowledge graphs and transform symbolic event information into continuous knowledge vectors. Second, since Transformer-based models suffer from more complexity that limits their scalability in long-range sequential modeling, we adopt Mamba as the backbone and integrate multimodal knowledge context through cross-attention fusion, balancing efficiency and contextual awareness. Third, since injecting knowledge only at a single stage risks losing semantic guidance in intermediate representations, we design the Triple-Stage Knowledge Injection Mechanism to propagate event context across the spatial encoder, bottleneck, and decoder levels, ensuring that structured knowledge influences feature learning throughout the entire prediction process. We evaluate KG-STMI on a high-resolution dataset spanning five years (2020–2024). Experimental results demonstrate that KG-STMI achieves superior performance over 15 competitive baselines, reaching a physical MAE

of 6.26 and establishing a SOTA performance for knowledge-driven sensing network analytics.

The main contributions of this work are as follows:

- We present KG-STMI, a knowledge-driven framework for air quality forecasting that integrates a purpose-built environmental event knowledge graph with a Mamba-based spatio-temporal backbone, enabling the model to capture both historical numerical patterns and event-driven pollutant disruptions such as those caused by public holidays and industrial activities.
- We propose the Prior Knowledge Reasoner (PKR), built upon a domain-specific knowledge graph for air quality forecasting. The key contribution lies in the knowledge graph construction strategy, we define a causal schema linking weather and social events to directional pollutant changes, and only establish a connection when an event and a pollutant trend appear together more than twice within the five-year observation window. This filtering step removes unstable associations and retains only those reflecting consistent physical patterns, and the structured knowledge is then transformed into continuous vectors through relational graph convolution.
- We introduce the Triple-Stage Knowledge Injection Mechanism, which addresses a fundamental limitation of existing knowledge-enhanced models that typically apply external knowledge at a single point in the network. Rather than treating knowledge as a one-time input, our design injects the Knowledge Context Vector at three distinct stages of the prediction process. At the spatial encoder, knowledge modulates early feature maps through channel-wise gating, suppressing irrelevant spatial patterns and amplifying event-relevant regions. At the bottleneck, knowledge conditions the temporal modeling process, ensuring that sequential dependencies are learned under the influence of current event context. At the decoder, knowledge applies a pollutant-wise correction to the final output, directly adjusting predicted concentration levels based on the detected event. This stage-specific design ensures that event knowledge shapes the model's behavior at every level, from spatial feature extraction to final prediction.
- We conduct extensive experiments on a high-resolution air quality dataset (2020–2024) to validate the effectiveness of KG-STMI. Our model achieves a physical MAE of 6.26, outperforming competitive baselines including SimVP (6.72 of MAE, 6.8% improvement) and KST-GCN (7.04 of MAE, 11.1% improvement) across all key metrics. Ablation studies further confirm the contribution of each component. Removing the Triple-Stage Injection causes the largest degradation, increasing MAE by 10.7%, replacing relational reasoning with simple embeddings raises MAE by 6.7%, and substituting the Mamba backbone with a Transformer increases MAE by 8.8%. Overall, the full framework improves Normalized MSE from 0.37 to 0.30 compared to the pure baseline (7.17 of MAE), representing an 18.9% reduction.

2 Related Work

2.1 Spatio-Temporal Prediction in Sensing Networks

Spatio-temporal prediction in sensing networks has evolved rapidly over the past decade. Early approaches relied on CNN-based architectures to capture local spatial correlations, including SimpleCNN [12], FlattenCNN [9], SimVP [7], and 3D U-Net [17]. Recurrent architectures such as SimpleLSTM [11], ConvLSTM [10], ConvGRU [13], and EncDecRNN [14] were then introduced to better model temporal dynamics. Transformer-based models such as Earthformer [8], ClimaX [4], and PredFormer [18] later extended this to longer-range Mamba [5] have been explored for sequential modeling. Despite these advances, all of the above models learn purely from historical data and lack awareness of the real world events [3], which often leads to substantial prediction errors when external factors such as public holidays or industrial disruptions occur.

2.2 Knowledge Integration and Structural Reasoning

Incorporating prior knowledge into predictive models has been shown to improve both accuracy and interpretability [21, 2]. Most existing methods inject knowledge by fusing additional input features with the raw data [2, 19], as modulation signals applied to hidden representations, or as correction terms added to the final output [24]. KST-GCN [21] combines knowledge graphs with graph convolution for traffic forecasting, UniST [22] uses text prompts to guide urban Spatio-Temporal modeling, and STKGformer [23] enhances spatial graph modeling with knowledge graph information. However, these methods tend to apply knowledge at a single stage or at the node level, which limits how deeply the semantic information can influence the model's behavior. KG-STMI takes a different approach by injecting knowledge into the spatial encoder, the bottleneck, and the decoder via relational reasoning over event-centric knowledge graphs using RGCN [16], allowing environmental knowledge to inform the model throughout the entire prediction process rather than serving as a shallow feature.

2.3 Efficient Sequential Modeling

With the increase in data, computational efficiency becomes a critical consideration. State Space Models (SSMs) such as Mamba [5] offer linear-time alternatives to Transformer architectures, capturing long-range dependencies with substantially lower computational overhead. Multimodal alignment techniques such as CLIP [6] further enable grounding of Spatio-Temporal patterns with textual environmental context. KG-STMI combines Mamba with multimodal knowledge reasoning, transforming the forecasting task from pure numerical modeling into a context-aware refinement process.

Table 1. Comparison with representative Spatio-Temporal forecasting approaches.

Method	External Knowledge	Graph Reasoning	Complexity
SimVP [7]	×	×	$O(n)$
STGCN [20]	×	✓(Spatial)	$O(n)$
ClimaX [4]	×	×	$O(n^2)$
UniST [22]	✓(Prompt)	×	$O(n)$
STKGformer [23]	✓(Multi-graph)	✓(Semantic)	$O(n^2)$
KG-STMI (Ours)	✓(KG)	✓(Event-centric)	$O(n)$

3 Methodology

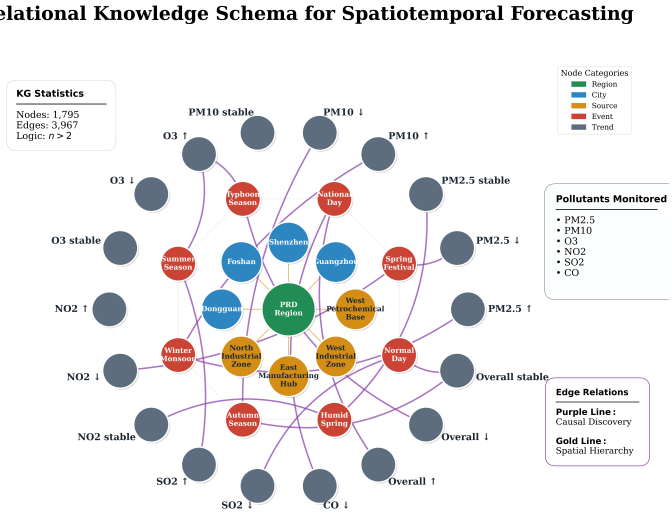


Fig. 1. Relational Knowledge Schema for Spatio-Temporal Forecasting. The graph organizes 1,795 nodes and 3,967 edges into a hierarchical structure. Purple lines denote causal discovery between events and pollutant trends based on the $n > 2$ logic, while gold lines represent the spatial hierarchy from the regional level to specific urban and industrial entities.

3.1 Data Preprocessing and Knowledge Graph Construction

We evaluate our framework on a high-resolution air quality dataset provided by the China Meteorological Administration, covering the Pearl River Delta region from 2020 to 2024 (1,827 days in total). The dataset records six pollutant channels ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) at a spatial resolution of $0.01° \times 0.01°$.

The dataset is divided into training (2020–2022), validation (2023), and test (2024) sets, corresponding to a 60% / 20% / 20% split. To reduce computational cost while preserving spatial structure, the original 300×400 grid is first downsampled to 192×256 via a 2×2 average pooling followed by bilinear interpolation. Since approximately 33.9% of the grid points cover ocean areas, we construct a global binary land mask $M \in \{0,1\}^{H \times W}$ by identifying pixels that contain valid observations across any time step. Ocean pixels are set to zero and excluded from both training and evaluation to prevent the model from learning spurious patterns from invalid sea regions. We then apply a standard three-step preprocessing pipeline. First, missing values are filled using temporal linear interpolation for short gaps and spatial interpolation for larger missing areas. Second, since the six pollutant channels have different physical units, we apply z-score normalization to each channel independently. Third, we use a sliding window with stride 1 to generate input-output pairs (X_{in}, Y_{gt}), where $X_{in} \in \mathbb{R}^{T_{in} \times C \times H \times W}$ is the historical input with $T_{in} = 45$ days, and $Y_{gt} \in \mathbb{R}^{T_{out} \times C \times H \times W}$ is the prediction target with $T_{out} = 30$ days, $C = 6$, $H = 192$, $W = 256$.

To operationalize structured guidance, we construct a domain-specific knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ as illustrated in Fig. 1. The resulting graph contains 1,795 nodes and 3,967 edges, synthesizing a relational metadata layer that exists in parallel with the numerical data stream. The entity set $\mathcal{V}$ is categorized into geographic regions and cities, industrial sources representing manufacturing hubs, social and meteorological events such as Spring Festival or Typhoon Season, trend entities capturing directional changes like $PM_{2.5}$ rise, and 1,752 temporal anchors serving as the interface to the time series. The relational set $\mathcal{R}$ encompasses six types, including shows_trend, has_event, and located_in. We distinguish between spatial hierarchy, which encodes geographic containment from the regional level down to specific sources, and causal discovery for event-trend associations. To ensure the robustness of the knowledge graph, we utilize a Spatio-Temporal co-occurrence filtering method where a correlates_with relation is established between an event and a trend entity only if their observed co-occurrence frequency n exceeds a threshold of 2 within the five year window. This logic effectively filters out stochastic fluctuations and ensures that the reasoning module operates on reliable physical associations.

3.2 Overall Workflow

The overall workflow of KG-STMI is illustrated in Fig. 2, and consists of two parallel streams that operate simultaneously. As shown in Fig. 2, the framework takes two parallel inputs: a text prompt P describing the current environmental context, and historical pollutant frames from the PRD region. The text prompt is fed into the Knowledge & Retrieval module, where it is encoded by the CLIP Encoder and used to retrieve top-k prototypes from the Memory Bank via vector search. At the same time, P is processed by the PKR, which performs relational reasoning over the Knowledge Graph to transform relational event information into a continuous Knowledge Context Vector (KCV). The historical frames are simultaneously passed into the Spatio-Temporal Backbone as

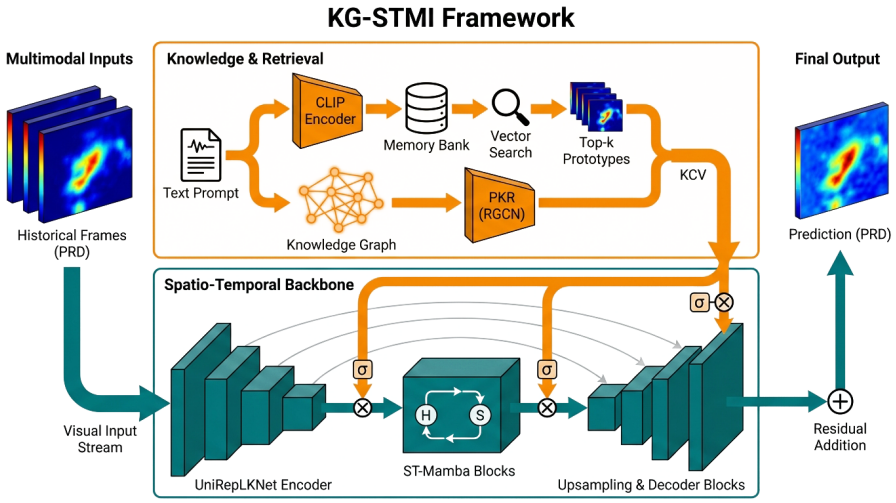


Fig. 2. Overall architecture of the KG-STMI Framework, combining a Knowledge & Retrieval module with a Spatio-Temporal Backbone via Triple-Stage Knowledge Injection (orange paths).

a visual input stream. This vector captures the semantic implications of environmental events and serves as a guidance signal for the rest of the network. At the same time, the Spatio-Temporal Backbone processes the historical pollutant frames through a hierarchical pipeline. The UniRepLKNet Encoder first extracts multi-scale spatial features, which are then passed to the ST-Mamba Blocks for temporal modeling. The resulting representations are finally reconstructed into future frames by the Upsampling and Decoder Blocks. The two streams are connected through the Triple-Stage Knowledge Injection mechanism, which uses gated connections to inject the KCV into the encoder, the bottleneck, and the decoder. This allows the environmental event knowledge to influence feature learning throughout the entire network rather than at a single stage. The final output is obtained by adding the predicted residuals to the last observed frame, producing the predicted sequence $\mathcal{Y} \in \mathbb{R}^{T' \times C \times H \times W}$.

3.3 Enhancements and Innovations for Knowledge-Driven Air Quality Forecasting

Prior Knowledge Reasoner: Reasoning over Knowledge Graphs To bridge the gap between discrete environmental events and numerical prediction, we propose the PKR, which applies relational graph convolution [16] to event-centric subgraphs extracted from a domain-specific knowledge graph designed for air quality forecasting. While the underlying graph convolution operation follows the RGCN formulation, the knowledge graph schema, the causal event-trend linking strategy, and the subgraph extraction procedure are designed specifi-

cally for this task. This module captures event-driven physical dependencies by performing relational message passing on event-centric subgraphs.

Let $\mathcal{G}_{sub} = (\mathcal{V}_{sub}, \mathcal{E}_{sub})$ be the extracted subgraph centered on the current event. We employ a Relational Graph Convolutional Network (RGCN) to update node embeddings. The feature update rule for node v_i at layer $l+1$ is defined as:

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{|\mathcal{N}_i^r|}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right) \tag{1}$$

where $\mathcal{N}_i^r$ denotes the neighbors of node i under relation r, and $W_r^{(l)}$ are learnable weights. After L layers, we synthesize the global context using a pooling operation:

$$\mathbf{KCV} = \frac{1}{|\mathcal{V}_{sub}|}\sum_{v\in\mathcal{V}_{sub}} h_v^{(L)} \tag{2}$$

This process is detailed in Algorithm 1

Algorithm 1 PKR: relational Knowledge Reasoning

Require: Global Knowledge Graph $\mathcal{G}$, Query Event e_{query}.
Require: RGCN Weights $\{W_r^{(l)}\}_{r\in\mathcal{R}}$.
Ensure: Knowledge Context Vector **KCV**.
// **Step 1: Subgraph Extraction**
$v_{center} \leftarrow \text{LookupNodeID}(\mathcal{G}, e_{query})$
$\mathcal{G}_{sub} \leftarrow \text{ExtractKHopSubgraph}(\mathcal{G}, v_{center}, k=1)$
// **Step 2: Relational Propagation (Eq. 1)**
for $l = 1$ to L **do**
  for each node $v_i \in \mathcal{V}_{sub}$ **do**
    $h_{msg} \leftarrow \sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{|\mathcal{N}_i^r|}W_r^{(l)}h_j^{(l-1)}$
    $h_i^{(l)} \leftarrow \sigma(h_{msg} + W_0^{(l)}h_i^{(l-1)})$
  end for
end for
// **Step 3: Aggregation (Eq. 2)**
KCV $\leftarrow \text{MeanPool}(\{h_v^{(L)} \mid v \in \mathcal{V}_{sub}\})$
return KCV

Knowledge-Driven Mamba with Multimodal Fusion To ensure computational efficiency while aligning multimodal context, we integrate a retrieval-based fusion mechanism [25] with the Mamba backbone [5]. The text embedding is obtained via CLIP [6], and the gated cross-attention design follows standard attention formulation [26], adapted here to fuse retrieved historical prototypes with visual spatio-temporal features for air quality forecasting.

Given the input visual features H_{enc} and text embedding E_{text}, we first retrieve the top-k historical prototypes $\mathcal{P}_{top-k}$ using semantic similarity. The

context C_{ctx} is fused into the Mamba input via Text-Gated Cross-Attention:

$$\text{Attn} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad \alpha = \sigma(W_{gate}C_{ctx}) \tag{3}$$

The fused feature $H_{fused} = H_{enc} + \alpha \cdot \text{Attn}$ is then processed by the Selective State Space Model (SSM), defined discretely as:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}H_{fused}[t], \quad y_t = \mathbf{C}h_t \tag{4}$$

This process is summarized in Algorithm 2.

Algorithm 2 Knowledge-Driven Mamba Fusion Process

Require: Input Features H_{enc}, Text Prompt P, Memory Bank $\mathcal{M}$.
Ensure: Refined Latent Features H_{latent}.
1: $E_{text} \leftarrow \text{CLIP}(P)$
2: $\mathcal{P}_{top-k} \leftarrow \text{Retrieve}(\mathcal{M}, E_{text})$
3: $C_{ctx} \leftarrow \text{Concat}(E_{text}, \mathcal{P}_{top-k})$
4: $\alpha \leftarrow \sigma(W_{gate}C_{ctx})$ {Compute Gate}
5: $H_{fused} \leftarrow H_{enc} + \alpha \cdot \text{CrossAttn}(H_{enc}, C_{ctx})$
6: $H_{latent} \leftarrow \text{SSM_Scan}(\overline{\mathbf{A}}, \overline{\mathbf{B}}, \mathbf{C}, H_{fused})$ {Eq. 4}
7: **return** H_{latent}

Triple-Stage Knowledge Injection Mechanism To fully integrate relational knowledge into the numerical prediction process, we employ a Triple-Stage Knowledge Injection strategy. Unlike existing knowledge-enhanced models that apply external knowledge at a single stage [21, 22], this mechanism injects the KCV at three distinct levels of the backbone via feature-wise modulation [24], namely the spatial encoder, the Mamba bottleneck [5], and the decoder output, ensuring that structured knowledge guides feature learning throughout the entire prediction process.

At Stage 1 (Spatial Encoder), the KCV is projected through an MLP to generate a channel-wise gating map, which is applied to the encoder feature maps via multiplicative modulation. This allows environmental event knowledge to selectively amplify or suppress spatial features at early encoding stages.

At Stage 2 (Mamba Bottleneck), the KCV is fused with the text embedding before being passed into the ST-Mamba Blocks, ensuring that the temporal modeling process is conditioned on the current environmental context throughout sequence processing.

At Stage 3 (Decoder Output), an adaptive bias correction term $B_{bias} = \text{MLP(KCV)}$ is applied directly to the final prediction, providing a pollutant-wise output correction that adjusts the predicted concentration levels based on event context.

The injection at each stage is governed by a unified Gated Attention operator. For a feature map F_{in}^s at stage $s \in \{\text{spatial, bottleneck, decoder}\}$, the updated

feature is:

$$G_s = \sigma(\mathbf{W}_g^s \cdot \mathbf{KCV} + b_g^s), \quad F_{out}^s = F_{in}^s \odot (1 + G_s) \tag{5}$$

This process is summarized in Algorithm 3.

Algorithm 3 Triple-Stage Knowledge Injection

Require: Backbone Features $F_{enc}, F_{bot}, F_{dec}$, Knowledge Vector **KCV**.
Ensure: Knowledge-Enhanced Prediction $\hat{\mathcal{Y}}$.
// **Stage 1: Spatial Injection**
$G_{spatial} \leftarrow \sigma(\mathbf{W}_g^s \cdot \mathbf{KCV} + b^s)$
$F_{enc} \leftarrow F_{enc} \odot (1 + G_{spatial})$
// **Stage 2: Bottleneck Injection**
$G_{bot} \leftarrow \sigma(\mathbf{W}_g^b \cdot \mathbf{KCV} + b^b)$
$F_{bot} \leftarrow \text{Mamba}(F_{enc}) \odot (1 + G_{bot})$
// **Stage 3: Decoder Injection**
$B_{bias} \leftarrow \text{MLP}_{bias}(\mathbf{KCV})$
$\hat{\mathcal{Y}} \leftarrow \text{Decoder}(F_{bot}) + B_{bias}$ {Eq. 5}
return $\hat{\mathcal{Y}}$

4 Experiments Application and Results

4.1 Multivariate Spatio-Temporal Data Analysis

We evaluate our framework on a high-resolution air quality dataset provided by the China Meteorological Administration, covering the Pearl River Delta (PRD) region from 2020 to 2024 (1,827 days in total). Table 2 provides a statistical summary of the 1,827-day sequence. The data exhibits significant variability and high peak concentrations (e.g., PM_{10} reaching 352.46 $\mu g/m^3$), indicating the presence of severe pollution episodes that challenge traditional numerical extrapolation.

Table 2. Statistical summary and specifications of the PRD air quality dataset (2020-2024). Units: $\mu g/m^3$ for $PM_{2.5,10}$, O_3, NO_2, SO_2; mg/m^3 for CO.

Variable	Unit	25% Q.	Max	Mean	Median	Dataset Specifications
$PM_{2.5}$	$\mu g/m^3$	12.36	210.20	20.93	18.74	**Spatial Res.:** $0.01^\circ \times 0.01^\circ$
PM_{10}	$\mu g/m^3$	23.86	352.46	36.38	32.94	**Grid Size:** 300×400 points
O_3	$\mu g/m^3$	65.33	326.56	88.21	84.17	**Temporal Range:**
NO_2	$\mu g/m^3$	10.08	228.17	16.08	13.63	2020-01-01 to 2024-12-31
SO_2	$\mu g/m^3$	6.12	44.71	7.05	6.79	(1,827 days)
CO	mg/m^3	0.54	6.66	0.61	0.60	**Ocean Points:** ∼33.9%

Figure 3 visualizes the spatial distribution of pollutants across the PRD region. Significant hotspots for $PM_{2.5}$, PM_{10}, and NO_2 are concentrated in urban

centers (e.g., Guangzhou and Foshan), while O_3 shows a more diffuse regional pattern. Importantly, the ocean regions (South China Sea) are effectively masked to prevent learning spurious features, highlighting the necessity of our geography-aware preprocessing and the Triple-Stage Knowledge Injection for distinguishing industrial zones from clean backgrounds.

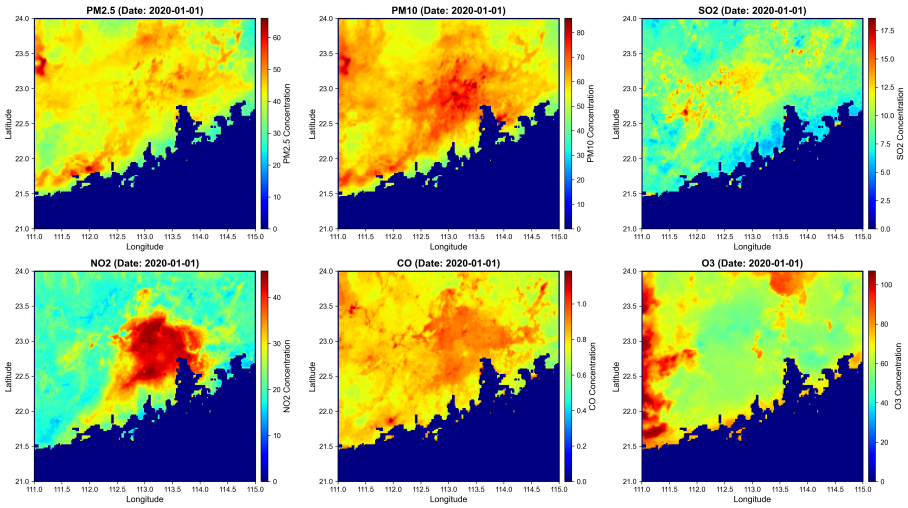


Fig. 3. Spatial distribution of six pollutants in the PRD region. High spatial heterogeneity is observed, with distinct urban hotspots and clear ocean-land boundaries handled by our masking strategy.

4.2 Experimental Results and Comparison

Fig. 4 and Fig. 5 further support the quantitative results above. As shown in Fig. 5, KG-STMI accurately reconstructs the high-concentration NO_2 hotspots in central urban areas, while most regions remain within a near-zero error range in the bias map. Fig. 4 shows that purely data-driven baselines such as Earthformer and ClimaX suffer from significant blurring of intensity peaks, and even knowledge-enhanced models like KST-GCN and UniST fail to restore fine-grained spatial boundaries. In contrast, KG-STMI consistently recovers spatial details with higher fidelity across all compared models.

To provide a comprehensive assessment, we evaluate the KG-STMI framework against a diverse set of 15 established models representing the technological evolution of spatio-temporal forecasting. Our comparison covers foundational recurrent networks such as SimpleLSTM, ConvLSTM, ConvGRU, and EncDecRNN, as well as spatial-centric convolutional architectures including SimVP, FlattenCNN, SimpleCNN, and 3D U-Net. We also benchmark against global-context Transformers like Earthformer, ClimaX, and PredFormer, along with the

Table 3. Quantitative comparison on the Air Quality dataset. Red indicates the best performance, and Blue indicates the second best. Our KG-STMI achieves superior performance across all metrics. Results are rounded to two decimal places.

Category	Method	Year	Physical Metrics						Normalized Metrics					
			MSE↓	MAE↓	RMSE↓	IOA↑	r↑	MAPE↓	MSE↓	MAE↓	RMSE↓	IOA↑	r↑	MAPE↓
CNNs	FlattenCNN	2021	236.92	7.89	10.47	0.53	0.35	22.98%	0.32	0.40	0.55	0.58	0.35	24.41%
	SimVP	2022	180.33	6.72	8.94	0.66	0.47	19.87%	0.34	0.42	0.56	0.66	0.37	22.98%
	SimpleCNN	2023	209.55	7.70	10.18	0.50	0.25	24.76%	0.53	0.53	0.70	0.39	0.09	29.02%
	3DUNet	2016	189.27	7.05	9.26	0.60	0.39	21.53%	0.36	0.43	0.57	0.43	0.24	24.57%
RNNs	SimpleLSTM	2019	204.27	7.35	9.69	0.57	0.34	22.98%	0.41	0.46	0.62	0.43	0.20	26.40%
	ConvLSTM	2023	185.94	7.08	9.30	0.57	0.35	23.08%	0.38	0.45	0.59	0.42	0.24	26.40%
	ConvGRU	2016	179.62	6.83	9.17	0.55	0.38	20.32%	0.38	0.43	0.59	0.40	0.28	23.14%
	EncDecRNN	2023	205.90	7.28	9.70	0.58	0.35	22.36%	0.41	0.46	0.62	0.45	0.22	25.73%
Trans.	Earthformer	2022	201.00	7.36	9.68	0.54	0.34	23.42%	0.43	0.48	0.63	0.38	0.19	26.77%
	ClimaX	2023	234.32	7.76	10.12	0.69	0.48	23.28%	0.44	0.47	0.63	0.53	0.25	26.91%
	PredFormer	2025	214.94	7.78	10.05	0.65	0.43	24.50%	0.49	0.50	0.66	0.47	0.16	28.43%
SSM/LLM	TimeLLM	2024	353.33	10.02	12.89	0.46	0.31	30.88%	0.95	0.77	0.97	0.37	-0.09	38.68%
KG-Baselines	KST-GCN	2022	204.42	7.04	9.55	0.67	0.45	20.93%	0.37	0.42	0.57	0.64	0.41	24.13%
	UniST	2024	253.82	7.63	10.49	0.66	0.43	21.46%	0.42	0.44	0.60	0.63	0.38	24.84%
	STKGformer	2025	247.37	7.51	10.32	0.67	0.44	20.92%	0.41	0.43	0.59	0.65	0.42	24.34%
Ours	**KG-STMI**	**2026**	161.26	6.26	8.56	0.69	0.51	18.44%	0.30	0.38	0.52	0.67	0.49	21.33%

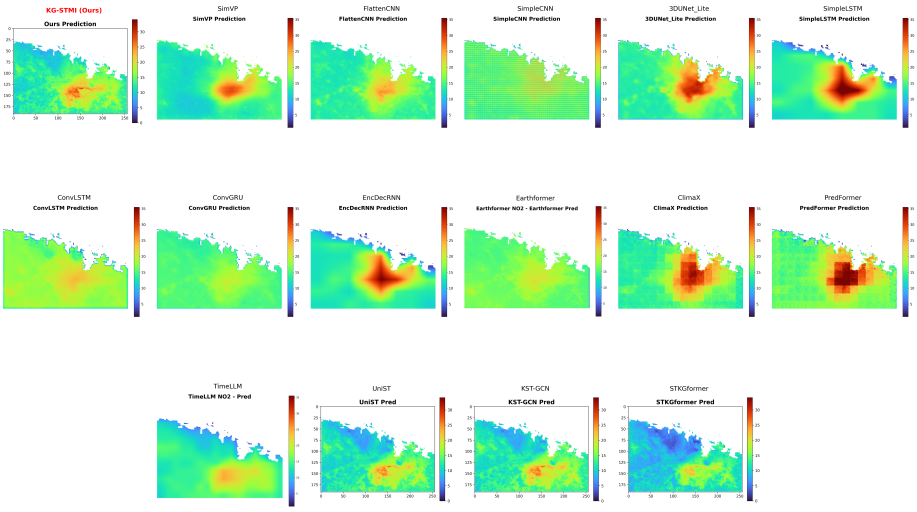


Fig. 4. Panoramic visual comparison across 16 models.

LLM-based Time-LLM. Crucially, to ensure a fair comparison, we include state-of-the-art knowledge-enhanced models: KST-GCN, UniST, and STKGformer.

As shown in Table 3, KG-STMI achieves state-of-the-art performance across all 12 metrics. In terms of physical metrics, KG-STMI attains a MAE of 6.26 and an MSE of 161.26, outperforming the best CNN-based baseline SimVP (6.72 of MAE, 180.33 of MSE) by 6.8% and 10.6% respectively. Compared to the best RNN-based baseline ConvGRU (6.83 of MAE, 179.62 of MSE), our model reduces MAE by 8.3%. Against Transformer-based models, KG-STMI surpasses Earthformer (7.36 of MAE) and ClimaX (7.76 of MAE) by significant mar-

gins. Notably, the LLM-based Time-LLM performs poorly on this task (10.02 of MAE), suggesting that general-purpose language models struggle with high-resolution spatiotemporal prediction without domain-specific knowledge. Among knowledge-enhanced baselines, KST-GCN (7.04 of MAE) and STKGformer (7.51 of MAE) both fall significantly behind KG-STMI, demonstrating that our triple-stage injection strategy provides substantially deeper knowledge integration than single-stage approaches.

In terms of normalized metrics, KG-STMI achieves a Normalized MSE of 0.30, compared to 0.34 for SimVP and 0.37 for KST-GCN, representing reductions of 11.8% and 18.9% respectively. Our model also achieves the best MAPE of 18.44%, outperforming all baselines including the second-best SimVP (22.98%) by 4.54 percentage points. The IOA and correlation coefficient r further confirm the superiority of KG-STMI, reaching 0.69 and 0.51 respectively.

The qualitative results in Fig. 5 further support these findings. KG-STMI accurately reconstructs the high-concentration NO_2 hotspots in central urban areas, while the prediction bias map (c) shows that most of the region remains within a near-zero error range, with only localized underestimation in the most extreme pollution zones.

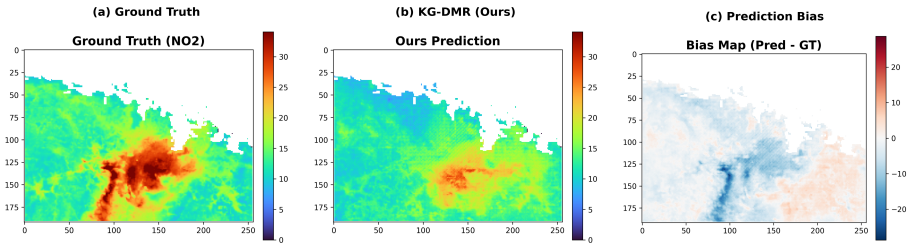


Fig. 5. Visual analysis of NO_2 prediction results. (a) Ground Truth distribution; (b) KG-STMI Prediction; (c) Prediction Bias map ($Pred - GT$). The high consistency between observation and prediction demonstrates the model's ability to maintain spatial precision under complex event influences.

4.3 Ablation Study

We conduct an ablation study to validate the effectiveness of each proposed component in KG-STMI. We evaluate six variants that represent a progressive technological hierarchy, specifically the Pure Baseline, the Single-Stage variant, the w/o Mamba version, the w/o KG model, the w/o Reasoning architecture, and the full KG-STMI. The Pure Baseline is defined as a simplified convolutional encoder-decoder architecture with linear temporal projection, lacking both advanced SSM-based sequence modeling and knowledge reasoning modules. Table 4 presents the results of all ablation variants across all metrics.

Table 4 presents the results of all ablation variants across all metrics. The Single-Stage variant shows the largest degradation among all variants, with Physical MSE rising to 201.79 and MAE to 6.93, an increase of 10.7% compared to the full model, indicating that multi-level knowledge injection is the most critical component in our framework. Removing Mamba (w/o Mamba) results in the second largest drop, with Physical MSE increasing to 197.64 and MAE to 6.81 (8.8% higher than the full model), confirming that Mamba's sequential modeling is essential for capturing temporal dynamics. Removing the knowledge graph entirely (w/o KG) raises the Physical MSE to 189.46 and MAE to 6.74 (7.7% degradation), while the Normalized MSE increases from 0.30 to 0.34 (13.3%). Replacing relational reasoning with simple embeddings (w/o Reasoning) yields the second-best variant with a Physical MAE of 6.68 and Normalized MSE of 0.34, yet still falls 6.7% short of the full model on MAE, demonstrating that structured relational reasoning over the knowledge graph is indispensable. Overall, the full KG-STMI reduces Physical MSE from 201.71 (Pure Baseline) to 161.26, a reduction of 20.1%, and Normalized MSE from 0.37 to 0.30, confirming the necessity of each proposed component.

Table 4. Ablation study results across twelve metrics. Red and Blue indicate the best and second-best results, respectively. The results confirm that each component is essential for bridging the semantic gap.

Variant	Physical Metrics						Normalized Metrics					
	MSE↓	MAE↓	RMSE↓	IOA↑	r↑	MAPE↓	MSE↓	MAE↓	RMSE↓	IOA↑	r↑	MAPE↓
Pure Baseline	201.71	7.17	9.44	0.58	0.36	22.24%	0.37	0.44	0.58	0.45	0.24	25.46%
Single-Stage	201.79	6.93	9.42	0.68	0.47	20.09%	0.35	0.41	0.56	0.64	0.42	23.25%
w/o Mamba	197.64	6.81	9.35	0.68	0.47	19.51%	0.35	0.40	0.56	0.64	0.43	22.61%
w/o KG	189.46	6.74	9.21	0.67	0.46	19.43%	0.34	0.40	0.55	0.63	0.42	22.56%
w/o Reasoning	186.73	6.68	9.13	0.68	0.47	19.37%	0.34	0.40	0.55	0.66	0.45	22.49%
KG-STMI (Full)	161.26	6.26	8.56	0.69	0.51	18.44%	0.30	0.38	0.52	0.67	0.49	21.33%

5 Conclusion and Future Work

In this paper, we proposed KG-STMI, a knowledge-driven multimodal framework for air quality forecasting that addresses the fundamental limitation of purely data-driven approaches, namely their inability to capture event-driven pollutant disruptions. The framework integrates three core components. The Prior Knowledge Reasoner (PKR) transforms environmental event information into continuous knowledge vectors. The Knowledge-Driven Mamba backbone provides efficient sequential modeling. The Triple-Stage Knowledge Injection mechanism then propagates event context across the encoder, bottleneck, and decoder levels of the network.

Extensive experiments on a high-resolution dataset (2020–2024) demonstrate that KG-STMI achieves state-of-the-art performance across all 12 metrics, attaining a physical MAE of 6.26 and MSE of 161.26, outperforming the best

CNN-based baseline SimVP (6.72 of MAE, 180.33 of MSE) by 6.8% and 10.6% respectively, and surpassing all knowledge-enhanced baselines including KST-GCN (7.04 of MAE) and STKGformer (7.51 of MAE). Normalized MSE is reduced from 0.37 to 0.30, a reduction of 18.9%, and MAPE improves from 22.24% to 18.44%. Ablation studies confirm that the Triple-Stage Injection contributes the most, with its removal causing a 25.1% increase in Physical MSE.

The main limitation of the proposed approach is that the current framework relies on a static knowledge graph constructed from historical co-occurrence patterns, which may not generalize well to unseen events or rapidly changing environmental conditions. In addition, the model has only been validated on a single geographic region, leaving its cross-regional transferability unexplored.

Future work will focus on incorporating dynamic knowledge updates to handle emerging events in real time, and developing cross-regional transfer learning strategies to reduce annotation effort when deploying to new geographic areas.

References

1. Hu, M., Wang, Y., Wang, S., Jiao, M., Huang, G., Xia, B.: Spatial-temporal heterogeneity of air pollution and its relationship with meteorological factors in the Pearl River Delta, China. Atmospheric Environment 254, 118415 (2021)
2. Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., Li, T.: Forecasting fine-grained air quality based on big data. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2267–2276 (2015)
3. Brimblecombe, P., Lai, Y.: Effect of Fireworks, Chinese New Year and the COVID-19 Lockdown on Air Pollution and Public Attitudes. Aerosol and Air Quality Research 20, 2318–2331 (2020)
4. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., Grover, A.: ClimaX: A foundation model for weather and climate. In: International Conference on Machine Learning, pp. 25901–25920 (2023)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021)
7. Gao, Z., Tan, C., Wu, L., Li, S. Z.: SimVP: Simpler yet better video prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19056–19066 (2022)
8. Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y., Li, M., Yeung, D.-Y.: Earthformer: Exploring space-time transformers for earth system forecasting. In: Advances in Neural Information Processing Systems, pp. 12905–12918 (2022)
9. Jeczmionek, E., Kowalski, P. A.: Flattening layer pruning in convolutional neural networks. Symmetry 13(7), 1147 (2021)
10. Ahmad, R., Yang, B., Ettlin, G., Berger, A., Rodríguez-Bocca, P.: A machine-learning based ConvLSTM architecture for NDVI forecasting. International Transactions in Operational Research 30(4), 2025–2048 (2023)

11. Chawla, P., Esteves, D., Pujar, K., Lehmann, J.: SimpleLSTM: A Deep-Learning Approach to Simple-Claims Classification. In: EPIA Conference on Artificial Intelligence, pp. 244–255. Springer (2019)
12. Lai, S., Zhang, H., Shao, W., Liu, H., Cai, D., Wang, W., Zhang, K.: Simple CNN for vision. arXiv preprint arXiv:2306.01320 (2023)
13. Ballas, N., Yao, L., Pal, C., Courville, A.: Delving Deeper into Convolutional Networks for Learning Video Representations. In: International Conference on Learning Representations (2016)
14. Banerjee, A., Arora, V.: Enc-Dec RNN Acoustic Word Embeddings learned via Pairwise Prediction. In: Proc. Interspeech 2023, pp. 1478–1482 (2023)
15. Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P. Y., Liang, Y., Li, Y. F., Pan, S., Wen, Q.: Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In: International Conference on Learning Representations (2024)
16. Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling Relational Data with Graph Convolutional Networks. In: The Semantic Web: 15th International Conference (ESWC), pp. 593–607 (2018)
17. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, pp. 424–432 (2016)
18. Tang, Y., Qi, L., Xie, F., Li, X., Ma, C., Yang, M.-H.: PredFormer: Transformers Are Effective Spatial-Temporal Predictive Learners. arXiv preprint arXiv:2410.04733 (2024)
19. Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T.: DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. International Journal of Forecasting 36(3), 1181–1191 (2020)
20. Yu, B., Yin, H., Zhu, Z.: Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pp. 3634–3640 (2018)
21. Zhu, J., Han, X., Deng, H., Tao, C., Zhao, L., Wang, P.: KST-GCN: A Knowledge-Driven Spatial-Temporal Graph Convolutional Network for Traffic Forecasting. IEEE Transactions on Intelligent Transportation Systems 23(9), 15055–15065 (2022)
22. Yuan, Y., Ding, J., Feng, J., Jin, D., Li, Y.: UniST: A Prompt-Empowered Universal Model for Urban Spatio-Temporal Prediction. In: Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2024), pp. 3638–3648 (2024)
23. Ji, Z., Hu, Z., Chang, S., Zhang, W.: STKGformer: A spatio-temporal knowledge graph enhanced transformer for passenger flow prediction. Information Sciences 735, 123055 (2026)
24. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer. In: AAAI Conference on Artificial Intelligence, pp. 3942–3951 (2018)
25. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)